

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

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

October 2012

802.1aq and 802.1Qbp Support over EVPN  
draft-allan-l2vpn-spbm-evpn-02

Abstract

This document describes how Ethernet Shortest Path Bridging MAC mode (802.1aq) and (802.1Qbp) can be combined with EVPN in a way that interworks with PBB-PEs as described in the PBB-EVPN solution in a way that permits operational isolation of each Ethernet network subtending an EVPN core while supporting full interworking between the 3 variations of Ethernet operation.

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 2013.

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

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

## 1. Introduction

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

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

### 1.1. Authors

David Allan, Jeff Tantsura, Don Fedyk, Ali Sajassi

### 1.2. Requirements Language

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

## 2. Conventions used in this document

### 2.1. Terminology

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

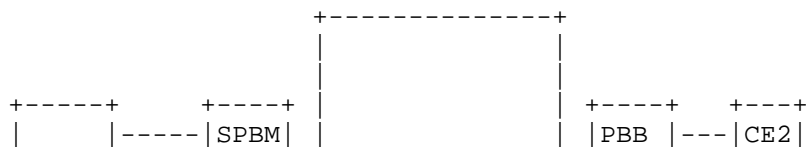
MVPN: Multicast VPN  
NLRI: Network layer reachability information  
PBBN: Provider Backbone Bridged Network  
PBB-PE: Co located BEB and PE  
PE: provider edge  
P2MP: Point to Multipoint  
P2P: Point to Point  
RD: Route Distinguisher  
SPB: Shortest path bridging  
SPBM: Shortest path bridging MAC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

### 3. Changes since previous version

- 1) Change of term from MES to PE to align with base draft.
- 2) Introduction of B-MAC advertisement route NLRI to compress B\_MAC associated I-SID information.

#### 4. Solution Overview

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



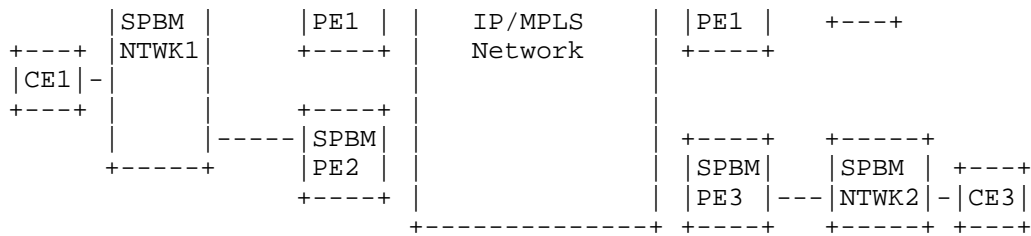


Figure 1: PBB and SPBM EVPN Network

Each EVPN is identified by a route target. The route target identifies the set of SPB islands and BEB-PEs that are allowed to communicate. This manifests itself as a set of Ethernet segments, where each Ethernet segment ID is unique within the route target.

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

Each SPBM island is administered to have an associated Ethernet Segment ID (ESI) associated with it.

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

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

## 5. Elements of Procedure

### 5.1. PE Configuration

At SPBM island commissioning a PE is configured with:

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

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

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

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

### 5.2. DF Election

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

### 5.3. Control plane interworking ISIS-SPB to EVPN

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

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

The format of the B-MAC advertisement route TLV is:

```

+-----+
| RD (8 octets) |
+-----+
| Ethernet Segment Identifier (10 octets) |
+-----+
| B-MAC Address Length (1 octet) |
+-----+
| B-MAC Address (6 octets) |
+-----+
| MPLS Label (n * 3 octets) |
+-----+
| version (4 bits) | 0 0 0 0 |
+-----+
| Base I-SID (3 octets) |
+-----+
| I-SID vector length (1 octet) |
+-----+
| I-SID vector |
// //
+-----+

```

- the Route Distinguisher (RD) is set to that of the PE
- the ESI is that of the SPBM island
- B-MAC address length/B-MAC address encode the MAC address of the advertising BEB
- The MPLS label encodes the label value to be used (does not have to be unique)
- The version identifies how the I-SID information is encoded. This is set to 0000b.
- The base I-SID value identifies the I-SID value associated with the start of the I-SID vector.
- The I-SID vector is a bit vector which uses 3 bits to define the interest in each I-SID value encoded (from base I-SID to base I-SID plus I-SID vector length-1)

The encoding is:

Bit 0 =1 if the BEB has registered interest in the I-SID, =0 otherwise.

Bit 1 =1 if the BEB has registered multicast transmit interest,  
= 0 otherwise, or if Bit 0 =0.

Bit 2 =1 if the BEB has registered multicast receive interest,  
= 0 otherwise or if Bit 0 =0.

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

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

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

#### 5.4. Control plane interworking EVPN to ISIS-SPB

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

The NLRI information advertised into ISIS-SPB is also used to locally populate a forwarding table indexed by B-MAC/I-SID that points to the label stack to impose on the SPBM frame. The bottom label being that offered in the NLRI.

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

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



#### 5.6. Data plane Interworking EVPN to 802.1aq SPBM island

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

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

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

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

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

#### 5.8. Dataplane interworking between 802.1Qbp islands and EVPN

For a future version of the document

#### 5.9. Multicast Stitching

For a future version of the document

### 6. Other Aspects

#### 6.1. Flow Ordering

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

#### 6.2. Transit

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

### 7. Acknowledgements

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

## 8. Security Considerations

For a future version of this document.

## 9. IANA Considerations

For a future version of this document.

## 10. References

### 10.1. Normative References

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

### 10.2. Informative References

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

## 11. Authors' Addresses

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

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

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

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

L2VPN Working Group  
Internet-draft  
Intended Status: Proposed Standard  
Expires: February 2013

Bhargav Bhikkaji  
Balaji Venkat Venkataswami  
Ramasubramani Mahadevan  
Shivakumar Sundaram  
Narayana Perumal Swamy  
DELL-Forcel0  
August 3, 2012

Connecting Disparate TRILL-based Data Center/PBB/Campus sites using BGP  
draft-balaji-l2vpn-trill-over-ip-multi-level-02

## Abstract

There is a need to connect (a) TRILL based data centers or (b) TRILL based networks which provide Provider Backbone like functionalities or (c) Campus TRILL based networks over the WAN using one or more ISPs that provide regular IP+GRE or IP+MPLS transport. A few solutions have been proposed as in [1] in the recent past that have not looked at the PB-like functionality. These solutions have not dealt with the details as to how these services could be provided such that multiple TRILL sites can be inter-connected with issues like nick-name collisions for unicast and multicast being taken care of. It has been found that with extensions to BGP the problem statement which we will define below can be handled. Both control plane and data plane operations can be driven into the solution to make it seamlessly look at the entire set of TRILL sites as a single entity which then can be viewed as one single Layer 2 cloud. MAC moves across TRILL sites and within TRILL sites can be realized. This document / proposal envisions the use of BGP-MAC-VPN vrfs both at the IP cloud PE devices and at the peripheral PEs within a TRILL site providing Provider Backbone like functionality. We deal in depth with the control plane and data plane particulars for unicast and multicast with nick-name election being taken care of as part of the solution.

In this version of the draft, we discuss how hierarchical MAC addresses can be doled out to the end stations thus reducing the size of the BGP-MAC-VPN VRFs in the IP+GRE or IP+MPLS edge devices. We also discuss how the MAC-Moves which involve changing the IP to MAC address associations where the IP addresses remain constant when VMs or physical servers (without VMs) are removed from one part of the network and moved to another even between Trill Data Center sites.

## Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the

provisions of BCP 78 and BCP 79.

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

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

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

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

#### Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

1	Introduction . . . . .	4
1.1	Acknowledgements . . . . .	4
1.2	Terminology . . . . .	4
1.2	Problem Statement . . . . .	5
1.2.1	TRILL Data Centers requiring connectivity over WAN . . .	5
1.2.2	Provider Backbone remote TRILL cloud requirements . . .	6
1.2.3	Campus TRILL network requirements . . . . .	7
2.	Architecture where the solution applies . . . . .	7
2.1	Proposed Solution . . . . .	7
2.1.1	Control Plane . . . . .	8

2.1.1.1 Nickname Collision Solution . . . . .	8
2.1.1.2 U-PE BGP-MAC-VPN VRFs . . . . .	9
2.1.1.3 Control Plane explained in detail. . . . .	11
2.1.2 Corresponding Data plane for the above control plane example. . . . .	12
2.1.2.1 Control plane for regular Campus and Data center sites . . . . .	13
2.1.2.2 Other Data plane particulars. . . . .	16
2.1.3 Encapsulations . . . . .	21
2.1.3.1 IP + GRE . . . . .	21
2.1.3.2 IP + MPLS . . . . .	21
2.2 Other use cases . . . . .	21
2.3 Novelty . . . . .	21
2.4 Uniqueness and advantages . . . . .	22
2.4.1 Multi-level IS-IS . . . . .	23
2.4.2 Benefits of the VPN mechanism . . . . .	23
2.4.3 Inter-working with other VXLAN, NVGRE sites . . . . .	23
2.4.4 Benefits of using Multi-level . . . . .	23
2.5 Comparison with OTV and VPN4DC and other schemes . . . . .	24
2.6 Multi-pathing . . . . .	24
2.7 TRILL extensions for BGP . . . . .	24
2.7.1 Format of the MAC-VPN NLRI . . . . .	24
2.7.2. BGP MAC-VPN MAC Address Advertisement . . . . .	25
2.7.2.1 Next hop field in MP_REACH_NLRI . . . . .	26
2.7.2.2 Route Reflectors for scaling . . . . .	26
2.7.3 Multicast Operations in Interconnecting TRILL sites . . . . .	26
2.7.4 Comparison with PBB-EVPN . . . . .	29
2.7.4.1 No nickname integration issues in our scheme . . . . .	29
2.7.4.2 Hierarchical Nicknames and their disadvantages in the PBB-EVPN scheme . . . . .	29
2.7.4.3 Load-Balancing issues with respect to PBB-EVPN . . . . .	30
2.7.4.4 Technology Agnostic for interworking between TRILL and Non-TRILL sites . . . . .	30
2.7.5 Conversational C-MACs only in the N-PE VRF MAC table . . . . .	30
2.7.5.1 VLAN filtering at U-PEs. . . . .	31
2.7.6 Table sizes in hardware will increase . . . . .	31
2.7.7 The N-PE and its implementation . . . . .	31
2.7.8 Hierarchical MAC addresses that shrink table sizes . . . . .	31
2.7.8.1 MAC-Moves with hierarchical MAC addresses . . . . .	32
3 Security Considerations . . . . .	33
4 IANA Considerations . . . . .	33
5 References . . . . .	33
5.1 Normative References . . . . .	33
5.2 Informative References . . . . .	33
Authors' Addresses . . . . .	34
A.1 Appendix I . . . . .	35

## 1 Introduction

There is a need to connect (a) TRILL based data centers or (b) TRILL based networks which provide Provider Backbone like functionalities or (c) Campus TRILL based networks over the WAN using one or more ISPs that provide regular IP+GRE or IP+MPLS transport. A few solutions have been proposed as in [1] in the recent past that have not looked at the Provider Backbone-like functionality. These solutions have not dealt with the details as to how these services could be provided such that multiple TRILL sites can be interconnected with issues like nick-name collisions for unicast (multicast is still TBD) being taken care of. It has been found that with extensions to BGP the problem statement which we will define below can be well handled. Both control plane and data plane operations can be driven into the solution to make it seamlessly look at the entire set of TRILL sites as a single entity which then can be viewed as one single Layer 2 cloud. MAC moves across TRILL sites and within TRILL sites can be realized. This document / proposal envisions the use of BGP-MAC-VPN vrfs both at the IP cloud PE devices and at the peripheral PEs within a TRILL site providing Provider Backbone like functionality. We deal in depth with the control plane and data plane particulars for unicast (multicast is still TBD) with nick-name election being taken care of as part of the solution.

### 1.1 Acknowledgements

The authors would like to thank Janardhanan Pathangi, Anoop Ghanwani for their inputs for this proposal.

### 1.2 Terminology

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

Legend :

U-PE / ARB : User-near PE device or Access Rbridge. U-PEs are edge devices in the Customer site or tier-2 site. This is a Rbridge with BGP capabilities. It has VRF instances for each tenant it is connected to in the case of Provider-Backbone functionality use-case.

U-Ps / CRB : Core Rbridges or core devices in the Customer site that do not directly interact with the Customer's Customer.

N-PE : Network Transport PE device. This is a device with RBridge capabilities in the non-core facing side. On the core facing side it is a Layer 3 device supporting IP+GRE and/or IP+MPLS. On the non-core

facing side it has support for VRFs one for each TRILL site that it connects to. It runs BGP to convey the BGP-MAC-VPN VRF routes to its peer N-PEs. It also supports IGP on the core facing side like OSPF or IS-IS for Layer 3 and supports IP+GRE and/or IP+MPLS if need be. A pseudo-interface representing the N-PE's connection to the Pseudo Level 2 area is provided at each N-PE and a forwarding adjacency is maintained between the near-end N-PE to its remote participating N-PEs pseudo-interface in the common Pseudo Level 2 area.

N-P : Network Transport core device. This device is IP and/or IP+MPLS core device that is part of the ISP / ISPs that provide the transport network that connect the disparate TRILL networks together.

## 1.2 Problem Statement

### 1.2.1 TRILL Data Centers requiring connectivity over WAN

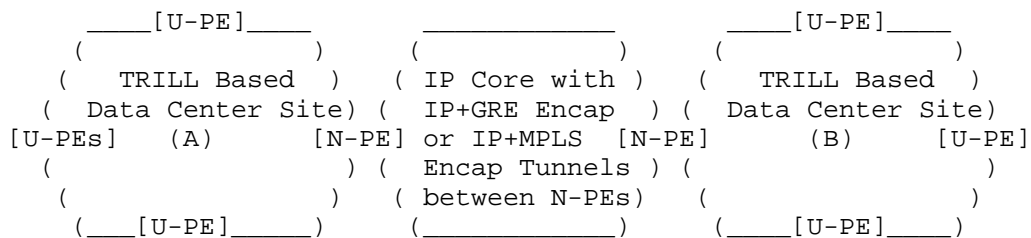


Figure 1.0 : TRILL based Data Center sites inter-connectivity.

- o Providing Layer 2 extension capabilities amongst different disparate data centers running TRILL.
- o Recognizing MAC Moves across data centers and within data centers to enjoin disparate sites to look and feel as one big Layer 2 cloud.
- o Provide a solution agnostic to the technology used in the service provider network
- o Provide a cost effective and simple solution to the above.
- o Provide auto-configured tunnels instead of pre-configured ones in the transport network.
- o Provide additional facilities as part of the transport network for eg., TE, QoS etc
- o Routing and forwarding state is to be maintained at the network edges and not within the site or the core of the transport network.



This requires minimization of the state explosion required to provide this solution.

o So connectivity for end-customers is through U-PE onto N-PE onto remote-N-PE and onto remote U-PE.

### 1.2.2 Provider Backbone remote TRILL cloud requirements

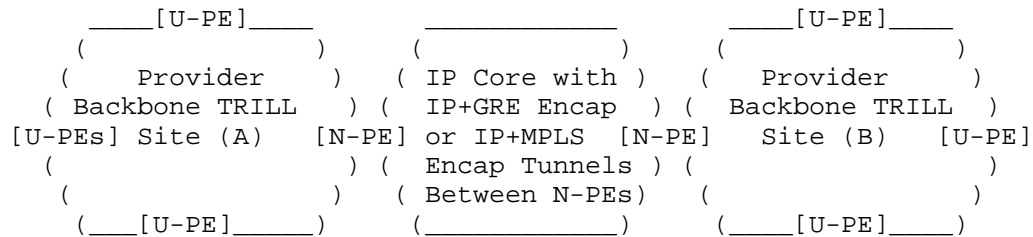


Figure 2.0 : TRILL based Provider backbone sites inter-connectivity

o Providing Layer 2 extension capabilities amongst different Provider Backbone Layer 2 clouds that need connectivity with each other.

o Recognizing MAC Moves across Provider Backbone Layer 2 Clouds and within a single site Layer 2 Cloud to enjoin disparate sites to look and feel as one big Layer 2 Cloud.

o Provide a solution agnostic to the technology used in the service provider network

o Provide a cost effective and simple solution to the above.

o Provide auto-configured tunnels instead of pre-configured ones in the transport network.

o Provide additional facilities as part of the transport network for eg., TE, QoS etc

o Routing and forwarding state is to be maintained at the network edges and not within the site or the core of the transport network. This requires minimization of the state explosion required to provide this solution.

o These clouds could be part of the same provider but be far away from each other. The customers of these clouds could demand connectivity to their sites through these TRILL clouds. These TRILL clouds could offer Provider Layer 2 VLAN transport for each of their customers. Hence Provide a seamless connectivity wherever these sites are placed.

- o So connectivity for end-customers is through U-PE onto N-PE onto remote-N-PE and onto remote U-PE.

### 1.2.3 Campus TRILL network requirements

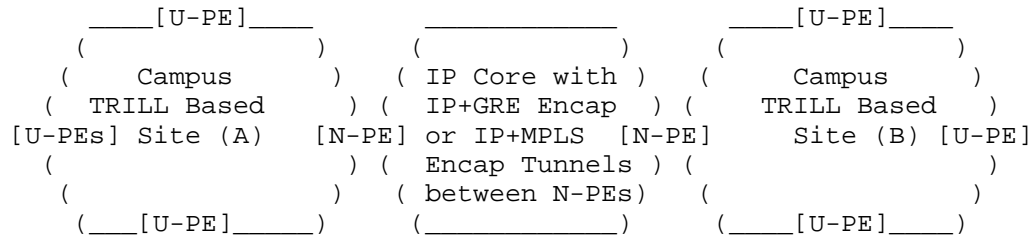


Figure 3.0 : TRILL based Campus inter-connectivity

- o Providing Layer 2 extension capabilities amongst different disparate distantly located Campus Layer 2 clouds that need connectivity with each other.
- o Recognizing MAC Moves across these Campus Layer 2 clouds and within a single site Campus cloud to enjoin disparate sites to look and feel as one Big Layer 2 Cloud.
- o Provide a solution agnostic to the technology used in the service provider network.
- o Provide a cost effective and simple solution to the above.
- o Provide auto-configured tunnels instead of pre-configured ones in the transport network.
- o Provide additional facilities as part of the transport network for eg., TE, QoS etc.
- o Routing and Forwarding state optimizations as in 1.2.1 and 1.2.2.
- o So connectivity for end-customers is through U-PE onto N-PE onto remote-N-PE and onto remote U-PE.

## 2. Architecture where the solution applies

### 2.1 Proposed Solution

The following section outlines (a) Campus TRILL topology or (b) TRILL Data Center topology or (c) Provider backbone Network topology for which solution is intended.

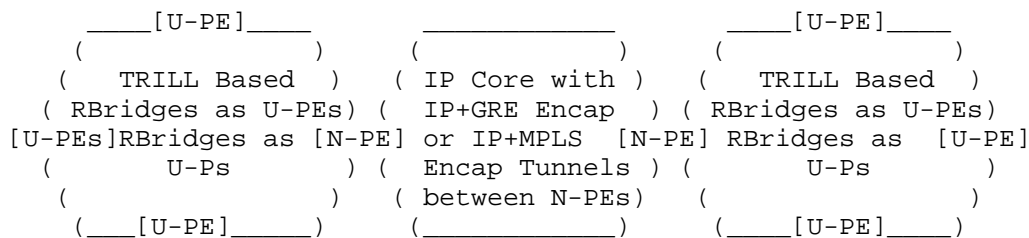


Figure 4.0 : Proposed Architecture

### 2.1.1 Control Plane

o Site network U-PEs still adopt learning function for source MACs bridged through their PE-CE links. For Campus TRILL networks (non-Provider-Backbone networks) the PE-CE links connect the regular hosts / servers. In the case of a data center the PE-CE links connect the servers in a rack to the U-PEs / Top of Rack Switches.

o End customer MACs are placed in BGP-MAC-VPN VRFs in the U-PE to customer PE-CE links. (at tier 2).

#### 2.1.1.1 Nickname Collision Solution

o The near-end N-PE for a site has a forwarding adjacency for the Pseudo Level 2 area Pseudo-Interface to obtain trill nicknames of the next hop far-end N-PE's Level 2 Pseudo-Interface. This forwarding adjacency is built up during the course of BGP-MAC-VPN exchanges between the N-PEs. This forwarding adjacency is a kind of targeted IS-IS adjacency through the IP+GRE or IP+MPLS core. This forwarding adjacency exchange is accomplished through tweaking BGP to connect the near-end N-PE with the far-end N-PEs. Nickname election is done with N-PE Rbridge Pseudo-Interfaces participating in nickname election in Level 2 Area and their non-core facing interfaces which are Level 1 interfaces in the sites in the site considered to be a Level 1 area.

o The Nicknames of each site are made distinct within the site since the nickname election process PDUs for Level 1 area are NOT tunneled across the transport network to make sure that each U-P or U-PE or N-PE's Rbridge interface have knowledge of the nickname election process only in their respective sites / domains. If a new domain is connected as a site to an already existing network then the election process NEED NOT be repeated in the newly added site in order to make sure the nicknames are distinct as Multi-Level IS-IS takes care of forwarding from one site / domain to another. It is only the Pseudo-interface of the N-PE of the newly added site that will have to partake in an election to generate a new Pseudo Level 2 area Nickname

for itself.

#### 2.1.1.2 U-PE BGP-MAC-VPN VRFs

o The Customer MACs are placed as routes in the MAC-VPN VRFs with Nexthops being the area number Nicknames of the U-PEs to which these customer MAC addresses are connected to. For MAC routes within the Level 1 area the Nicknames are those of the local U-PE itself while the MAC routes learnt from other sites have the area number of the site to which the remote U-PE belongs to. When the source learning happens the BGP-MAC-VPN-NLRI are communicated to the participating U-PEs in all the sites of the said customer. Refer to section A.1.1 in Appendix A.1 for more details on how forwarding takes place between the sites through the multi-level IS-IS mechanism orchestrated over the IP core network.

Format of the BGP-MAC-VPN VRF on a U-PE / ARB

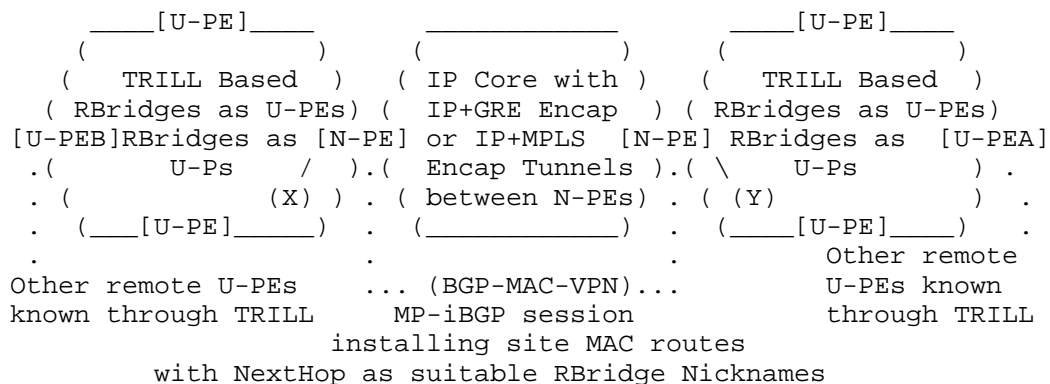
MAC address	U-PE Nickname
00:be:ab:ce:fg:9f (local)	<16-bit U-PE Nickname>
00:ce:cb:fe:fc:0f (Non-local)	<16-bit U-PE Area Num>
....	....

o A VRF is allocated for each customer who in turn may have multiple VLANs in their end customer sites. So in theory a total of 4K VLANs can be supported per customer. The P-VLAN or the provider VLAN in the case of a Provider Backbone category can also be 4K VLANs. So in effect in this scheme upto 4K customers could be supported if P-VLAN encapsulation is to be used to differentiate between multiple customers.

o ISIS for Layer 2 is run atop the Rbridges in the site / Tier-2 network

o ISIS for Layer 2 disseminates MACs reachable via the TRILL nexthop nicknames of site / Tier-2 network Rbridges amongst the Rbridges in the network site.

o N-PEs have VRFs for each tier-2 access network that gain connectivity through the IP+GRE or IP+MPLS core.



Legend :

(X) - Customer A Site 1 MAC-VPN-VRF

(Y) - Customer A Site 2 MAC-VPN-VRF

U-PEs are edge devices a.k.a Access Rbridges (ARBs)

U-Ps a.k.a Core Rbridges (CRBs) are core devices that interconnect U-PEs.

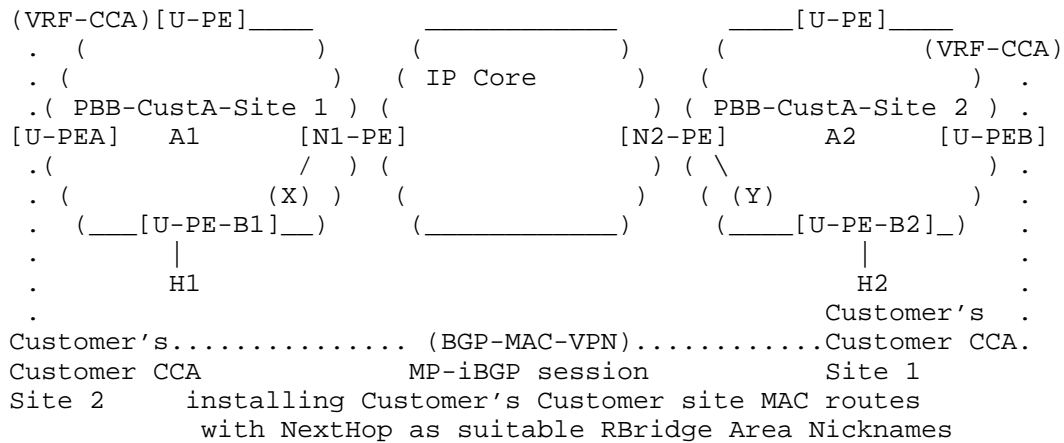
Figure 5.0 : BGP-MAC-VPN VRFs amongst N-PEs

o N-PEs re-distribute the MAC routes in their respective VRFs into the IS-IS Level 1 area after export / import amongst the N-PEs is done. The reverse re-distribution from IS-IS to BGP is also done at each N-PE for its tier-2 customer site.

o N-PEs exchange BGP information through route-targets for various customer sites with other N-PEs. The MAC routes for the various customer sites are placed in the BGP-MAC-VPN VRF of each N-PE for each customer site it connects to on the same lines as U-PE MAC-VPN-VRFs. The MAC routes placed in the VRFs of the N-PEs indicate the MAC addresses for the various Rbridges of the remote tier-2 customer sites with the respective next-hops being the Nicknames of the Level 2 pseudo-interface of the far-end N-PE through which these MAC routes are reachable.

o U-PE and U-P Rbridges MACs and TRILL nicknames are placed in BGP-MAC-VPN vrf on the N-PEs.

o Routes to various end customer MACs within a tier-2 customer's sites are exchanged through BGP MAC-VPN sessions between U-PEs. IP connectivity is provided through IP addresses on same subnet for participating U-PEs.



#### Legend :

A1, A2 - Area Nicknames of the customer sites in TRILL  
 N1, N2 - These are the N-PEs connecting A1 and A2 running BGP sessions  
 B1, B2 - U-PEs in A1 and A2 respectively running BGP sessions  
 H1, H2 - Hosts connected to B1 and B2 U-PEs.  
 Figure 6.0 : BGP-MAC-VPN VRFs between U-PE amongst various sites

#### 2.1.1.3 Control Plane explained in detail.

- 1) B1 and B2 exchange that MACs of H1 and H2 are reachable via BGP. Example., H2-MAC is reachable via B2-MAC through area Nickname A2.
- 2) N1 and N2 exchange that A1 and A2 are reachable through N1 Nickname and N2 Nickname respectively via BGP.
- 3) N1 and N2 also exchange the MACs of U-PEs B1 and B2.
- 4) The routes in the N1 and N2 are re-distributed into IS-IS to end up with the following correlated routing state.

Now the correlated route in B1 is that H2 -> reachable via -> B2 -> reachable via A2 -> reachable via N1 Nickname.

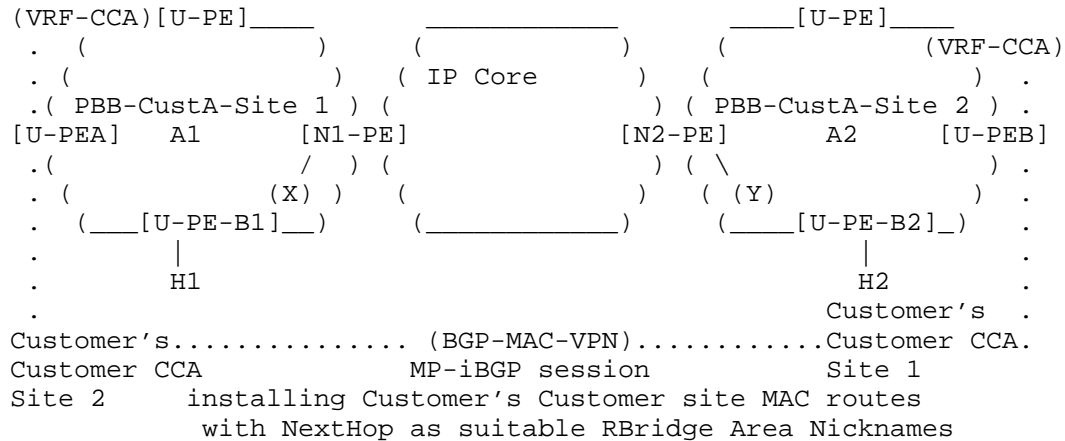
And the correlated route in B2 is that H1 -> reachable via -> B1 -> reachable via A1 -> reachable via N2 Nickname.

And the correlated route in N1 is that B2 -> reachable via -> A2 -> reachable via Nickname N2

And the correlated route in N2 is that B1 -> reachable via -> A1 ->

reachable via Nickname N1

### 2.1.2 Corresponding Data plane for the above control plane example.



#### Legend :

A1, A2 - Area Nicknames of the customer sites in TRILL  
 N1, N2 - These are the N-PEs connecting A1 and A2 running BGP sessions  
 B1, B2 - U-PEs in A1 and A2 respectively running BGP sessions  
 H1, H2 - Hosts connected to B1 and B2 U-PEs.

Figure 6.0 : BGP-MAC-VPN VRFs between U-PE amongst various sites

- 1) H1 sends a packet to B1 with SourceMac as H1-MAC and DestMac as H2-MAC and C-VLAN as C1. This frame is named F1.
- 2) B1 encapsulates this packet in a P-VLAN (Provider VLAN) packet with outer SourceMac as B1-MAC and DestMac as B2-MAC with P-VLAN PV1. This frame is named F2.
- 3) B1 being and Rbridge encapsulates a TRILL header on top of F2, with Ingress Rbridge as B1 and Egress Rbridge as A2.
- 4) This reaches N1 where N1 decapsulates the TRILL header and sends frame F2 inside a IP+GRE header with GRE key as Cust-A's VRF id.
- 5) Packet reaches N2 where N2 looks up the GRE key to identify which customer / VRF to be looked into.
- 6) In that VRF table N2 looks up B2 and encapsulates F2 with TRILL header with Ingress Rbridge as A1 and Egress Rbridge being B2.
- 7) Finally the packet reaches B2 and is decapsulated and sends F1 to

the host.

#### 2.1.2.1 Control plane for regular Campus and Data center sites

For non-PBB like environments one could choose the same capabilities as a PBB like environment with all TORs for e.g in a data center having BGP sessions through BGP Route Reflectors with other TORs. By manipulating the Route Targets specific TORs could be tied in together in the topology within a site or even across sites. The easier way to go about the initial phase of deployment would be to restrict the MP-BGP sessions between N-PEs alone within Campus networks and Data centers and let IS-IS do the job of re-distributing into BGP. Flexibility however can be achieved by letting the U-PEs in the Campus or data center networks too to have MP-BGP sessions. Different logical topologies could be achieved as the result of the U-PE BGP sessions.

##### 2.1.2.1.1 First phase of deployment for Campus and Data Center sites

For the first phase of deployment it is recommended that MP-BGP sessions be constructed between N-PEs alone in case of Data Center and Campus sites. This is necessary as PBB tunnels are not involved. The exchanges remain between the N-PEs about the concerned sites alone and other peering sessions of BGP are not needed since connectivity is the key. When TOR silo based topologies need to be executed then MP-BGP sessions between TORs on the near site and the remote sites can be considered. This will be explored in other documents in the future.

##### 2.1.2.1.2 Control Plane for Data Centers and Campus

- 1) N1 and N2 exchange that A1 and A2 are reachable through N1 Nickname and N2 Nickname respectively via BGP.
- 2) N1 and N2 also exchange that B1 and B2 are within A1 and A2 and that H1 and H2 are attached to B1 and B2 respectively.
- 3) N1 and N2 also exchange the MACs of ARBs B1 and B2.
- 4) The routes in the N1 and N2 are re-distributed into IS-IS to end up with the following correlated routing state.
- 5) The corresponding ESADI protocol routes for end stations will also be exchanged between N-PEs using BGP. The Nickname of the nexthop will be the Area number from which the route originated.

Now the correlated route in B1 is that H2 -> reachable via -> B2 -> reachable via A2 -> reachable via N1 Nickname.



And the correlated route in B2 is that H1 -> reachable via -> B1 -> reachable via A1 -> reachable via N2 Nickname.

And the correlated route in N1 is that B2 -> reachable via -> A2 -> reachable via Nickname N2

And the correlated route in N2 is that B1 -> reachable via -> A1 -> reachable via Nickname N1

#### 2.1.2.1.3 Data Plane for Data Centers and Campus

- 1) H1 sends a packet to B1 with SourceMac as H1-MAC and DestMac as H2-MAC and C-VLAN as C1. This frame is named F1.
- 2) B1 encapsulates this packet with outer SourceMac as B1-MAC and DestMac as B2-MAC. This frame is named F2.
- 3) B1 being an Rbridge encapsulates a TRILL header on top of F2, with Ingress Rbridge as B1 and Egress Rbridge as A2.
- 4) This reaches N1 where N1 decapsulates the TRILL header and sends frame F2 inside a IP+GRE header with GRE key as Cust-A's VRF id.
- 5) Packet reaches N2 where N2 looks up the GRE key to identify which customer / VRF to be looked into.
- 6) In that VRF table N2 looks up B2 and encapsulates F2 with TRILL header with Ingress Rbridge as A1 and Egress Rbridge being B2.
- 7) Finally the packet reaches B2 and is decapsulated and sends F1 to the host.

#### 2.1.2.1.4 Control Plane for Data Centers and Campus networks with more optimizations

In order to avoid double encapsulations at the U-PE / Access Rbridge level it can also be proposed that the U-PE/ARB contain only the Customer MAC addresses and include the N-PE in the default / unknown flood tree. This way the unknown MACs are sent across all participating sites in a VPN. The response will point to the nearest N-PE. The Customer MACs will be learnt by the N-PE for over the core conversations. This way we get rid of the double encapsulations at the ARB level.

- 1) N1 and N2 exchange that A1 and A2 are reachable through N1 Nickname and N2 Nickname respectively via BGP.
- 2) N1 and N2 also exchange that B1 and B2 are within A1 and A2 and

that H1 and H2 are attached to B1 and B2 respectively.

4) The routes in the N1 and N2 are re-distributed into IS-IS to end up with the following correlated routing state.

5) The corresponding ESADI protocol routes for end stations will also be exchanged between N-PES using BGP. The Nickname of the nexthop will be the Area number from which the route originated.

Now the correlated route in B1 is that H2 -> -> reachable via A2 -> reachable via N1 Nickname.

And the correlated route in B2 is that H1 -> -> reachable via A1 -> reachable via N2 Nickname.

#### 2.1.2.1.5 Data Plane Optimizations for Data Centers and Campus

1) H1 sends a packet to B1 with SourceMac as H1-MAC and DestMac as H2-MAC and C-VLAN as C1. This frame is named F1.

2) B1 being an Rbridge encapsulates a TRILL header on top of F2, with Ingress Rbridge as B1 and Egress Rbridge as A2.

3) This reaches N1 where N1 decapsulates the TRILL header and sends frame F2 inside a IP+GRE header with GRE key as Cust-A's VRF id.

5) Packet reaches N2 where N2 looks up the GRE key to identify which customer / VRF to be looked into.

6) In that VRF table N2 looks up H2-MAC and encapsulates F1 with TRILL header with Ingress Rbridge as A1 and Egress Rbridge being B2.

7) Finally the packet reaches B2 and is decapsulated and sends F1 to the host.

## 2.1.2.2 Other Data plane particulars.

Default Dtree which is spanning all sites is setup for P-VLAN for Customer's Customer CCA supported on all Tier-2 sites. Denoted by ==, //.

```
(VRF-CCA)[U-PE]_____
. ( ) ( ) (VRF-CCA)
. ( TRILL Based ) ( IP Core with ) ( TRILL Based ) .
.( Customer A Site 1) ( IP+GRE Encap ) ( Customer A Site 2) .
[U-PEA]=====[N-PE]=====[N-PE]=====[U-PEB]
.( / ) ( Encap Tunnels ) ( \ // ) .
. ( (X) ) ( between N-PEs ) ( (Y) // ) .
. ( ____[U-PE]_____ ) ( ____[U-PE]_____ ) ( ____[U-PE]_____ ) (VRF-CCA)
.
Customer's..... (BGP-MAC-VPN).....Customer CCA.
Customer CCA MP-iBGP session Site 1
Site 2 installing Customer's Customer site MAC routes
with NextHop as suitable RBridge Area Nicknames
```

Legend :

```
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3
```

Figure 8.0 : Dtree spanning all U-PEs for unknown floods.

(1) When a packet comes into a U-PE from the near-end the source MAC is learned and placed in the near-end U-PE BGP-MAC-VPN VRF. This is done in a sub-table depending on which VLAN they belong to in the end-customer's VLANs. The destination MAC if unknown is flooded through a default Spanning tree (could be a dtree) constructed for that provider VLAN which is mapped to carry traffic for the end-customer VLAN in the customer's network sites involved.

Default Dtree which is spanning all sites is setup for P-VLAN for Customer's Customer CCA supported on all Tier-2 sites.

Denoted by ==, //.

Forwarding for unknown frames using the default Dtree spanning all customer sites and their respective U-PEs and onto their customers.

```
(VRF-CCA)[U-PE]_____
. ( ) ( ) (VRF-CCA)
. ( TRILL Based ) ( IP Core with ) ( TRILL Based ) .
.( Customer A Site 1) ( IP+GRE Encap ) ( Customer A Site 2) .
( ) ( ) ( )
[U-PEA]=====[N-PE]=====[N-PE]=====[U-PEB]
.( / ) ( Encap Tunnels ) ( \ // ) .
. ( (X) ) ( between N-PEs ) ( (Y) // ) .
. (____[U-PE]_____) (_____) (____[U-PE]....(VRF-CCA)
.
Customer's..... (BGP-MAC-VPN).....Customer CCA.
Customer CCA MP-iBGP session Site 1
Site 2 installing Customer's Customer site MAC routes
with NextHop as suitable RBridge Area Nicknames
```

Legend :

(X) - Customer A Site 1 MAC-VPN-VRF

(Y) - Customer A Site 2 MAC-VPN-VRF

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

Figure 9.0 : Unknown floods through Dtree spanning for that P-VLAN

(2) The Spanning tree (which could be a dtree for that VLAN) carries that packet through site network switches all the way to N-PEs bordering that network site. U-PEs can drop the packet if there exist no ports for that customer VLAN on that U-PE. The Spanning tree includes auto-configured IP-GRE tunnels or MPLS LSPs across the IP+GRE and/or IP+MPLS cloud which are constituent parts of that tree and hence the unknown flood is carried over to the remote N-PEs participating in the said Dtree. The packet then heads to that remote-end (leaf) U-PEs and on to the end customer sites. For purposes of connecting multiple N-PE devices for a Dtree that is being used for unknown floods, a mechanism such as PIM-Bidir overlay using the MVPN mechanism in the core of the IP network can be used. This PIM-Bidir tree would stitch together all the N-PEs of a specific customer.

(3) BGP-MAC-VPN VRF exchanges between N-PEs carry the routes for MACs

of the near-end Rbridges in the near-end site network to the remote-end site network. At the remote end U-PE a correlation between near-end U-PE and the customer MAC is made after BGP-MAC-VPN VRF exchanges between near-end and far-end U-PEs. The MPLS inner label or the GRE key indicates which VRF to consult for an incoming encapsulated packet at an ingress N-PE and at the outgoing N-PE in the IP core.

(4) From thereon the source MAC so learnt at the far end is reachable just like a Hierarchical VPN case in MPLS Carrier Supporting Carrier. The only difference is that the nicknames of the far-end U-PEs/U-Ps may be the same as the nicknames of the near-end U-PEs/U-Ps. In order to overcome this, the MAC-routes exchanged between the U-PEs have the next-hops as Area nicknames of the far-end U-PE and then the Area number nickname is resolved to the near-end N-PE/N-PEs in the local site that provide connectivity to the far-end U-PE in question.

<srcMac, DstMac> srcMac is known at U-PEA, so advertize to other U-PEs through BGP in the other customer sites for Customer A that srcMAC is reachable via U-PEA. This is received at the BGP-MAC-VPN VRFs in U-PEB and U-PEC.

```
(VRF-CCA)[U-PE]_____ (_____[U-PE]_____
. (_____) (_____) (_____) (VRF-CCA)
. ( TRILL Based ) ( IP Core with ) ( TRILL Based ) .
.( Customer A Site 1) ( IP+GRE Encap ) ( Customer A Site 2) .
( ..... ) ( ..... ) ( ..... ) .
[U-PEA]===== [N-PE]===== [N-PE]===== [U-PEB]
.( / ) ( Encap Tunnels ) ( \ // ) .
. ( (X) ) ( between N-PEs ) ( (Y) // ) .
. (____[U-PE]_____) (_____) (____[U-PEC]....(VRF-CCA)
.
Customer's..... (BGP-MAC-VPN).....Customer CCA.
Customer CCA MP-iBGP session Site 1
Site 2 installing Customer's Customer site MAC routes
with NextHop as suitable RBridge Area Nicknames
```

Legend :

```
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3
```

Figure 10.0 : Distributing MAC routes through BGP-MAC-VPN

<srcMac, DstMac>

Flooding when DstMAC is unknown. The flooding reaches all U-PEs and is forwarded to the customer devices (Customer's customer devices).

```
(VRF-CCA)[U-PE]_____
. ( ) ( ) (VRF-CCA)
. ( TRILL Based ) ( IP Core with ) ( TRILL Based ) .
.( Customer A Site 1) ( IP+GRE Encap ) ( Customer A Site 2) .
( ..... ) ( ..... ) ( ..... ) .
[U-PEA]=====[N-PE]=====[N-PE]=====[U-PEB]
.( / ) ( Encap Tunnels ) ( \ // . ) .
. ( (X) ) ( between N-PEs ) ( (Y) // . ) .
. (____[U-PE]____) ( ) (____[U-PE]....(VRF-CCA)
.
Customer's..... (BGP-MAC-VPN).....Customer CCA.
Customer CCA MP-iBGP session Site 1
Site 2 installing Customer's Customer site MAC routes
with NextHop as suitable RBridge Area Nicknames
```

Legend :

(X) - Customer A Site 1 MAC-VPN-VRF

(Y) - Customer A Site 2 MAC-VPN-VRF

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

Figure 11.0 : Forwarding when DstMAC is unknown.

<srcMac, DstMac>

When DstMAC is known. Payload is carried in the following fashion in the IP core.

(<Outer Ethernet Header, IP+GRE,VRF in GRE key>,  
In PBB like environments / sites interconnected, the payload is P-VLAN headers encapsulating actual payload.

<Outer Ethernet header, P-VLAN header>

<Payload = Ethernet header, Inner VLAN header>, <Actual Payload>)

In Campus and Data Center environments only the latter is carried. There is no P-VLAN header required.

```
(VRF-CCA)[U-PE]_____ (_____) (_____) [U-PE]_____
. ( ) ( ) (VRF-CCA)
. ( TRILL Based ) ( IP Core with ) ( TRILL Based ) .
.( Customer A Site 1) ( IP+GRE Encap ) ( Customer A Site 2) .
( ) ( ) ( ) .
[U-PEA]=====[N-PE]=====[N-PE]=====[U-PEB]
.( / ) ( Encap Tunnels ) ( \ // ) .
. ( (X) ) ( between N-PEs ) ( (Y) // ) .
. (____[U-PE]____) (_____) (____[U-PE]....(VRF-CCA)
. Customer's .
Customer's..... (BGP-MAC-VPN).....Customer CCA.
Customer CCA MP-iBGP session Site 1
Site 2 installing Customer's Customer site MAC routes
with NextHop as suitable RBridge Area Nicknames
```

Legend :

(X) - Customer A Site 1 MAC-VPN-VRF

(Y) - Customer A Site 2 MAC-VPN-VRF

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2

(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

Figure 12.0 : Forwarding when the DstMAC is known.

(5) The reverse path would do the same for reachability of the near-end from the far-end.

(6) Connectivity is thus established between end customer-sites through site networks and through the IP+GRE and/or IP+MPLS core.

(7) End customer packets are carried IP+GRE tunnels or IP+MPLS LSPs through access network site to near-end N-PE in the near-end. N-PE encapsulates this in auto-configured MPLS LSPs or IP+GRE tunnels to

far-end N-PES through the IP+GRE and/or IP+MPLS core. The label is stripped at the far-end N-PE and the inner frame continues to far-end U-PE and onto the customer.

### 2.1.3 Encapsulations

#### 2.1.3.1 IP + GRE

(<Outer Ethernet Header, IP+GRE, VRF in GRE key>,

In PBB like environments...

<Outer Ethernet header, P-VLAN header>,

<Payload = Ethernet header, Inner VLAN header>, <Actual Payload>)

In non-PBB like environments such as Campus and Data Center the Ethernet header with P-VLAN header is not required.

#### 2.1.3.2 IP + MPLS

(<Outer Ethernet Header, MPLS header, VRF in Inner MPLS label>,

In PBB like environments...

<Outer Ethernet header, P-VLAN header>,

<Payload = Ethernet header, Inner VLAN header>, <Actual Payload>)

In non-PBB like environments such as Campus and Data Center the Ethernet header with P-VLAN header is not required.

### 2.2 Other use cases

o Campus to Campus connectivity can also be achieved using this solution. Multi-homing where multiple U-Pes connect to the same customer site can also facilitate load-balancing if a site-id (can use ESI for MAC-VPN-NLRI) is incorporated in the BGP-MAC-VPN NLRI. Mac Moves can be detected if the site-id of the advertised MAC from U-Pes is different from the older ones available.

### 2.3 Novelty

o TRILL MAC routes and their associated nexthops which are TRILL nicknames Are re-distributed into BGP from IS-IS



- o Thus BGP-MAC-VPNs on N-Pes in the transport network contain MAC routes with nexthops as TRILL Area nicknames.
- o The customer edge Rbridges / Provider bridges too contain MAC routes with associated nexthops as TRILL nicknames. This proposal is an extension of BGP-MAC-VPN I-D to include MAC routes with TRILL Area nicknames as Nexthops.

## 2.4 Uniqueness and advantages

- o Uses existing protocols such as IS-IS for Layer 2 and BGP to achieve this. No changes to IS-IS except for redistribution into BGP at the transport core edge and vice-versa.
- o Uses BGP-MAC-VPNs for transporting MAC-updates of customer devices between edge devices only.
- o Employs a hierarchical MAC-route hiding from the core Rbridges of the site. Employs a hierarchical VPN like solution to avoid routing state of sites within the transport core.
- o Multi-tenancy through the IP+GRE or IP+MPLS core is possible when N-PEs at the edge of the L3 core place various customer sites using the VPN VRF mechanism. This is otherwise not possible in traditional networks and using other mechanisms suggested in recent drafts.
- o The VPN mechanism also provides ability to use overlapping MAC address spaces within distinct customer sites interconnected using this proposal.
- o Multi-tenancy within each data center site is possible by using VLAN separation within the VRF.
- o Mac Moves can be detected if source learning / Gratuitous ARP combined with the BGP-MAC-VPN update triggers a change in the concerned VRF tables.
- o PBB like functionality supported where P-VLAN and Customer VLAN are different spaces.
- o Uses regular BGP supporting MAC-VPN features, between transport core edge devices and the Tier-2 customer edge devices.
- o When new TRILL sites are added then no re-election in the Level 1 area is needed. Only the Pseudo-interface of the N-PE has to be added to the mix with the transport of the election PDUs being done across the transport network core.

#### 2.4.1 Multi-level IS-IS

Akin to TRILL IS-IS multi-level draft where each N-PE can be considered as a ABR having one nickname in a customer site which in turn is a level-1 area and a Pseudo Interface facing the core of the transport network which belongs to a Level 2 Area, the Pseudo Interface would do the TRILL header decapsulation for the incoming packet from the Level 1 Area and throw away the TRILL header within the Pseudo Level 2 Area and transport the packets across the Layer 3 core (IP+GRE and/or IP+MPLS) after an encapsulation in IP+GRE or IP+MPLS. Thus we should have to follow a scheme with the NP-E core facing Pseudo-interface in the Level 2 Pseudo-Area doing the TRILL encapsulation and decapsulation for outgoing and incoming packets respectively from and to the transport core. The incoming packets from the Level 1 area are subject to encapsulation in IP+GRE or IP+MPLS by the sending N-PE's Pseudo-Interface and the outgoing packets from the transport core are subject to decapsulation from their IP+GRE or IP+MPLS headers by the Pseudo-Interface on the receiving N-PE.

#### 2.4.2 Benefits of the VPN mechanism

Using the VPN mechanism it is possible that MAC-routes are placed in distinct VRFs in the N-PEs thus providing separation between customers. Assume customer A and customer B have several sites that need to be interconnected. By isolating the routes within specific VRFs multi-tenancy across the L3 core can be achieved. Customer A's sites talk to customer A's sites alone and the same is applicable with Customer B.

The same mechanism also provides for overlapping MAC addresses amongst the various customers. Customer A could use the same MAC-addresses as Customer B. This is otherwise not possible with other mechanisms that have been recently proposed.

#### 2.4.3 Inter-working with other VXLAN, NVGRE sites

Without TRILL header it is possible to inter-work with STP sites, VXLAN sites, NVGRE sites and with other TRILL sites.

For this purpose if for example TRILL site has to inter-operate with VXLAN sites then the VXLAN site has to have a VXLAN gateway that translated plain Ethernet packets coming in from the WAN core into VXLAN packets with the VRF signifying the VXLAN-ID or the VNI.

#### 2.4.4 Benefits of using Multi-level

The benefits of using Multi-level are choosing appropriate Multicast

Trees in other sites through the inter-area multicast method as proposed by Radia Perlman et.al.

## 2.5 Comparison with OTV and VPN4DC and other schemes

- o OTV requires a few proprietary changes to IS-IS. There are less proprietary changes required for this scheme with regard to IS-IS compared to OTV.

- o VPN4DC is a problem statement and is not yet as comprehensive as the scheme proposed in this document.

- o [4] deals with Pseudo-wires being setup across the transport core. The control plane protocols for TRILL seem to be tunneled through the transport core. The scheme in the proposal we make do NOT require anything more than Pseudo Level 2 area number exchanges and those for the Pseudo-interfaces. BGP takes care of the rest of the routing. Also [4] does not take care of nick-name collision detection since the control plane TRILL is also tunneled and as a result when a new site is sought to be brought up into the inter-connection amongst existing TRILL sites, nick-name re-election may be required.

- o [5] does not have a case for TRILL. It was intended for other types of networks which exclude TRILL since [5] has not yet proposed TRILL Nicknames as nexthops for MAC addresses.

## 2.6 Multi-pathing

By using different RDs to export the BGP-MAC routes with their appropriate Nickname next-hops from more than one N-PE we could achieve multi-pathing over the transport IP+GRE and/or IP+MPLS core.

## 2.7 TRILL extensions for BGP

### 2.7.1 Format of the MAC-VPN NLRI

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

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

The Length field indicates the length in octets of the Route Type specific field of MAC-VPN NLRI.

This document defines the following Route Types:

- + 1 - Ethernet Tag Auto-Discovery (A-D) route
- + 2 - MAC advertisement route
- + 3 - Inclusive Multicast Ethernet Tag Route
- + 4 - Ethernet Segment Route
- + 5 - Selective Multicast Auto-Discovery (A-D) Route
- + 6 - Leaf Auto-Discovery (A-D) Route
- + 7 - MAC Advertisement Route with Nexthop as TRILL Nickname

Here type 7 is used in this proposal.

#### 2.7.2. BGP MAC-VPN MAC Address Advertisement

BGP is extended to advertise these MAC addresses using the MAC advertisement route type in the MAC-VPN-NLRI.

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

RD	(8 octets)
MAC Address	(6 octets)
GRE key / MPLS Label rep.	VRF(3 octets)
Originating Rbridge's IP Address	
Originating Rbridge's MAC address	(8 octets)

The RD MUST be the RD of the MAC-VPN instance that is advertising the NLRI. The procedures for setting the RD for a given MAC VPN are described in section 8 in [3].

The encoding of a MAC address is the 6-octet MAC address specified by IEEE 802 documents [802.1D-ORIG] [802.1D-REV].

If using the IP+GRE and/or IP+MPLS core networks the GRE key or MPLS label MUST be the downstream assigned MAC-VPN GRE key or MPLS label that is used by the N-PE to forward IP+GRE or IP+MPLS encapsulated ethernet packets received from remote N-PEs, where the destination MAC address in the ethernet packet is the MAC address advertised in

the above NLRI. The forwarding procedures are specified in previous sections of this document. A N-PE may advertise the same MAC-VPN label for all MAC addresses in a given MAC-VPN instance. Or a N-PE may advertise a unique MAC-VPN label per MAC address. All of these methodologies have their tradeoffs.

Per MAC-VPN instance label assignment requires the least number of MAC-VPN labels, but requires a MAC lookup in addition to a GRE key or MPLS lookup on an egress N-PE for forwarding. On the other hand a unique label per MAC allows an egress N-PE to forward a packet that it receives from another N-PE, to the connected CE, after looking up only the GRE key or MPLS labels and not having to do a MAC lookup.

The Originating Rbridge's IP address MUST be set to an IP address of the PE (U-PE or N-PE). This address SHOULD be common for all the MAC-VPN instances on the PE (e.g., this address may be PE's loopback address).

#### 2.7.2.1 Next hop field in MP\_REACH\_NLRI

The Next Hop field of the MP\_REACH\_NLRI attribute of the route MUST be set to the Nickname of the N-PE or in the case of the U-PE the Area Nickname of the Rbridge one whose MAC address is carried in the Originating Rbridge's MAC Address field.

The BGP advertisement that advertises the MAC advertisement route MUST also carry one or more Route Target (RT) attributes.

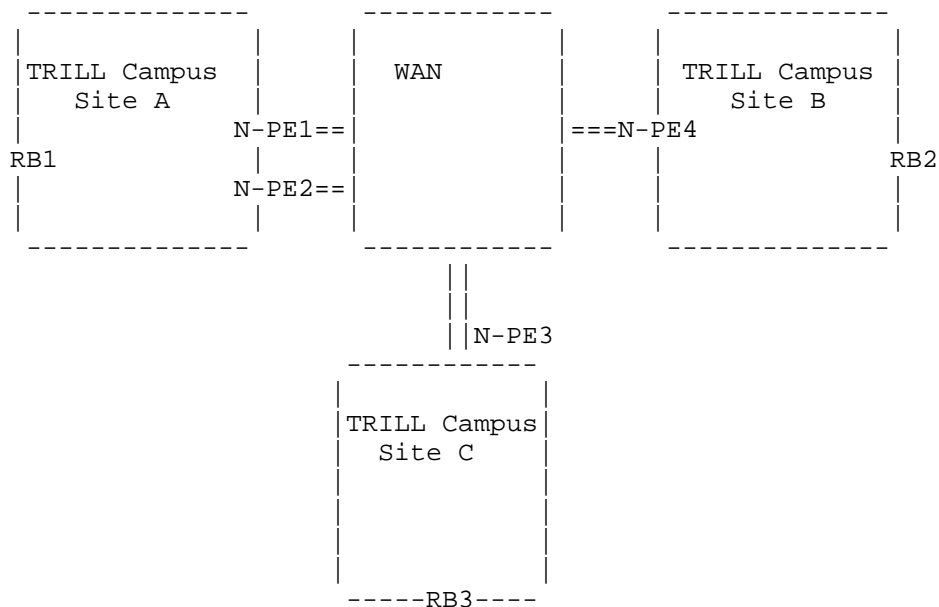
It is to be noted that document [3] does not require N-PEs/U-PEs to create forwarding state for remote MACs when they are learned in the control plane. When this forwarding state is actually created is a local implementation matter. However the proposal in this document requires that forwarding state be established when these MAC routes are learned in the control plane.

#### 2.7.2.2 Route Reflectors for scaling

It is recommended that Route Reflectors SHOULD be deployed to mesh the U-PEs in the sites with other U-PEs at other sites (belonging to the same customer) and the transport network also have RRs to mesh the N-PEs. This takes care of the scaling issues that may arise if full mesh is deployed amongst U-PEs or the N-PEs.

#### 2.7.3 Multicast Operations in Interconnecting TRILL sites

For the purpose of multicast it is possible that the IP core can have a Multicast-VPN based PIM-bidir tree (akin to Rosen or NGEN-MVPN) for each customer that will connect all the N-PEs related to a customer and carry the multicast traffic over the transport core thus connecting site to site multicast trees. Each site that is connected to the N-PE would have the N-PE as the member of the MVPN PIM-Bidir Tree connecting that site to the other sites' chosen N-PE. Thus only one N-PE from each site is part of the MVPN PIM-Bidir tree so constructed. If there exists more than one N-PE per site then that other N-PE is part of a different MVPN PIM-Bidir tree. Consider the following diagram that represents three sites that have connectivity to each other over a WAN. The Site A has 2 N-PEs connected from the WAN to itself and the others B and C have one each. It is to be noted that two MVPN Bidir-Trees are constructed one with Site A's N-PE1 and Site B and C's N-PE respectively while the other MVPN Bidir-tree is constructed with Site A's N-PE2 and site B and C's respective N-PEs. It is possible to load-balance of multicast groups among the sites. The method of interconnecting trees from the respective Level 1 areas (that is the sites) to each other is akin to stitching the Dtrees that have the N-PEs as their stitch end-points in the Pseudo-Level 2 area with the MVPN Bidir tree acting as the conduit for such stitching. The tree-ids in each site are non-unique and need not be distinct across sites. It is only that the N-PEs which have their one foot in the Level 1 area are stitched together using the MVPN Bidir overlay in the Layer 3 core.



Here N-PE1, N-PE3 and N-PE4 form a MVPN Bidir-tree amongst themselves

to link up the multilevel trees in the 3 sites. While N-PE2, N-PE3 and N-PE4 form a MVPN Bidir-tree amongst themselves to up the multilevel trees in the 3 sites.

There exist 2 PIM-Bidir overlay trees that can be used to load-balance say Group G1 on the first and G2 on the second. Lets say the source of the Group G1 lies within Site A and the first overlay tree is chosen for multicasting the stream. When the packet hits the WAN link on N-PE1 the packet is replicated to N-PE3 and N-PE4. It is important to understand that a concept like Group Designated Border Rbridge (GDBR) is applied in this case where group assignments are made to specific N-PEs such that only one of them is active for a particular group and the other does not send it across the WAN using the respective MVPN PIM-Bidir tree. Now Group G2 could then use the MVPN PIM-bidir based tree for its transport. The procedures for election of Group Designated Border Rbridge within a site will be further discussed in detail in future versions of this draft or may be taken to a separate document. VLAN based load-balancing of multicast groups is also possible and feasible in this scenario. It also can be VLAN, Multicast MAC-DA based. The GDBR scheme is applicable only for packets that N-PEs receive as TRILL decapsulated MVPN PIM-Bidir tree frames from the Layer 3 core. If a TRILL encapsulated multicast frame arrives at a N-PE only the GDBR for that group can decapsulate the TRILL header and send it across the Layer 3 core. The other N-PEs can however forward these multi-destination frames coming from N-PEs across the core belonging to a different site.

When the packet originates from the source host the Egress Nickname of the multicast packet is set to the Dtree root at the Level 1 area where the source is originating the stream from. The packet flows along the multicast distribution tree to all Rbridges which are part of the Dtree. Now the N-PE that provides connectivity to the Pseudo-Level 2 area and to other sites beyond it, also receives the packet. The MVPN PIM-bidir tree is used by the near end N-PE to send the packet to all the other member N-PEs of the customer sites and appropriate TRILL encapsulation is done at the ingress N-PE for this multicast stream with the TRILL header containing a local Dtree root on the receiving site and packet streamed to the said receivers in that site. Source suppression such that the packet is not put back on the core, is done by looking at the Group Designated Border Rbridge information at the receiving site. If then other N-PEs which connect the site to the Layer 3 core receive the multicast packet sent into the site by the GDBR for that group then the other N-PEs check if they are indeed the GDBR for the said group and if not they do not forward the traffic back into the core.

It is to be noted that the Group Address TLV is transported by BGP

from across the other sites into a site and it is the GDBR for that group from the remote side that enables this transport. This way the MVPN PIM-bidir tree is pointed to from within each site through the configured GDBR N-PEs for a said group. The GDBR thus lies as one of the receivers in the Dtree for a said group within the site where the multicast stream originates.

#### 2.7.4 Comparison with PBB-EVPN

With respect to PBB-EVPN scheme outlined in [PBB-EVPN], the scheme explained in this document has the following advantages over and above the PBB-EVPN scheme.

##### 2.7.4.1 No nickname integration issues in our scheme

Existing TRILL based sites can be brought into the interconnect without any re-election / re-assignment of nicknames. The one benefit it seems to have vs PBB-EVPN is that adding a new site to a VPN, or merging 2 VPNs, doesn't cause issues with nickname clashes. This is a major advantage since the new TRILL site can hit the ground running without any interruptions to the existing sites in the interconnect.

##### 2.7.4.2 Hierarchical Nicknames and their disadvantages in the PBB-EVPN scheme

The PBB-EVPN scheme advocates the use of Hierarchical Nicknames where the nickname is split into the Site-ID and the Rbridge-ID. The use of the nicknames has the following corollary disadvantages.

(a) The nickname is a 16 bit entity. With a interconnect where there are for eg., 18 sites the PBB-EVPN scheme has to use 5 bits in the nickname bitspace for Site-ID. It wastes  $(32 - 18) = 14$  Site-IDs. The number of sites is also limited to say at best 255 sites.

(b) The nickname is a 16 bit entity. With a interconnect where there are at least 4K Rbridges in each site, the nickname space has to set aside 12 bits at the least in the nickname space for the Rbridge-ID. This means that the Sites cannot be more than  $2^4 = 16$ .

Thus the use of the hierarchical scheme limits the Site-IDs and also the number of Rbridges within the site. If we want to have more Sites we set aside more bits for the Site-ID thus sacrificing maximum number of Rbridge-IDs within the site. If there are more Rbridges within each site, then allocating more bits for the RBridge-ID would sacrifice the maximum number of Site-IDs possible.

For eg., in a branch office scenario if there are 32 sites and more than 255 Rbridges in each of the branch offices it would be difficult



to accomodate the set of sites along with the number of Rbridges using the hierarchical nickname scheme.

In the scheme outlined in this document, it is possible to set aside 1000 nicknames or 2000 nicknames or even 200 nicknames depending on the number of sites (since this is a range of nicknames without hierarchy in the nickname space), without compromising on the maximum number of Rbridges within each site. If M were the number of sites to be supported then the number of Rbridges would be  $2^{16} - M = X$ . This X number would be available to all sites since the nickname is site-local and not globally unique.

It would be possible to set aside a sizeable number within the nickname space for future expansion of sites without compromising on the number of Rbridges within the site.

#### 2.7.4.3 Load-Balancing issues with respect to PBB-EVPN

While PBB-EVPN allows for active/active load-balancing the actual method of distributing the load leads to pinning the flow onto one of the multi-homed N-PEs for a specific site rather than the multi-path hashing based scheme that is possible with our scheme.

#### 2.7.4.4 Technology Agnostic for interworking between TRILL and Non-TRILL sites

Our scheme provides a Technology agnostic method for inter-working between TRILL and non-TRILL sites such as STP-based sites, and other NVO3 schemes for example. This is because the TRILL header is not carried over the L3 core. This is provided as an option in the initial capability exchange between N-PEs when a said pair of N-PEs handshake for BGP. The PBB-EVPN scheme doesnt offer this capability.

#### 2.7.5 Conversational C-MACs only in the N-PE VRF MAC table

It is possible to maintain only conversational MACs on the N-PE table in the case of Campus and Data Center networks by installing the C-MACs in the hardware learned through the site interface or through the Core facing interface only if there arises evidence of across-core conversations. Thus those C-MAC addresses that have been learnt as a result of conversations between Rbridges across sites connecting to hosts that actively communicate with each other are installed in the hardware. Locally switched conversations are not learnt. This is an optimization that will reduce the disadvantage of learning all possible C-MACs located in all the various sites of a VPN on the N-PE. If a one-sided C-MAC is evidenced in the data plane, the learnt

C-MAC is not installed in the hardware unless a reverse path conversation is heard across sites. This C-MAC initially is placed in the software table and the wait begins to hear a reverse path conversation flow. If the wait results in learning that a two-way conversation exists across sites then the software learns are actually programmed in the hardware.

#### 2.7.5.1 VLAN filtering at U-PEs.

It is further possible for the U-PE to filter based on VLANs that it possesses and thus exclude those MAC addresses for VLANs that it does not converse with for the hosts attached to it. This optimizes on the table-size to a large degree since the U-PEs need to know only what they need and not hold all the C-MACs that are in vogue in that site for those conversing VLANs for that Rbridge.

#### 2.7.6 Table sizes in hardware will increase

There may be a concern that table sizes in hardware may be a problem with respect to the C-MAC scaling. With the possibility of having more table sizes in merchant silicon this may no longer be a issue. Also with enhanced lookup tables which may be external to the merchant silicon this problem may no longer be a downside to the scheme proposed in this document.

#### 2.7.7 The N-PE and its implementation

It is possible that the N-PE placed as the border Rbridge and router-PE device respectively on either side of the L3 core, the actual implementation would be in the form of two devices one acting as the border Rbridge and the other as the plain Provider Edge router. The link between the two would be an attachment circuit.

#### 2.7.8 Hierarchical MAC addresses that shrink table sizes

In this section, we discuss how hierarchical MAC addresses can be doled out to the end stations thus reducing the size of the BGP-MAC-VPN VRFs in the IP+GRE or IP+MPLS edge devices. We also discuss how the MAC-Moves which involve changing the IP to MAC address associations where the IP addresses remain constant when VMs of physical servers (without VMs) are removed from one part of the network and moved to another even between Trill Data Center sites.

Consider the case where the end stations with either Virtual Machines managed by hypervisors or physical servers exist behind the U-PEs at each Data Center site. The Hypervisor or the physical server when they are booted up and join the cloud behind the U-PE send Active Directory Requests to an AD-Service. These AD requests are

intercepted by a smart endnode proximal to the U-PE (ARB). The smart endnode (as in a cloudlet specified in [RadiaCloudlet]) requests the AD-Service with information on the U-PEs (ARBs) Rbridge Nickname in that site. This AD-Service is available for each site and has discontinuous sets of Hierarchical MAC address prefixes of length 20 bits to dole out for each U-PE (behind which end stations exist) within a site. These discontinuous sets are unique for all U-PEs in all sites belonging to a particular Trill Data Center inter-connection. The AD-Service replies with hierarchical prefix and the nodes are assigned their addresses based on arbitration from the smart endnode. It is also possible that the AD-service will return the complete MAC address with the hierarchical prefix of 20 bits at the beginning of the address. The smart endnode returns this AD-service request to the end station requesting it. The VM or the physical server whichever the case may be absorbs this address and uses this address to reach out to the other end stations within the site or across sites.

The N-PE begins to learn MAC prefixes alone of the MAC addresses passing through it and the ingress Rbridge nickname to which this prefix was reported from. For MAC prefixes belonging to other sites of the Data Center VPN the Area nickname of the other site from which it came from is learnt from. This is also conveyed to other N-PEs belonging or having BGP-MAC-VPN VRFs of the said VPN with participating DC sites. Thus the table sizes of the BGP-MAC-VPN VRFs is reduced to having only prefixes rather than having complete MAC addresses.

#### 2.7.8.1 MAC-Moves with hierarchical MAC addresses

When the VMs or Physical servers are moved from one site to another VPN site then appropriate Gratuitous ARP requests are sent from the moved VM or end station which then helps the communicating end stations or VMs to that moving end station or VM to re-assign their IP to MAC address mapping. This is because the move would have changed the hierarchical prefix of the moving stations MAC address based on the U-PE to which the end station attaches to after the move. Appropriate mechanisms are already present to make this change.

### 3 Security Considerations

TBD.

### 4 IANA Considerations

A few IANA considerations need to be considered at this point. A proper AFI-SAFI indicator would have to be provided to carry MAC addresses as NLRI with Next-hops as Rbridge Nicknames. This one AFI-SAFI indicator could be used for both U-PE MP-iBGP sessions and N-PE MP-iBGP sessions. For transporting the Group Address TLV suitable extensions to BGP must be done and appropriate type codes assigned for the transport of such TLVs in the BGP-MAC-VPN VRF framework.

### 5 References

#### 5.1 Normative References

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

#### 5.2 Informative References

- [1] draft-xl-trill-over-wan-00.txt, XiaoLan. Wan et.al December 11th ,2011 Work in Progress
- [2] draft-perlman-trill-rbridge-multilevel-03.txt, Radia Perlman et.al October 31, 2011 Work in Progress
- [3] draft-raggarwa-mac-vpn-01.txt, Rahul Aggarwal et.al, June 2010, Work in Progress.
- [4] draft-yong-trill-trill-o-mps, Yong et.al, October 2011, Work in Progress.
- [5] draft-raggarwa-sajassi-l2vpn-evpn Rahul Aggarwal et.al, September 2011, Work in Progress.

[RadiaCloudlet] draft-perlman-trill-cloudlet-00, Radia Perlman et.al, July 30 2012, Work in Progress.

[EVILBIT] Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, April 1 2003.

[RFC5513] Farrel, A., "IANA Considerations for Three Letter Acronyms", RFC 5513, April 1 2009.

[RFC5514] Vyncke, E., "IPv6 over Social Networks", RFC 5514, April 1 2009.

#### Authors' Addresses

Bhargav Bhikkaji,  
Dell-Force10,  
350 Holger Way,  
San Jose, CA  
U.S.A

Email: Bhargav\_Bhikkaji@dell.com

Balaji Venkat Venkataswami,  
Dell-Force10,  
Olympia Technology Park,  
Fortius block, 7th & 8th Floor,  
Plot No. 1, SIDCO Industrial Estate,  
Guindy, Chennai - 600032.  
TamilNadu, India.  
Tel: +91 (0) 44 4220 8400  
Fax: +91 (0) 44 2836 2446

Email: BALAJI\_VENKAT\_VENKAT@dell.com

Ramasubramani Mahadevan,  
Dell-Force10,  
Olympia Technology Park,  
Fortius block, 7th & 8th Floor,  
Plot No. 1, SIDCO Industrial Estate,  
Guindy, Chennai - 600032.  
TamilNadu, India.

Tel: +91 (0) 44 4220 8400  
Fax: +91 (0) 44 2836 2446

EMail: Ramasubramani\_Mahade@dell.com

Shivakumar Sundaram,  
Dell-Force10,  
Olympia Technology Park,  
Fortius block, 7th & 8th Floor,  
Plot No. 1, SIDCO Industrial Estate,  
Guindy, Chennai - 600032.  
TamilNadu, India.  
Tel: +91 (0) 44 4220 8400  
Fax: +91 (0) 44 2836 2446

EMail: Shivakumar\_sundaram@dell.com

Narayana Perumal Swamy,  
Dell-Force10,  
Olympia Technology Park,  
Fortius block, 7th & 8th Floor,  
Plot No. 1, SIDCO Industrial Estate,  
Guindy, Chennai - 600032.  
TamilNadu, India.  
Tel: +91 (0) 44 4220 8400  
Fax: +91 (0) 44 2836 2446

Email: Narayana\_Perumal@dell.com

## A.1 Appendix I

### A.1.1 Extract from Multi-level IS-IS draft made applicable to scheme

In the following picture, RB2 and RB3 are area border RBridges. A source S is attached to RB1. The two areas have nicknames 15961 and 15918, respectively. RB1 has a nickname, say 27, and RB4 has a nickname, say 44 (and in fact, they could even have the same nickname, since the RBridge nickname will not be visible outside the area).



field will be 15918. Also RB2 learns that S is attached to nickname 27 in area 15961 to accommodate return traffic.

- The frame is forwarded through Level 2, to RB3, which has advertised, in Level 2, reachability to the nickname 15918.
- RB3, when forwarding into area 15918, replaces the egress nickname in the TRILL header with RB4's nickname (44). So, within the destination area, the ingress nickname will be 15961 and the egress nickname will be 44.
- RB4, when decapsulating, learns that S is attached to nickname 15961, which is the area nickname of the ingress.

Now suppose that D's location has not been learned by RB1 and/or RB3. What will happen, as it would in TRILL today, is that RB1 will forward the frame as a multi-destination frame, choosing a tree. As the multi-destination frame transitions into Level 2, RB2 replaces the ingress nickname with the area nickname. If RB1 does not know the location of D, the frame must be flooded, subject to possible pruning, in Level 2 and, subject to possible pruning, from Level 2 into every Level 1 area that it reaches on the Level 2 distribution tree.

UNQUOTE...

In the current proposal that we outline in this document, the TRILL header is done away with completely in the IP+GRE or IP+MPLS core. A re-look into the inner headers after decapsulation gives the appropriate information to carry the frame from the N-PE towards the destination U-PE.



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 22, 2013

S. Sivabalan  
S. Boutros  
L. Martini  
N. McGill  
Cisco Systems, Inc.  
October 19, 2012

MAC Address Withdrawal over Static Pseudowire  
draft-boutros-l2vpn-mpls-tp-mac-wd-01.txt

Abstract

This document specifies a mechanism to signal MAC address withdrawal notification using PW Associated Channel (ACH). Such notification is useful when statically provisioned PWs are deployed in VPLS/H-VPLS environment.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 22, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

## Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	3
3. MAC Withdraw OAM Message . . . . .	4
4. Operation . . . . .	5
4.1. Operation of Sender . . . . .	5
4.2. Operation of Receiver . . . . .	5
5. Security Considerations . . . . .	6
6. IANA Considerations . . . . .	6
7. References . . . . .	6
7.1. Normative References . . . . .	6
7.2. Informative References . . . . .	6
Authors' Addresses . . . . .	7

## 1. Introduction

An LDP-based MAC Address Withdrawal Mechanism is specified in [RFC4762] to remove dynamically learned MAC addresses when the source of those addresses can no longer forward traffic. This is accomplished by sending an LDP Address Withdraw Message with a MAC List TLV containing the MAC addressed to be removed to all other PEs over LDP sessions. When the number of MAC addresses to be removed is large, empty MAC List TLV may be used. [MAC-OPT] describes an optimized MAC withdrawal mechanism which can be used to remove only the set of MAC addresses that need to be re-learned in H-VPLS networks. The solution also provides optimized MAC Withdrawal operations in PBB-VPLS networks.

A PW can be signaled via LDP or can be statically provisioned. In the case of static PW, LDP based MAC withdrawal mechanism cannot be used. This is analogous to the problem and solution described in [RFC4762] where PW OAM message has been introduced to carry PW status TLV using in-band PW Associated Channel. In this document, we propose to use PW OAM message to withdraw MAC address(es) learned via static PW.

## 2. Terminology

The following terms are defined in this document:

ACK: Acknowledgement for MAC withdraw message.

LDP: Label Distribution Protocol.

MAC: Media Access Control.

PE: Provide Edge Node.

MPLS: Multi Protocol Label Switching.

PW: PseudoWire.

PW OAM: PW Operations, Administration and Maintenance.

TLV: Type, Length, and Value.

VPLS: Virtual Private LAN Services.



MAC List TLV are governed by [RFC4762], and the corresponding rules of MAC Flush Parameter TLV are governed by [MAC-OPT].

"TLV Length" is the total length of all TLVs in the message, and "Sequence Number TLV Length" is the length of the sequence number field.

A single bit (called A-bit) is set to indicate if a MAC withdraw message is for ACK. Also, ACK does not include MAC TLV(s).

Only half of the sequence number space is used. Modular arithmetic is used to detect wrapping of sequence number. When sequence number wraps, all MAC addresses are flushed and the sequence number is reset.

#### 4. Operation

This section describes how the initial MAC withdraw OAM messages are sent and retransmitted, as well as how the messages are processed and retransmitted messages are identified.

##### 4.1. Operation of Sender

Each PW is associated with a counter to keep track of the sequence number of the transmitted MAC withdrawal messages. Whenever a node sends a new set of MAC TLVs, it increments the transmitted sequence number counter, and include the new sequence number in the message.

The sender expects an ACK from the receiver within a time interval which we call "Retransmit Time" which can be either a default or configured value. If the ACK arrives within the Retransmit Time, the sender assumes that the message transmission is successful. Otherwise, it retransmits the message with the same sequence number as the original message.

##### 4.2. Operation of Receiver

Each PW is associated with a register to keep track of the sequence number of the MAC withdrawal message received last. Whenever a MAC withdrawal message is received, and if the sequence number on the message is greater than the value in the register, the MAC address(es) contained in the MAC TLV(s) is/are removed, and the register is updated with the received sequence number. The receiver sends an ACK whose sequence number is the same as that in the received message.

If the sequence number in the received message is smaller than or

equal to the value in the register, the MAC TLV(s) is/are not processed. However, an ACK whose sequence number is the same as that in the received message is sent.

As mentioned above, since only half of the sequence number space is used, the receiver MUST use modular arithmetic to detect wrapping of the sequence number.

## 5. Security Considerations

This document does not introduce any additional security constraints.

## 6. IANA Considerations

The proposed mechanism requests IANA to assign new channel type (recommended value 0x0028) from the registry named "Pseudowire Associated Channel Types". The description of the new channel type is "Pseudowire MAC Withdraw OAM Channel".

IANA needs to create a new registry for Pseudowire Associated Channel TLVs, and create an entry for "Sequence Number TLV". The recommended value is 0x0001.

## 7. References

### 7.1. Normative References

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

### 7.2. Informative References

[MAC-OPT] Pranjal, P., Balus, F., and G. Calvignac, "LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS", draft-ietf-l2vpn-vpls-ldp-mac-opt-07.txt (work in progress), 2012.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC6478] Martini, L., Swallow, G., Heron, G., and M. Bocci, "Pseudowire Status for Static Pseudowires", RFC 6478, May 2012.

Authors' Addresses

Siva Sivabalan  
Cisco Systems, Inc.  
2000 Innovation Drive  
Kanata, Ontario K2K 3E8  
Canada

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

Sami Boutros  
Cisco Systems, Inc.  
170 West Tasman Dr.  
San Jose, CA 95134  
USA

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

Luca Martini  
Cisco Systems, Inc.  
170 West Tasman Dr.  
San Jose, CA 95134  
USA

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

Neil McGill  
Cisco Systems, Inc.  
7100-9 Kit Creek Road, PO Box 14987  
RESEARCH TRIANGLE PARK, NC 27709-4987  
USA

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





INTERNET-DRAFT  
Intended Status: Standard Track

Dennis Cai  
Sami Boutros  
Samer Salam  
Reshad Rahman  
October 13, 2012

Expires: April 16, 2013

VLAN Aware VPLS services  
draft-cai-l2vpn-vpls-vlan-aware-bundling-00.txt

## Abstract

This document specifies VPLS extensions to support the new VLAN aware bundling service interface type. The new service interface type provides advantages in reducing the provisioning overhead, as well as pseudowire scalability in environments where a large number of VLANs need to be extended over an MPLS/IP network while maintaining traffic segregation among those VLANs.

The VLAN aware bundling service interface can handle the high scale requirements of today's Data Centers by bundling different VLANs over a single WAN VPLS instance used to interconnect sites. Furthermore, this document specifies an extension to the LDP MAC Withdrawal mechanisms to allow per-VLAN MAC flushing for the new service interface type.

## Status of this Memo

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

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

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

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

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

## Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1	Introduction . . . . .	3
1.1	Terminology . . . . .	3
2.	VLAN-aware-bundling PW . . . . .	4
3.	PW VLAN Vector TLV . . . . .	4
4.	LDP Capability Negotiation . . . . .	5
5.	Operation . . . . .	6
5.1.	Packet forwarding, MAC learning, aging and flushing . . . . .	7
5.2.	Multicast Pruning . . . . .	7
5.3.	OAM . . . . .	7
5.4.	VLAN translation . . . . .	7
6.	Security Considerations . . . . .	7
7.	IANA Considerations . . . . .	8
8	References . . . . .	8
8.1	Normative References . . . . .	8
8.2	Informative References . . . . .	8
	Authors' Addresses . . . . .	9

## 1 Introduction

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

These new VLAN aware bundling service interfaces MUST:

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

As discussed in [EVPN-REQ], two new service interface types are defined for VLAN aware bundling: with and without translation. The new service interfaces maintain data-plane separation, per VLAN, while sharing one L2VPN VPN instance. In this document, we focus on the scenario where VPLS is the L2VPN technology. This document defines a new PW VLAN Vector TLV to be included in the LDP PW FEC label mapping messages for the VPLS service, using the mechanisms specified in RFC 4762, as well as a new LDP capability by which a PE can specify its ability to support this new VLAN aware bundling service interface type. Furthermore, This document defines extension to the PWE3 control protocol [RFC4447] to set up the new VLAN aware bundling type service in MPLS networks. The document specifies as well an extension to the MAC Withdrawal mechanisms to allow per VLAN service MAC flushing for this new VLAN aware bundling service.

### 1.1 Terminology

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

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.

## 2. VLAN-aware-bundling PW

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

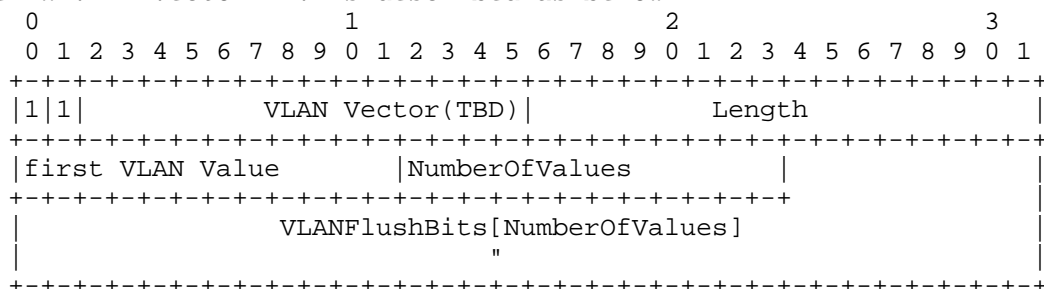
PW type port or raw mode will be used for the VLAN aware bundling interface type service.

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

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

## 3. PW VLAN Vector TLV

The PW VLAN Vector TLV is described as below:



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

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

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

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

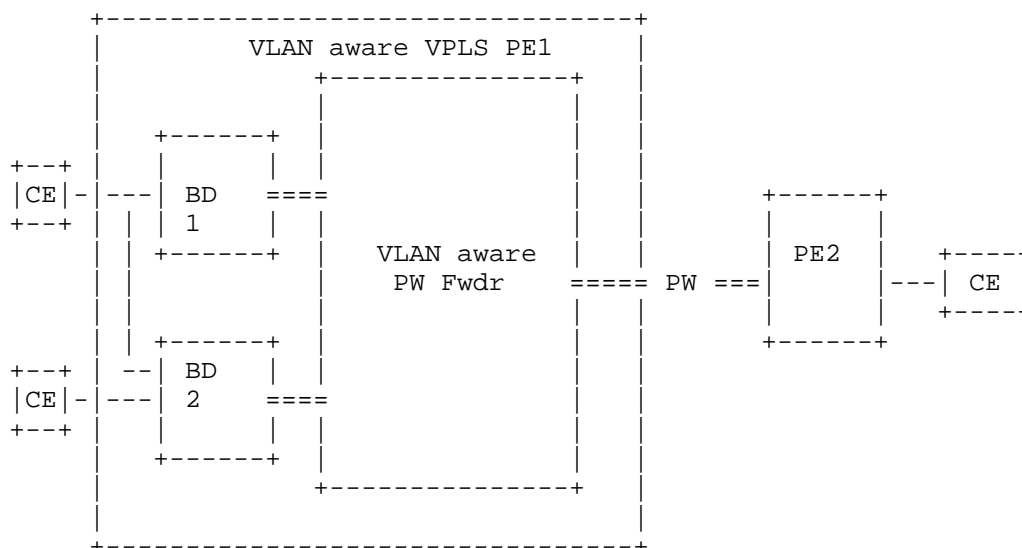


0 - The TLV is withdrawing the capability specified by the TLV Code Point.

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

## 5. Operation

The following figure shows the VPLS PE model for supporting the VLAN aware service interface.



One VPLS instance has been set up between two sites to extend multiple customer VLANs. On each site, multiple CE devices could be connected to the PE. The link between the CE and the PE could be 802.1q or 802.1ad, setup with multiple VLANs. Unlike a classic VPLS solution that requires a dedicated VPLS instance for each customer VLAN, only a single VPLS instance has been set up to carry customer VLANs between the two sites. The use of two sites in the above figure is for illustration; however, this could be extended to many sites. In order to quantify the benefit of the approach, let's assume N data center sites, with M customer VLANs. Classic VPLS full mesh solution would require M VPLS instances and  $M \cdot (N-1)$  PWs on each PE. While with the new VLAN aware interface service type, the solution would require one VPLS instance and will only require  $(N-1)$  PWs on each PE. To maintain data-plane separation per customer VLAN, with the new VLAN aware interface service, each PE will create a bridge-domain per customer VLAN. As well, a customer VLAN on each CE port will represent a unique bridge port in the customer bridge-domain. Only one VPLS instance would be signaled in the core and will be used to carry multiple customer bridge-domains (or customer VLANs) as long as

those customer VLANs need to be extended to the same set of sites. Unlike classic VPLS, where the VPLS PW is presented as a bridge port, the VFI and the customer VLAN would map to the customer bridge-domain.

#### 5.1. Packet forwarding, MAC learning, aging and flushing

Given the data-plane separation, packet forwarding in the scope of one bridge-domain will remain unchanged. When sending traffic over the PW, a qualifying VLAN tag MUST be present on the packet. This VLAN tag has global significance across all sites connected to the VPLS instance and is used to identify the customer bridge domain in all sites. MAC learning, aging and flushing per bridge-domain will remain un-changed. Extensions to MAC withdrawal mechanisms, as described in section 4, would allow the MAC flushing to occur on a subset of the customer bridge-domains.

#### 5.2. Multicast Pruning

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

#### 5.3. OAM

Customer Ethernet OAM frames (e.g. CFM [802.1ag]) will be carried transparently over the shared VPLS instance by the customers bridge-domains. Current VCCV mechanisms can be used to verify PWs connectivity in the VPLS instance shared by the customer bridge-domains. VPLS OAM framework as defined in [RFC6136] applies to this new service with no changes.

#### 5.4. VLAN translation

As mentioned above, the VLAN tag carried across the PWs for the new VLAN aware bundling VPLS instance MUST have network global significance within the scope of the VPLS instance. As such, VLAN translation can happen at each PE attached to the VPLS instance to translate between the global VLAN tag identifying the customer bridge-domain and the local VLAN tag used by the customer bridge-domain on this PE.

### 6. Security Considerations

This document does not introduce any additional security constraints.

## 7. IANA Considerations

Two new types field for the VLAN Vector TLV type and VLAN aware Bundling Capability TLV type are to be assigned by IANA from the LDP standard [RFC5036] "TLV type name space".

## 8 References

### 8.1 Normative References

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

### 8.2 Informative References

- [RFC 4762] Mark Lassere, et. al, "Virtual Private LAN Service (LAN) Using Label Distribution Protocol (LDP) Signaling", RFC4762, January 2007.
- [RFC 5036] Andersson, L., et al. "LDP Specification", RFC5036, October 2007.
- [RFC 4447] Martini. and et al., "Pseudowire Setup and Maintenance Using Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC 4448] Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.
- [EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-00.txt.
- [RFC5561] B.Thomas, K.Raza, S.Aggarwal, R.Agarwal, JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.
- [RFC-6136] Layer 2 Virtual Private Network (L2VPN) Operations, Administration, and Maintenance (OAM) Requirements and



Framework.

Authors' Addresses

Dennis Cai  
Cisco Systems

EMail: dcai@cisco.com

Sami Boutros  
Cisco Systems

EMail: sboutros@cisco.com

Samer Salam  
Cisco Systems

EMail: ssalam@cisco.com

Reshad Rahman  
Cisco Systems

EMail: rrahman@cisco.com

Network Working Group  
INTERNET-DRAFT  
Category: Standards Track

A. Sajassi  
Cisco

N. Bitar  
Verizon

R. Aggarwal  
Arktan

S. Boutros  
K. Patel  
S. Salam  
Cisco

W. Henderickx  
F. Balus  
Alcatel-Lucent

Aldrin Isaac  
Bloomberg

J. Drake  
R. Shekhar  
Juniper Networks

J. Uttaro  
AT&T

Expires: April 22, 2013

October 22, 2012

BGP MPLS Based Ethernet VPN  
draft-ietf-l2vpn-evpn-02

Status of this Memo

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

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

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

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

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

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Abstract

This document describes procedures for BGP MPLS based Ethernet VPNs (E-VPN).

## Table of Contents

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

9.3.1.2. P2MP MPLS LSPs . . . . .	17
9.3.1.3. MP2MP LSPs . . . . .	18
9.4 Aliasing . . . . .	18
9.4.1 Constructing the Ethernet A-D Route per EVI . . . . .	19
9.4.1.1 Ethernet A-D Route Targets . . . . .	20
9.5 Designated Forwarder Election . . . . .	20
9.5.1 Default DF Election Procedure . . . . .	22
9.5.2 DF Election with Service Carving . . . . .	22
10. Determining Reachability to Unicast MAC Addresses . . . . .	23
10.1. Local Learning . . . . .	23
10.2. Remote learning . . . . .	24
10.2.1. Constructing the BGP E-VPN MAC Address Advertisement . . . . .	24
11. ARP and ND . . . . .	26
11.1 Default Gateway . . . . .	26
12. Handling of Multi-Destination Traffic . . . . .	27
12.1. Construction of the Inclusive Multicast Ethernet Tag Route . . . . .	27
12.2. P-Tunnel Identification . . . . .	28
13. Processing of Unknown Unicast Packets . . . . .	29
13.1. Ingress Replication . . . . .	30
13.2. P2MP MPLS LSPs . . . . .	30
14. Forwarding Unicast Packets . . . . .	30
14.1. Forwarding packets received from a CE . . . . .	30
14.2. Forwarding packets received from a remote PE . . . . .	31
14.2.1. Unknown Unicast Forwarding . . . . .	31
14.2.2. Known Unicast Forwarding . . . . .	32
15. Load Balancing of Unicast Frames . . . . .	32
15.1. Load balancing of traffic from an PE to remote CEs . . . . .	32
15.1.1 Active-Standby Redundancy Mode . . . . .	32
15.1.2 All-Active Redundancy Mode . . . . .	33
15.2. Load balancing of traffic between an PE and a local CE . . . . .	35
15.2.1. Data plane learning . . . . .	35
15.2.2. Control plane learning . . . . .	35
16. MAC Mobility . . . . .	35
17. Multicast . . . . .	37
17.1. Ingress Replication . . . . .	37
17.2. P2MP LSPs . . . . .	37
17.3. MP2MP LSPs . . . . .	37
17.3.1. Inclusive Trees . . . . .	38
17.3.2. Selective Trees . . . . .	38
17.4. Explicit Tracking . . . . .	39
18. Convergence . . . . .	39
18.1. Transit Link and Node Failures between PEs . . . . .	39
18.2. PE Failures . . . . .	39
18.2.1. Local Repair . . . . .	40
18.3. PE to CE Network Failures . . . . .	40
19. LACP State Synchronization . . . . .	40
20. Acknowledgements . . . . .	41

21. Security Considerations . . . . .	42
22. IANA Considerations . . . . .	42
23. References . . . . .	42
23.1 Normative References . . . . .	42
23.2 Informative References . . . . .	42
24. Author's Address . . . . .	43

## 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

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

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

Ethernet segment identifier (ESI): If a CE is multi-homed to two or more PEs, the set of Ethernet links that attaches the CE to the PEs is an 'Ethernet segment'. Ethernet segments MUST have a unique non-zero identifier, the 'Ethernet Segment Identifier'.

Ethernet Tag: An Ethernet Tag identifies a particular broadcast domain, e.g., a VLAN. An E-VPN instance consists of one or more broadcast domains. Ethernet tag(s) are assigned to the broadcast domains of a given E-VPN instance by the provider of that E-VPN, and each PE in that E-VPN instance performs a mapping between broadcast

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

Link Aggregation Control Protocol (LACP):

Multipoint to Multipoint (MP2MP):

Point to Multipoint (P2MP):

Point to Point (P2P):

## 5. BGP MPLS Based E-VPN Overview

This section provides an overview of E-VPN.

An E-VPN comprises CEs that are connected to PEs that form the edge of the MPLS infrastructure. A CE may be a host, a router or a switch. The PEs provide virtual Layer 2 bridged connectivity between the CEs. There may be multiple E-VPNs in the provider's network.

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

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

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

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

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

## 6. Ethernet Segment

If a CE is multi-homed to two or more PEs, the set of Ethernet links constitutes an "Ethernet Segment". An Ethernet segment may appear to the CE as a Link Aggregation Group (LAG). Ethernet segments have an identifier, called the "Ethernet Segment Identifier" (ESI) which is encoded as a ten octets integer. A single-homed CE is considered to be attached to an Ethernet segment with ESI 0. Otherwise, an Ethernet segment MUST have a unique non-zero ESI. The ESI can be assigned using various mechanisms:

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

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

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



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

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

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

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

## 7. Ethernet Tag

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

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

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

The following subsections discuss the relationship between Ethernet Tags, EVIs and broadcast domain identifiers as well as the setting of the Ethernet Tag Identifier, in the various E-VPN BGP routes (defined in section 8), for the different types of service interfaces

described in [EVPN-REQ].

### 7.1 VLAN Based Service Interface

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

### 7.2 VLAN Bundle Service Interface

With this service interface, there is a many-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there is a single bridge domain per PE for the EVI. Different CEs connected to different PE ports MUST use the same broadcast domain identifiers (e.g. CE-VIDs) for the same EVI. The MPLS encapsulated frames MUST remain tagged with the originating CE's broadcast domain identifier (e.g. CE-VID). Tag translation is NOT permitted. The Ethernet Tag Identifier in all E-VPN routes MUST be set to 0.

#### 7.2.1 Port Based Service Interface

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

### 7.3 VLAN Aware Bundle Service Interface

With this service interface, there is a many-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there are multiple bridge domains per PE for the EVI: one broadcast domain per CE broadcast domain identifier. In the case where the CE broadcast domain identifiers are not consistent for different CEs, a normalized Ethernet Tag MUST be carried in the MPLS encapsulated frames and a tag translation function MUST be supported in the data path. This translation MUST be performed on both the imposition as well as the disposition PEs. The Ethernet Tag Identifier in all E-VPN routes MUST be set to the normalized Ethernet Tag assigned by the E-VPN provider.

### 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 TBD and an SAFI of E-VPN (To be assigned by IANA). The NLRI field in the MP\_REACH\_NLRI/MP\_UNREACH\_NLRI attribute contains the E-VPN NLRI (encoded as specified above).

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

### 8.1. Ethernet Auto-Discovery Route

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

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

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

### 8.2. MAC Advertisement Route

A MAC advertisement route type specific E-VPN NLRI consists of the following:

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

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

### 8.3. Inclusive Multicast Ethernet Tag Route

An Inclusive Multicast Ethernet Tag route type specific E-VPN NLRI

consists of the following:

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

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

#### 8.4 Ethernet Segment Route

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

RD (8 octets)
Ethernet Segment Identifier (10 octets)

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

#### 8.5 ESI MPLS Label Extended Community

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

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

0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
+-----																															

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 Active-Active mode; whereas, a value of 1 means that the multi-homed site is operating in Active-Standby 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 Extended Community

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

```

  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
  +-----+-----+-----+-----+-----+-----+-----+-----+
  | 0x44          | Sub-Type          | ES-Import          |
  +-----+-----+-----+-----+-----+-----+-----+-----+
  |                                     ES-Import Cont'd      |
  +-----+-----+-----+-----+-----+-----+-----+-----+

```

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

## 8.7 MAC Mobility Extended Community

This extended community is a new transitive extended community. It may be advertised along with MAC Advertisement routes. The procedures for using this Extended Community are described in section 16 "MAC Mobility".

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

```

  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
  +-----+-----+-----+-----+-----+-----+-----+-----+
  | 0x44          | Sub-Type          | Reserved=0        |
  +-----+-----+-----+-----+-----+-----+-----+-----+
  |                                     Sequence Number      |
  +-----+-----+-----+-----+-----+-----+-----+-----+

```

## 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 MPLS label used for split-horizon filtering (as discussed in section 9.2). Support of this route flavor is MANDATORY.

Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by 0. The reason for such encoding is that the RD cannot be that of a given EVI since the ESI can span across one or more EVIs.

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

The Ethernet Tag ID MUST be set to 0.

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

The "ESI MPLS Label Extended Community" MUST be included in the route. If all-active multi-homing is desired, then the "Active-



Standby" bit in the flags of the ESI MPLS Label Extended Community MUST be set to 0 and the MPLS label in that extended community MUST be set to a valid MPLS label value. The MPLS label in this Extended Community is referred to as an "ESI label". This label MUST be a downstream assigned MPLS label if the advertising PE is using ingress replication for receiving multicast, broadcast or unknown unicast traffic from other PEs. If the advertising PE is using P2MP MPLS LSPs for sending multicast, broadcast or unknown unicast traffic, then this label MUST be an upstream assigned MPLS label. The usage of this label is described in section 9.2.

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

#### 9.2.1.1. Ethernet A-D Route Targets

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

### 9.3 Split Horizon

Consider a CE that is multi-homed to two or more PEs on an Ethernet segment ES1. If the CE sends a multicast, broadcast or unknown unicast packet to a particular PE, say PE1, then PE1 will forward that packet to all or subset of the other PEs in the EVI. In this case the PEs, other than PE1, that the CE is multi-homed to MUST drop the packet and not forward back to the CE. This is referred to as "split horizon" filtering in this document.

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

#### 9.3.1 ESI MPLS Label Assignment

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

#### 9.3.1.1 Ingress Replication

An PE that is using ingress replication for sending broadcast, multicast or unknown unicast traffic, distributes to other PEs, that belong to the Ethernet segment, a downstream assigned "ESI MPLS label" in the Ethernet A-D route. This label MUST be programmed in the platform label space by the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for.

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

#### 9.3.1.2. P2MP MPLS LSPs

An PE that is using P2MP LSPs for sending broadcast, multicast or unknown unicast traffic, distributes to other PEs, that belong to the Ethernet segment or have an E-VPN in common with the Ethernet Segment, an upstream assigned "ESI MPLS label" in the Ethernet A-D route. This label is upstream assigned by the PE that advertises the route. This label MUST be programmed by the other PEs, that are connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for. This label MUST also be programmed by the other PEs, that import the route but are not connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must be a POP with no other associated action.

Consider PE1 and PE2 that are multi-homed to CE1 on ES1. Also consider PE3 that is in the same EVI as one of the EVIs to which ES1 belongs. Further, assume that PE1 is using P2MP MPLS LSPs to send broadcast, multicast or unknown unicast packets. When PE1 sends a multicast, broadcast or unknown unicast packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that it has assigned for the ESI that the packet was received on. The resulting packet is further encapsulated in the P2MP MPLS label stack necessary to transmit the packet to the other PEs. Penultimate hop popping MUST be disabled on the P2MP LSPs used in the MPLS transport infrastructure for E-VPN. When PE2 receives this packet, it de-capsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES11, then PE2 MUST NOT forward the packet onto ES11. When PE3 receives this packet, it de-capsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES11 and PE3 is not connected to ES11, then PE3 MUST pop the label and flood the packet over all local ESIs in the EVI.

#### 9.3.1.3. MP2MP LSPs

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

#### 9.4 Aliasing

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

To alleviate this issue, E-VPN introduces the concept of 'Aliasing'. Aliasing refers to the ability of an PE to signal that it has reachability to a given locally attached Ethernet segment, even when it has learnt no MAC addresses from that segment. The Ethernet A-D route per EVI is used to that end. Remote PEs which receive MAC

advertisement routes with non-zero ESI SHOULD consider the advertised MAC address as reachable via all PEs which have advertised reachability to the relevant Segment using Ethernet A-D routes with the same ESI (and Ethernet Tag if applicable).

This flavor of Ethernet A-D route 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. Therefore, if Ether A-D route associated with aliasing arrives ahead of the route associated with mass-withdraw, then former must NOT be processed into the FIB till the latter arrives. This will take care of corner cases and race conditions where the Ether A-D route associated with mass-withdraw is withdrawn but a PE still receives the route associated with aliasing routes.

#### 9.4.1 Constructing the Ethernet A-D Route per EVI

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

Route-Distinguisher (RD) MUST be set to the RD of the EVI that is advertising the NLRI. An RD MUST be assigned for a given EVI on an PE. This RD MUST be unique across all EVIs on an PE. It is RECOMMENDED to use the Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by a number unique to the PE. This number may be generated by the PE. Or in the Unique VLAN E-VPN case, the low order 12 bits may be the 12 bit VLAN ID, with the remaining high order 4 bits set to 0.

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

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

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

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

disposition.

- + One Ethernet A-D route per <ESI, EVI> (where the Ethernet Tag ID is set to 0). This is applicable when the PE uses MAC-based disposition, or when the PE uses MPLS-based disposition when no VLAN translation is required.

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

The Next Hop field of the MP\_REACH\_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

#### 9.4.1.1 Ethernet A-D Route Targets

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

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

##### 9.4.1.1.1 Auto-Derivation from the Ethernet Tag ID

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

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

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

#### 9.5 Designated Forwarder Election

Consider a CE that is a host or a router that is multi-homed directly to more than one PE in an E-VPN on a given Ethernet segment. One or

more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

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

Note that this behavior, which allows selecting a DF at the granularity of <ESI, EVI> for multicast, broadcast and unknown unicast traffic, is the default behavior in this specification. Optional mechanisms, which will be specified in the future, will allow selecting a DF at the granularity of <ESI, EVI, S, G>.

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

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

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

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

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

single PE MUST be elected as the DF for a particular EVI on that Ethernet segment.

The detailed procedures for DF election are described next.

#### 9.5.1 Default DF Election Procedure

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

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

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

#### 9.5.2 DF Election with Service Carving

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

1. When a PE discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute.
2. The PE then starts a timer to allow the reception of Ethernet Segment routes from other PE nodes connected to the same Ethernet Segment.
3. When the timer expires, each PE builds an ordered list of the IP addresses of all the PE nodes connected to the Ethernet Segment (including itself), in increasing numeric value. Every PE is then given an ordinal indicating its position in the ordered list, starting with 0 as the ordinal for the PE with the numerically lowest

IP address. The ordinals are used to determine which PE node will be the DF for a given EVI on the Ethernet Segment using the following rule: Assuming a redundancy group of N PE nodes, the PE with ordinal i is the DF for EVI V when  $(V \bmod N) = i$ .

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

4. The PE that is elected as a DF for a given EVI will unblock traffic for that EVI only if the EVI is configured/enabled on the Segment. Note that the DF PE unblocks multi-destination traffic in the egress direction towards the Segment. All non-DF PEs continue to drop multi-destination traffic (for the associated EVIs) in the egress direction towards the Segment.

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

#### 10. Determining Reachability to Unicast MAC Addresses

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

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

##### 10.1. Local Learning

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

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

- DHCP requests
- ARP request for its own MAC.



- ARP request for a peer.

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

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

## 10.2. Remote learning

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

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

### 10.2.1. Constructing the BGP E-VPN MAC Address Advertisement

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

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

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

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

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

the case if the CE Ethernet tags, e.g., VLAN ID, for a particular bridge domain are different on different CEs.

The MAC address length field is typically set to 48. However this specification enables specifying the MAC address as a prefix; in which case, the MAC address length field is set to the length of the prefix. This provides the ability to aggregate MAC addresses if the deployment environment supports that. The encoding of a MAC address MUST be the 6-octet MAC address specified by [802.1D-ORIG] [802.1D-REV]. If the MAC address is advertised as a prefix then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as 6 octets.

The IP Address Length field value is set to the number of octets in the IP Address field.

The IP Address field is optional. By default, the IP Address Length field is set to 0 and the IP address field is omitted from the route. When a valid IP address is included, it is encoded as specified in section 12.

The MPLS label field carries one or more labels (that corresponds to the stack of labels [MPLS-ENCAPS]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [MPLS-ENCAPS]). The MPLS label stack MUST be the downstream assigned E-VPN MPLS label stack that is used by the PE to forward MPLS-encapsulated Ethernet frames received from remote PEs, where the destination MAC address in the Ethernet frame is the MAC address advertised in the above NLRI. The forwarding procedures are specified in section "Forwarding Unicast Packets" and "Load Balancing of Unicast Packets".

An PE may advertise the same single E-VPN label for all MAC addresses in a given EVI. This label assignment methodology is referred to as a per EVI label assignment. Alternatively, an PE may advertise a unique E-VPN label per <ESI, Ethernet Tag> combination. This label assignment methodology is referred to as a per <ESI, Ethernet Tag> label assignment. As a third option, an PE may advertise a unique E-VPN label per MAC address. All of these methodologies have their tradeoffs.

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

ID translation on egress to the CE.

The Next Hop field of the MP\_REACH\_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

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

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

## 11. ARP and ND

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

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

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

### 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 Active-Standby Redundancy Mode

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

These are referred to as the backup PEs.

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

#### 15.1.2 All-Active Redundancy Mode

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

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

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

The following example explains the above.

Consider a CE (CE1) that is dual-homed to two PEs (PE1 and PE2) on a LAG interface (ES1), and is sending packets with MAC address MAC1 on

VLAN1. A remote PE, say PE3, is able to learn that MAC1 is reachable via PE1 and PE2. Both PE1 and PE2 may advertise MAC1 in BGP if they receive packets with MAC1 from CE1. If this is not the case, and if MAC1 is advertised only by PE1, PE3 still considers MAC1 as reachable via both PE1 and PE2 as both PE1 and PE2 advertise a Ethernet A-D route per ESI for ESI1 as well as an Ethernet A-D route per EVI for <ESI1, VLAN1>.

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

The MPLS label stack to send packets to PE2 is the MPLS LSP stack to get to PE2 and the MPLS label in the Ethernet A-D route advertised by PE2 for <ESI1, VLAN1>, if PE2 has not advertised MAC1 in BGP.

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

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

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

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

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

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

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

### 15.2.1. Data plane learning

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

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

### 15.2.2. Control plane learning

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

## 16. MAC Mobility

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

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

If local learning is performed using the data plane, these PEs will

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

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

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

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

- A PE advertising a MAC address for the first time advertises it with no MAC Mobility extended community attribute.
- A PE detecting a locally attached MAC address for which it had previously received a MAC Advertisement route with a different Ethernet segment identifier advertises the MAC address in a MAC Advertisement route tagged with a MAC Mobility extended community attribute with a sequence number one greater than the sequence number in the MAC mobility attribute of the received MAC Advertisement route. In the case of the first mobility event for a given MAC address, where the received MAC Advertisement route does not carry a MAC Mobility attribute, the value of the sequence number in the received route is assumed to be 0 for purpose of this processing.
- A PE detecting a locally attached MAC address for which it had previously received a MAC Advertisement route with the same Ethernet segment identifier advertises it with:
  - i. no MAC Mobility extended community attribute, if the received route did not carry said attribute.
  - ii. a MAC Mobility extended community attribute with the sequence number equal to the sequence number in the received MAC Advertisement route, if the received route is tagged with a MAC Mobility extended community attribute.

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

## 17. Multicast

The PEs in a particular E-VPN may use ingress replication or P2MP LSPs to send multicast traffic to other PEs.

### 17.1. Ingress Replication

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

### 17.2. P2MP LSPs

An PE may use an "Inclusive" tree for sending an unknown unicast, broadcast or multicast packet or a "Selective" tree. This terminology is borrowed from [VPLS-MCAST].

A variety of transport technologies may be used in the SP network. For inclusive P-Multicast trees, these transport technologies include point-to-multipoint LSPs created by RSVP-TE or mLDP. For selective P-Multicast trees, only unicast PE-PE tunnels (using MPLS or IP/GRE encapsulation) and P2MP LSPs are supported, and the supported P2MP LSP signaling protocols are RSVP-TE, and mLDP.

### 17.3. MP2MP LSPs

The root of the MP2MP LDP LSP advertises the Inclusive Multicast Tag route with the PMSI Tunnel attribute set to the MP2MP Tunnel identifier. This advertisement is then sent to all PEs in the E-VPN. Upon receiving the Inclusive Multicast Tag routes with a PMSI Tunnel attribute that contains the MP2MP Tunnel identifier, the receiving PEs initiate the setup of the MP2MP tunnel towards the root using the procedures in [MLDP].

### 17.3.1. Inclusive Trees

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

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

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

### 17.3.2. Selective Trees

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

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

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

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

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

#### 17.4. Explicit Tracking

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

#### 18. Convergence

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

##### 18.1. Transit Link and Node Failures between PEs

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

##### 18.2. PE Failures

Consider a host host1 that is dual homed to PE1 and PE2. If PE1 fails, a remote PE, PE3, can discover this based on the failure of the BGP session. This failure detection can be in the sub-second



range if BFD is used to detect BGP session failure. PE3 can update its forwarding state to start sending all traffic for host1 to only PE2. It is to be noted that this failure recovery is potentially faster than what would be possible if data plane learning were to be used. As in that case PE3 would have to rely on re-learning of MAC addresses via PE2.

#### 18.2.1. Local Repair

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

#### 18.3. PE to CE Network Failures

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

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

#### 19. LACP State Synchronization

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

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

This includes at least the following LACP specific configuration parameters:

- System Identifier (MAC Address): uniquely identifies a LACP speaker.
- System Priority: determines which LACP speaker's port priorities are used in the Selection logic.
- Aggregator Identifier: uniquely identifies a bundle within

- a LACP speaker.
- Aggregator MAC Address: identifies the MAC address of the bundle.
- Aggregator Key: used to determine which ports can join an Aggregator.
- Port Number: uniquely identifies an interface within a LACP speaker.
- Port Key: determines the set of ports that can be bundled.
- Port Priority: determines a port's precedence level to join a bundle in case the number of eligible ports exceeds the maximum number of links allowed in a bundle.

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

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

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

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

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

## 20. Acknowledgements

We would like to thank Yakov Rekhter, Pedro Marques, Kaushik Ghosh, Nischal Sheth, Robert Raszuk, Amit Shukla and Nadeem Mohammed for discussions that helped shape this document. We would also like to thank Han Nguyen for his comments and support of this work. We would also like to thank Steve Kensil for his review.

## 21. 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

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.



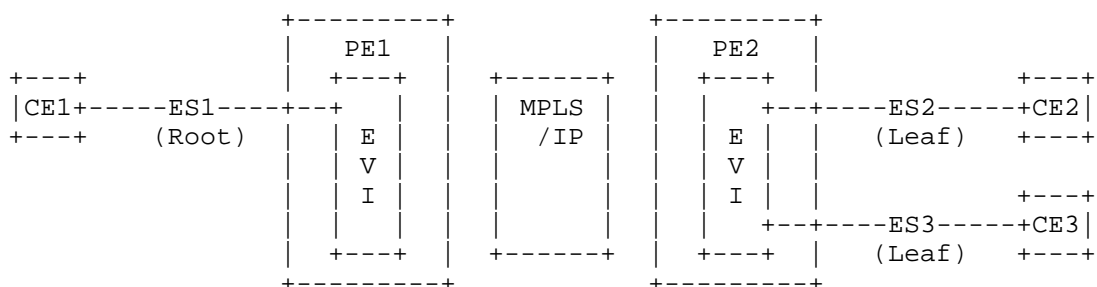


Figure 1: Scenario 1

## 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.

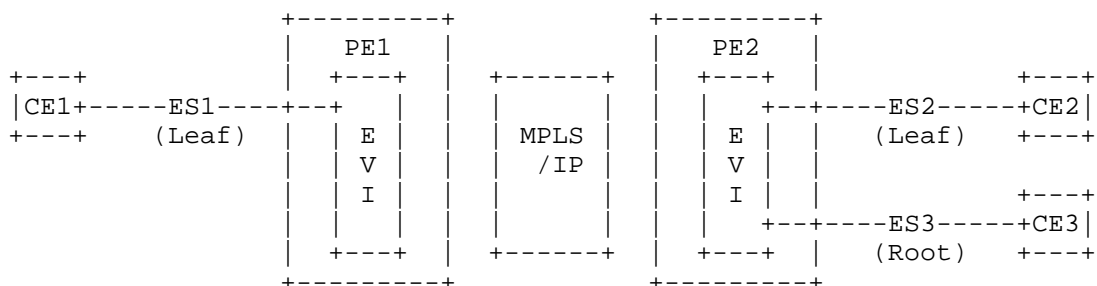


Figure 2: Scenario 2

## 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.

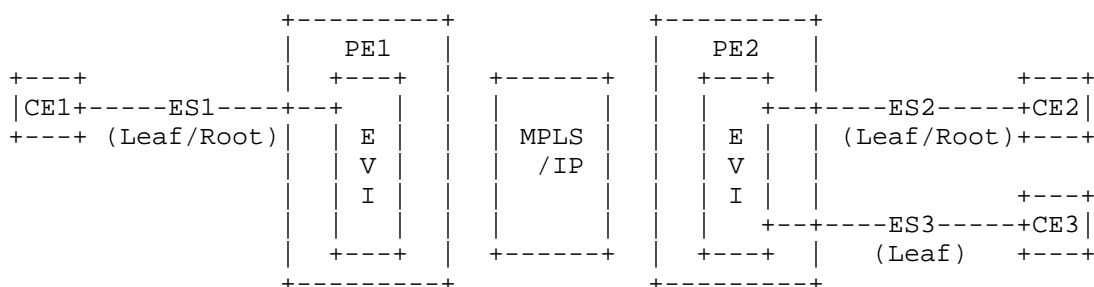


Figure 3: Scenario 3

### 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](mailto:ssalam@cisco.com)

Wim Henderickx  
Alcatel-Lucent  
Email: [wim.henderickx@alcatel-lucent.com](mailto:wim.henderickx@alcatel-lucent.com)

Jim Uttaro  
AT&T  
Email: [jul738@att.com](mailto:jul738@att.com)

L2VPN Workgroup  
INTERNET-DRAFT  
Intended Status: Standards Track

Ali Sajassi  
Samer Salam  
Keyur Patel  
Cisco

Wim Henderickx  
Alcatel-Lucent

Nabil Bitar  
Verizon

John Drake  
Yakov Rakhter  
Juniper

Aldrin Isaac  
Bloomberg

Jim Uttaro  
AT&T

Expires: April 22, 2012

October 22, 2012

E-VPN Seamless Interoperability with IP-VPN  
draft-sajassi-l2vpn-evpn-ipvpn-interop-01

## Abstract

E-VPN can be an integral part of an Integrated Routing and Bridging (IRB) solution which is capable of performing optimum unicast and multicast forwarding not just for L2 traffic but also for L3 traffic. This document describes how an IRB solution based on E-VPN can interoperate seamlessly with the IP-VPN solution over MPLS and IP networks.

## 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	Shortcomings of L2-Only Solution . . . . .	3
1.2	Shortcomings of L3-Only Solution . . . . .	4
1.3	Combined L2 & L3 Solution: IRB . . . . .	6
1.1	Terminology . . . . .	6
2	Seamless Interoperability with IP-VPN PEs . . . . .	6
2.1	Interoperability Use-Cases . . . . .	6
2.1.1	IP-VPN Clients Access to Cloud Services . . . . .	7
2.1.2	Communication with IP-VPN NVEs . . . . .	7
2.1.3	Communication with IP-VPN GWs . . . . .	8
2.2	Characteristics of Seamless Interoperability . . . . .	8
3	An IRB Solution Based on E-VPN . . . . .	9
3.1	E-VPN PE Model for Seamless Interoperability . . . . .	9
3.2	IP-VPN BGP support on E-VPN PEs . . . . .	11
3.3	Handling Multi-Destination Traffic: . . . . .	12
3.2.1	Further optimization on RR . . . . .	12
5	Acknowledgement . . . . .	12
6	Security Considerations . . . . .	12
7	IANA Considerations . . . . .	13
8	References . . . . .	13
8.1	Normative References . . . . .	13
8.2	Informative References . . . . .	13
	Authors' Addresses . . . . .	13



## 1 Introduction

E-VPN can be an integral part of an Integrated Routing and Bridging (IRB) solution which is capable of performing optimum unicast and multicast forwarding not just for L2 traffic (intra-subnet forwarding), as described in the baseline draft [E-VPN], but also is capable of performing optimum unicast and multicast forwarding for L3 traffic (inter-subnet forwarding) as described in [DC-MOBILITY].

Such IRB capability is of high relevance in data center applications where performing either L2 or L3 forwarding alone may not be sufficient.

### 1.1 Shortcomings of L2-Only Solution

Figure-1 depicts a Data Center Network (DCN) using IP overlay where the PE functionality (and IP tunnel encapsulation) are either residing on physical Top of Rack (ToR) switches or on virtual hypervisor-based switches. In this document, we refer to these PE devices (either physical or virtual) that provide IP overlay tunneling as Network Virtualized Endpoints (NVEs). The DCN is connected to the Internet and/or enterprise/SP MPLS/IP core network via gateway (GW) nodes.

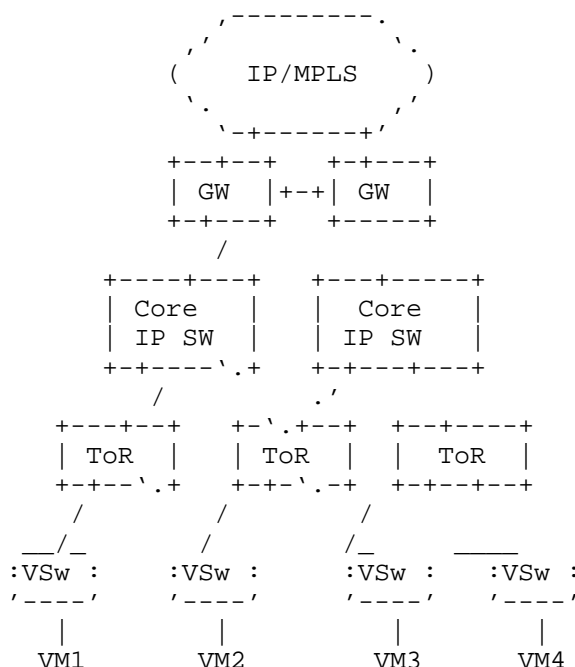


Figure 1: A typical DC network

If Network Virtualization Endpoints (NVEs) were only to provide L2 service (and forwarding), then for two VMs on two different subnets, which need to communicate with each other, their packets need to be forwarded to a router (either physical or virtual). In the above diagram, the packets from the VMs need to be forwarded all the way to one of the GW devices to perform L3 forwarding. This is generally sub-optimal because the two VMs may be connected to the same virtual switch or the same TOR where L3 switching could have been performed locally. Even if the two VMs are located in different PODs within the same DC, and the traffic between the two VMs requires transitioning a core switch, adding a GW for L3 switching adds additional hops to the data path. However, if an NVE has IRB capability, then it can perform optimum L2 forwarding for intra-subnet traffic and optimum L3 forwarding for inter-subnet traffic, delivering optimum forwarding of unicast and multicast packets at all time.

## 1.2 Shortcomings of L3-Only Solution

Consider the scenario where a server is multi-homed to several ToR devices using an Ethernet Link Aggregation Group with LACP [802.1AX]

and the VMs are connected to a virtual bridge on the server - i.e., there is an Ethernet bridge on the data path between the VMs and the TORs. The TORs are acting as NVEs. In this scenario, the LAG spans across multiple PE devices (NVEs) and IGMP joins for the same multicast group can arrive at both PEs. As such, DF election and split-horizon filtering functions are required on the TORs belonging to the same LAG in order to avoid loops and packet duplication. However, the existing IP-VPN solution does not provide such capabilities that are available in the E-VPN solution. Therefore, these TOR devices cannot be simple L3VPN PEs.

Assuming that the above shortcoming is addressed by adding DF election and split-horizon filtering to IP-VPN, several other issues will continue to exist with L3-only solution, particularly when attempting to rely on L3 forwarding for intra-subnet traffic:

1. With L3 forwarding, in the absence of a default route, unknown IP destination addresses are dropped. Furthermore, an IP default route directs a particular traffic flow to a single next-hop or outbound interface. This means that L3 forwarding cannot support the forwarding semantics of a subnet broadcast.
2. With L3 forwarding, the MAC header is link-local and MAC addresses are swapped on a hop-by-hop basis. This means that if an NVE resorts to L3 forwarding of intra-subnet traffic, then all hosts within the same subnet will receive traffic with the source MAC addresses set to the NVE's address(es) instead of the originating hosts' MAC addresses. As a result, any higher layer application which relies on the source MAC address for identifying the communicating endpoint will break, as it will no longer be able to tell apart the hosts within the subnet based on their MAC addresses. This essentially creates an address aliasing problem. A related issue, that results from the MAC address being rewritten by the NVE, is that the hosts can no longer perform duplicate MAC address detection.
3. With L3 forwarding, the IP TTL is decremented with every routed hop. Some applications rely on this fundamental behavior to confine traffic to the originating subnet, by setting the TTL to 1 on transmission. Such applications will no longer work when intra-subnet traffic is L3 forwarded.
4. IPv6 link-local addressing and duplicate address detection [RFC4862] assumes and relies upon L2 connectivity within the subnet. These mechanisms will break if the NVE performs L3 intra-subnet forwarding.
5. Finally last but not least, there are non-IP applications that require L2 forwarding or there are applications that rely on end host

MAC addresses.

### 1.3 Combined L2 & L3 Solution: IRB

An IRB solution based on E-VPN can address the shortcomings of L2-only as well as L3-only solutions, and provide optimum forwarding for both inter and intra subnet switching, not only within a DCN but across different DCNs. This E-VPN based solution fits well for DCN overlay and DCI applications, but typical deployments will include IP-VPN PE's that E-VPN PE's need to inter-operate with, such as:

- 1) IP-VPN client sites accessing cloud services
- 2) Communication with IP-VPN ToRs/VSw
- 3) Communication with IP-VPN GWs

Therefore, interoperability with IP-VPN PE's is of paramount importance.

### 1.1 Terminology

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

## 2 Seamless Interoperability with IP-VPN PE's

### 2.1 Interoperability Use-Cases

There are three use-cases that require interoperability between E-VPN and IP-VPN. Those are discussed next.

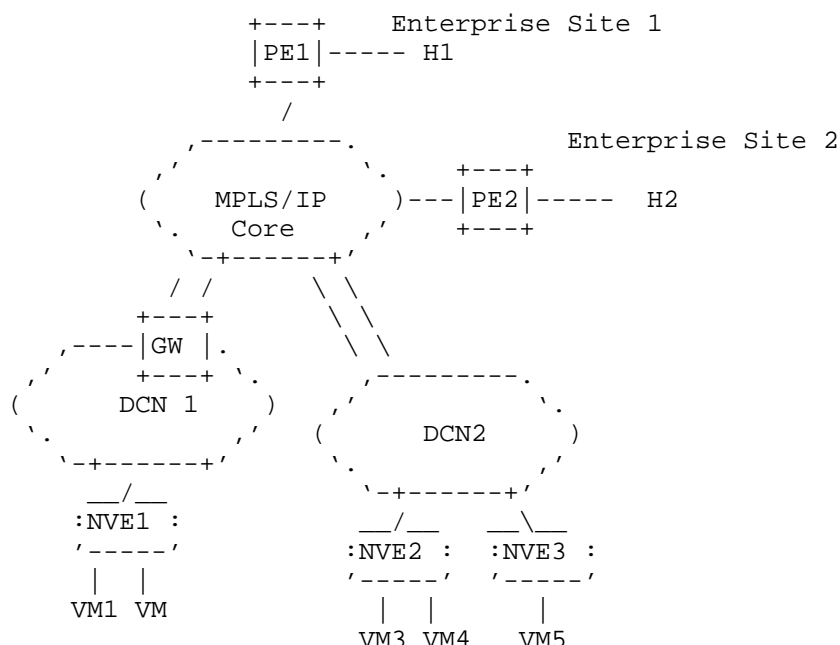


Figure 2: Interoperability Use-Cases

### 2.1.1 IP-VPN Clients Access to Cloud Services

An SP offering IP-VPN services to an enterprise may wish to expand its service offering to include Cloud services, while leveraging its existing MPLS/IP infrastructure. The SP may deploy E-VPN on the NVE in order to support L2 connectivity between VMs. If the number of VPNs and routes per VPN is not high, the SP may extend the L3 edge to the NVE and implement that function on the ToR switch. An alternative would be to have the NVE be on the hypervisor, in higher scale scenarios. Either way, distributing the L3 edge to the NVE renders it possible to avoid having an IP-VPN GW for the DCN. For this scenario, interoperability between the E-VPN NVE and IP-VPN PE is required in order to enable the new service offering.

For e.g., consider Figure 1 where an IP-VPN service is being offered between Enterprise sites 1 and 2. PE1 and PE2 act as IP-VPN PEs. Furthermore, assume that DCN2 employ E-VPN (i.e. NVE2 and NVE3 are E-VPN PEs). For the SP to offer Cloud service, interoperability between the IP-VPN PEs and E-VPN NVEs is required.

### 2.1.2 Communication with IP-VPN NVEs

In certain deployments, where only L3 connectivity is required by

certain hosts (e.g. VMs), the NVEs associated with those hosts may employ IP-VPN functionality only. An example of this would be running the IP-VPN PE functionality on the hypervisor using the mechanisms of [L3VPN-ENDSYSTEM]. Other VMs may require both L2 as well as L3 connectivity. The NVEs associated with those latter VMs would employ E-VPN. In order to allow for inter subnet communication between both categories of VMs (i.e. those which require L3 connectivity only and those requiring both L2 as well as L3 connectivity), interoperability is required between the IP-VPN and the E-VPN NVEs.

To illustrate this with an example, consider the network of Figure 1. VM5 requires L3 connectivity only, and subsequently NVE3 employs IP-VPN PE functionality solely. VM3 requires both L2 and L3 connectivity, hence, NVE2 is employing E-VPN PE functionality. For VM3 to be able to optimally communicate with VM5, seamless interoperability between IP-VPN and E-VPN is required.

### 2.1.3 Communication with IP-VPN GWs

The DCN may include an IP-VPN GW in order to confine the routing tables of the NVEs to L3 routes that are local to the DCN. The NVEs, in this case, would have default routes pointing to the GW. When the NVEs need to provide L2 as well as L3 connectivity to the associated VMs, they must run E-VPN PE functionality. In order for the IP-VPN GW to learn reachability to the VMs local to the DCN, interoperability is required between E-VPN NVEs and the IP-VPN GW.

As an example, consider the network of Figure 1 where the GW of DCN1 is an IP-VPN gateway. If NVE1 employs E-VPN PE functionality, then interoperability between E-VPN and IP-VPN is required for connectivity between NVE1 and the GW.

## 2.2 Characteristics of Seamless Interoperability

Seamless interoperability between E-VPN and IP-VPN must meet the following characteristics:

- Be completely transparent to the operation of the IP-VPN PE. In other words, the IP-VPN PE would not even be aware that it is communicating with an E-VPN endpoint. As such, no upgrade to the IP-VPN nodes is required, not even a software upgrade.
- Be optimal from data-plane forwarding perspective. This means that a gateway function is not required in order to normalize the encapsulation to Ethernet in order to support the interoperability. To elaborate on this: it is always possible to have an E-VPN PE interoperate with an IP-VPN PE using a normalized Ethernet L2 hand-off between the two. This however, requires that the MPLS

encapsulation be terminated on each PE, with the added overhead of unnecessarily performing MPLS imposition and disposition on both PEs. A side-effect of this gateway approach is that the host MAC addresses will be visible to the E-VPN, and this may create scalability bottlenecks, especially in virtualized data center environments because of sheer number of host MAC addresses.

### 3 An IRB Solution Based on E-VPN

An IRB solution based on E-VPN can meet data center network requirements in terms of:

- Providing optimal forwarding for intra-subnet (L2) traffic.
- Providing optimal forwarding for inter-subnet (L3) traffic, by avoiding the need for a centralized L3 GW. This is because the E-VPN MAC Advertisement route can carry an IP address in addition to the MAC address.
- Support for light-weight multicast using ingress replication, in cases where multicast applications are not required or dominant.
- Support for optimal multicast delivery through P2MP tunnels, when required, to optimize DCN resources.
- Support for multi-homing with active/active redundancy and per-flow load-balancing using multi-chassis LAG.
- Support for network-based as well as host-based overlay models.
- Support for consistent policy-based forwarding for both L2 and L3 forwarded traffic.

#### 3.1 E-VPN PE Model for Seamless Interoperability

This section describes the PE data-plane model required to achieve seamless interoperability.

The E-VPN PE establishes a many-to-one mapping between EVIs and a VRF. For a given EVI, it is possible to have multiple associated bridge-domains using the VLAN-aware bundling service interface, as defined in [EVPN-REQ]. Each bridge-domain maps to a unique IP subnet within a VRF context. The following figure depicts the model where there are N VRFs corresponding to N tenants, with each tenant having 2 EVIs and up to M subnets (bridge domains) per EVI.

Note that this PE model provides flexibility for a wide gamut of deployment options. For example, one end of the spectrum would be

with a single EVI per tenant being mapped to a single VRF. Each EVI hosts multiple bridge-domains (one bridge-domain per subnet). This model allows for L2 traffic segregation between different subnets in addition to L3 connectivity among those subnets, as long as global Service VLANs are assigned per tenant (this uses VLAN-aware bundling service in E-VPN). The other end of the spectrum is with multiple EVIs per tenant all mapped to a single VRF. Each EVI hosts a single bridge-domain in this latter case. This model allows for L2 traffic segregation between subnets in addition to L3 connectivity among those subnets without the need for globally assigned Service VLANs (this uses VLAN-based service in E-VPN).

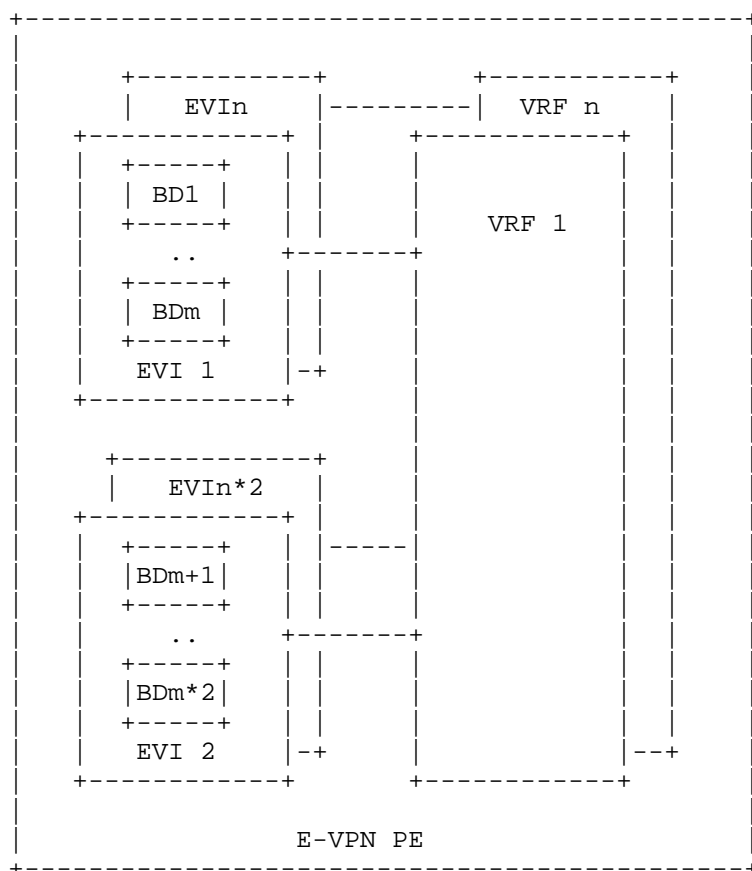


Figure 3: E-VPN PE Model for Seamless Interoperability with IP-VPN

One way to visualize this model is to consider a bridged virtual interface (BVI) to be associated with every bridge-domain in a given EVI. The BVI is an L3 routed interface (hence terminates L2). All the



BVIs associated with a given EVI are placed in the same VRF.

The IP forwarding table in a given VRF is shared in between E-VPN and IP-VPN. When an E-VPN MAC advertisement route is received by the PE, the MAC address associated with the route is used to populate the bridge-domain MAC table, whereas the IP address associated with the route is used to populate the corresponding VRF. For intra-subnet forwarding, the PE consults the bridge-domain MAC table whereas for inter-subnet forwarding the PE performs the lookup in the associated VRF.

When an E-VPN packet is received by a PE, it decapsulates the MPLS header and then performs a lookup on the destination MAC address. If the MAC address corresponds to one of its BVI interfaces, the PE deduces that the packet must be inter-subnet routed. Hence, the PE performs an IP lookup in the associated VRF table. However, if the destination MAC address does not correspond to a BVI, then the PE concludes that this packet needs to be intra-subnet switched, and no further IP lookup is needed.

### 3.2 IP-VPN BGP support on E-VPN PEs

The E-VPN PE learns host (e.g. VM) MAC addresses via normal bridge learning, and host IP addresses either via snooping of control traffic (e.g. ARP, DHCP...) or gleaning of data traffic. Once the PE learns a new MAC/IP address tuple, it advertises two routes for that tuple:

- An E-VPN MAC Advertisement route using the E-VPN AFI/SAFI and associated NLRI, which is used to advertise reachability to other remote E-VPN nodes. The MAC route advertises both the IP and MAC addresses of the host.
- An IP-VPN route using IP-VPN AFI/SAFI and associated NLRIs, which is used to advertise reachability to remote IP-VPN speakers. The IP-VPN route advertises only the IP address of the end-station.

Given that on the E-VPN PEs there is a one-to-one mapping between an E-VPN Instance (EVI) and a VRF, the same BGP RT and RD are used for both E-VPN and IP-VPN routes. Received E-VPN routes carry both IP and MAC addresses. The MAC addresses are injected into BD tables whereas the IP addresses are injected into VRFs. When an E-VPN speaker receives an IP-VPN route from a remote IP-VPN speaker, it installs the associated IP address in the appropriate VRF. It should be noted that when a MAC address is installed in the EVI, it is only installed in a single BD associated with the subnet corresponding to the Ethernet Tag encoded in the E-VPN MAC route.

If, for a given tenant, the IP-VPN PEs only need to share IP-VPN routes for a subset of the subnets with their E-VPN PEs counterparts, then one RT is used as a common RT between IP-VPN and E-VPN PEs for the common subnets and a different one or more RTs are used by the E-VPN PEs for the other tenant subnets that don't need to share routes with the IP-VPN PEs. If further topology constraint is needed among E-VPN and IP-VPN PEs, then instead of a common RT, one can use additional RTs to satisfy the topology constraint.

### 3.3 Handling Multi-Destination Traffic:

A key issue is how to handle multi-destination traffic, since E-VPN uses an MPLS label for split-horizon, and the equivalent does not exist in IP-VPN. This can be solved in two different ways, depending on whether the network uses LSM or Ingress Replication:

For LSM, two different sets of P2MP multicast trees can be used by the E-VPN PEs. One tree set encompasses only the IP-VPN endpoints whereas the second set includes only the E-VPN speakers. When an E-VPN PE receives a multi-destination frame, it sends a copy on each of the two trees associated with a given EVI/VRF. When the PE sends traffic on the IP-VPN tree, it does not include the split-horizon label since the IP-VPN endpoints do not understand this label. Note that this does not create any adverse side-effects because an E-VPN PE and an IP-VPN will never be combined in the same Redundancy Group (i.e. will never be multi-homed to the same Ethernet Segment), and as such the split-horizon filtering is never required on the IP-VPN PEs.

For ingress replication, the E-VPN PE sends the right label stack depending on the capability of the receiving (i.e. egress) PE. When replicating to IP-VPN endpoints, the ingress PE simply does not include any split-horizon labels.

#### 3.2.1 Further optimization on RR

It is possible to optimize the number of routes that are advertised by a given E-VPN speaker for a specific host address, by leveraging extra intelligence on the BGP route reflector. A future version of this document will describe the detailed procedures to achieve this.

## 5 Acknowledgement

## 6 Security Considerations

## 7 IANA Considerations

## 8 References

### 8.1 Normative References

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

### 8.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-12vpn-evpn-00.txt, work in progress, February, 2012.

[DC-MOBILITY] Aggarwal et al., "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-03.txt, work in progress, June, 2012.

## Authors' Addresses

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

Samer Salam  
Cisco  
595 Burrard Street  
Vancouver, BC V7X 1J1, Canada  
Email: [ssalam@cisco.com](mailto:ssalam@cisco.com)

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

Nabil Bitar  
Verizon Communications  
Email : [nabil.n.bitar@verizon.com](mailto:nabil.n.bitar@verizon.com)

Aldrin Isaac  
Bloomberg  
aldrin.isaac@gmail.com

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

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

Yakov Rekhter  
Juniper Networks  
Email: yakov@juniper.net

NVO3 Workgroup  
INTERNET-DRAFT  
Intended Status: Standards Track

Ali Sajassi  
Samer Salam  
Keyur Patel  
Cisco

Nabil Bitar  
Verizon

Wim Henderickx  
Alcatel-Lucent

Expires: April 22, 2013

October 22, 2012

A Network Virtualization Overlay Solution using E-VPN  
draft-sajassi-nvo3-evpn-overlay-01

#### Abstract

This document describes how E-VPN can be used as an NVO solution and explores the various tunnel encapsulation options and their impact on the E-VPN control-plane and procedures. In particular, the following three 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 . . . . .	4
2	E-VPN Main Features . . . . .	5
2.1	Multi-homed Ethernet Segment Auto-Discovery . . . . .	5
2.2	Fast Convergence and Mass Withdraw . . . . .	5
2.3	Split-Horizon . . . . .	5
2.4	Aliasing . . . . .	6
2.5	DF Election . . . . .	6
3	Encapsulation Options for E-VPN Overlays . . . . .	7
3.1	MPLS over GRE . . . . .	7
3.1.1	Benefits of MPLS over GRE . . . . .	7
3.2	VXLAN/NVGRE Encapsulation . . . . .	8
3.2.1	Impact on E-VPN Routes for VXLAN/NVGRE Encapsulation . . . . .	8
3.2.2	Impact on E-VPN Procedures for VXLAN/NVGRE Encapsulation . . . . .	9
3.2.2.1	NVE with No Redundancy . . . . .	9
3.2.2.2	NVE with Active/Standby Redundancy . . . . .	10
3.2.2.3	NVE with All-Active Redundancy . . . . .	10
3.2.3	Support for Multicast . . . . .	13
3.2.4	Inter-AS Challenges . . . . .	13
4	Comparison between MPLSoGRE and VXLAN/NVGRE Encapsulation . . . . .	14
5	Acknowledgement . . . . .	15
6	Security Considerations . . . . .	15
7	IANA Considerations . . . . .	15
8	References . . . . .	15
8.1	Normative References . . . . .	15
8.2	Informative References . . . . .	15
	Authors' Addresses . . . . .	16



## 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 (i.e. Virtual Machines or VMs). The key requirements of such a solution as described in [Problem-Statement] are:

- Isolation of network traffic per tenant
- Support of 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

The underlay network for NVO solutions is assumed to provide IP connectivity.

This document describes how E-VPN can be used as an NVO solution and explores the various tunnel encapsulation options for E-VPN over IP, and their impact on the E-VPN control-plane and procedures. 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. This may not be desirable, for e.g., when the NVE resides on the hypervisor. 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:

- MPLS over GRE
- VXLAN and NVGRE

Before getting into the description of the different encapsulation options for E-VPN over IP, it is important to highlight the E-VPN solution main features, how those features are currently supported, and any impact that the encapsulation may have 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].



## 2 E-VPN Main Features

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

### 2.1 Multi-homed Ethernet Segment Auto-Discovery

E-VPN NV Edge devices (NVEs) connected to the same Ethernet segment (e.g. server) can automatically discover each other with minimal to no configuration through the exchange of BGP routes.

### 2.2 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. 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).

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

frame is allowed to egress a specific Ethernet segment.

## 2.4 Aliasing

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 PEs are connected to the multi-homed segment. As a result, the remote PEs 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 PEs 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 PE. Another scenario where this occurs is when the PEs rely on control plane learning on the access (e.g. using ARP), since ARP traffic will be hashed to a single link in the LAG.

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

## 2.5 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.

### 3 Encapsulation Options for E-VPN Overlays

#### 3.1 MPLS over GRE

The E-VPN data-plane is modeled as an E-VPN MPLS client layer sitting over an MPLS PSN tunnel. The Split-Horizon and Aliasing functions of E-VPN are tied to the MPLS client layer. In order to keep the E-VPN procedures intact and data-plane operation as is, an ideal encapsulation would allow the E-VPN MPLS client layer to be carried over an IP PSN tunnel transparently - i.e., without any changes. The existing standards-based GRE encapsulation as defined by [RFC2890] and [RFC2784] provides such a solution:

```

 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
+-----+-----+-----+-----+-----+-----+-----+-----+
|C| |K|S| Reserved0          | Ver |               Protocol Type          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Key                               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The Key field can be used to provide 32-bit entropy field.

The C (Checksum Present) and S (Sequence Number Present) bits in the GRE header are set to zero. The K bit is set to 1.

[MPLSoUDP] discusses using a UDP header instead of the GRE header to transport MPLS client layer over an IP PSN tunnel. The main advantage for doing so is for better load-balancing capabilities over existing IP networks, where some core routers can perform ECMP based on the UDP header but not based on the GRE Key field. However, the routers that are capable of supporting [NVGRE] encapsulation, can also perform load-balancing based on the GRE key which accommodates a 32-bit entropy value; whereas, UDP encapsulation accommodates a 16-bit entropy value.

##### 3.1.1 Benefits of MPLS over GRE

The benefits of using the MPLS over GRE encapsulation are as follows:

- Uses existing standard for transporting MPLS over IP.
- Uses E-VPN control plane (BGP routes and attributes), as well as E-VPN procedures and functions exactly as is.
- Consistent with L3VPN over IP (RFC 4797)
- The MPLS label can be a global value (instead of downstream

assigned) just like VXLAN or NVGRE service-instance ID.  
- Provides seamless interoperability with E-VPN PEs. There is no need for a gateway device.

### 3.2 VXLAN/NVGRE Encapsulation

If either the VXLAN or NVGRE encapsulation were to be used with the E-VPN control plane, there will be an impact on the E-VPN client layer and the associated procedures and BGP routes. In order to assess this impact, the first step is to identify which subset of the service interfaces defined in [E-VPN] is needed for the NVO solutions defined in [VXLAN] and [NVGRE]. Then we need to examine how the E-VPN BGP routes and procedures should be modified to support these service interfaces with the new encapsulation.

[E-VPN] defines the following four service interface types:

- VLAN Based Service Interface
- VLAN Bundle Service Interface
- Port-based Service Interface
- VLAN Aware Bundle Service Interface

For a detailed description of these service interface types, refer to [EVPN-REQ] and [E-VPN]. As described in [E-VPN], the first three service interface types don't require encoding the VLAN Tag in the BGP routes, because there is a one-to-one mapping between an EVI and a broadcast domain represented by a virtual network or a virtual segment.

[NVGRE] requires only VLAN-based service interface and it clearly describes that the tenant VLAN Tag (inner VLAN Tag) is not part of the encapsulated frames because there is a one-to-one mapping between Virtual Subnet Identifier (VSID) and the inner VLAN ID.

The [VXLAN] default mode of operation only requires VLAN-based service interface, as it specifies that the VTEP does not include an inner VLAN tag upon encapsulation; moreover, the decapsulated frames with an inner VLAN tag should get discarded. However, [VXLAN] provides an option of including an inner VLAN tag in the encapsulated packet if it is configured explicitly at the VTEP. If an inner VLAN tag is included, then VXLAN requires a VLAN-bundle service interface. However, as discussed above, this service interface type does not require that the tenant VLAN tag be sent in the BGP routes.

#### 3.2.1 Impact on E-VPN Routes 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 32-bit Ethernet tag field in the E-VPN BGP routes can be used to represent NVGRE VSID or VXLAN VNI. This is not accidental, but rather by design: The Ethernet Tag field in E-VPN was designed not just for C-tagged or S-tagged interfaces [802.1Q] but also for I-tagged interfaces [802.1ah] where an I-SID is a 24-bit entity representing a virtual segment just like VSID or VNI. Therefore, there is no need to re-purpose the MPLS label field in the E-VPN BGP routes and this field can be omitted in the E-VPN BGP routes. The length field of the NLRI in E-VPN routes will be three octets shorter for VXLAN and NVGRE encapsulations.

Since 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 can be set to zero in the E-VPN routes. However, in the scenarios where different group 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 VNI or VSID, then a non-zero RD MUST be used.

### 3.2.2 Impact on E-VPN Procedures for VXLAN/NVGRE Encapsulation

In order to analyze the impact of the VXLAN/NVGRE encapsulation on E-VPN procedures, we must distinguish three NVE redundancy models:

- No redundancy
- Active/Standby redundancy
- All-active redundancy

The impact of the encapsulation varies depending on the employed model.

#### 3.2.2.1 NVE with No Redundancy

This is the scenario where, for e.g., the NVE is implemented on the hypervisor. In this case, neither the Split-Horizon nor the Aliasing functions are required or applicable. Therefore, the choice of VXLAN/NVGRE encapsulation has no impact on E-VPN procedures.

For all practical purposes, in this scenario, the only difference

between the choice of GRE or VXLAN/NVGRE encapsulation is in the size of the entropy field (32-bits vs. 16 bits).

#### 3.2.2.2 NVE with Active/Standby Redundancy

This is the scenario 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 as well, the Split-Horizon function is not required. However, in order to support fast convergence in case where the primary NVE fails, the Aliasing function of E-VPN is needed. Note that Aliasing in this scenario is used to quickly identify the backup NVE rather than being used for traffic load-balancing. In this case, the impact of the use of the VXLAN/NVGRE encapsulation on the E-VPN procedures is as discussed in Section 3.2.2.3.2, with the difference being that a remote NVE uses the received Ethernet A-D routes to build primary and backup paths to the advertising NVEs, instead of a load-balancing path-list.

If fast convergence is not required or not used, then the VXLAN/NVGRE encapsulation would have no impact on the E-VPN procedures.

#### 3.2.2.3 NVE with All-Active Redundancy

Out of the E-VPN features listed in section 2, 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.

##### 3.2.2.3.1 Split Horizon

In E-VPN, an MPLS label is used for split-horizon filtering to support active/active multi-homing where an ingress NV Edge device (NVE) adds a label corresponding to the site of origin (aka ESI MPLS Label) when encapsulating the packet. The egress NVE 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. One way of supporting this function is to assign an IP address for each site of origin (e.g., for each ESI in the E-VPN terminology) and advertise this IP address in the BGP Remote-Next-Hop attribute associated with the E-VPN Ethernet A-D route (refer to section 3.2.3 for details). The "Active-Standby" bit in the flags of

the ESI MPLS Label Extended Community MUST be set to 0 to indicate active/active multi-homing and the MPLS label field MUST be set to zero to indicate that IP address in the BGP Remote-Next-Hop attribute will be used for split-horizon filtering. The ingress NVE uses the IP address associated with a given site as the source IP address for all traffic originating from said site. The egress NVE will program its egress ACL with this IP address for the interfaces corresponding to that same site.

Although the impact in control plane is minimal and the existing E-VPN BGP routes can be used with minimum modifications to its corresponding procedures, the same cannot be said in terms of network operations, management, and data plane. The use of IP addresses to represent the site of origin requires many IP addresses to be allocated and configured on a single NVE. For example a TOR with N interfaces may require one IP address per interface in worst case which may impact management and operational aspects of the Data Center Network. Also, the data-plane operation for Split-Horizon filtering will be different from that of MPLS client layer and it cannot be assumed that platforms/ASICs that support Split-Horizon filtering based on MPLS label can also support such function based on IP addresses. However, there are alternative options for performing such Split-Horizon filtering function when doing VXLAN/NVGRE encapsulation, while retaining a single IP address per NVE, and those will be described in a future revision of this document.

It should be noted that such filtering function is not required when doing active/standby multi-homing where load-balancing from a tenant can still be performed on a per VLAN basis - e.g., different VLANs are active on different NVEs connected to a multi-homed site. Furthermore, active/active multi-homing is primarily applicable when NVEs are on physical devices as opposed to on the hypervisor. For example, [VXLAN] describes the use of physical devices as VXLAN gateways to connect a legacy network with a VXLAN overlay network. In such scenarios, one would expect: a) that the number of such gateways is not very large and/or b) that not all of them require active/active multi-homing.

#### 3.2.2.3.2 Aliasing

In E-VPN, the NVEs connected to a multi-homed site optionally advertise a VPN label used to load-balance traffic between NVEs, even when a given MAC address is learnt by only a single NVE connected to the site. In the case where VXLAN or NVGRE encapsulation is used, some alternative means that does not rely on MPLS labels is required to support aliasing. One solution would be to rely on the IP address per site assignment depicted in the previous section for aliasing as well: Effectively every NVE advertises an Ethernet A-D route for a

given site with the BGP Remote-Next-Hop attribute set to an IP address that has a 1:1 mapping to the site. The remote NVEs resolve an ESI (site ID) to a list of IP addresses corresponding to that site. Furthermore, a given MAC address that is associated with an ESI, in turn, gets resolved to this list of IP addresses. When a remote NVE wants to forward a packet for a given MAC address, it selects one of IP addresses from the list (using a hash value for load balancing) and encapsulates the packet using that IP address as the destination IP address in the VXLAN or NVGRE encapsulation. The source IP address will be that of the source multi-homed site. In case where the source site is single homed, the source IP address will be the loopback address of the NVE.

### 3.2.2.3.3 Tunnel Endpoint Identification

To accommodate the Split Horizon as well as Aliasing functions of E-VPN, multiple IP tunnel endpoints (one per site) must be associated with the same NVE. As such, the mechanisms of [RFC5512] cannot be used to specify the tunnel endpoint and encapsulation, since those mechanisms only allow a single tunnel endpoint IP address to be associated with the BGP speaker. To alleviate this, the BGP Remote-Next-Hop attribute defined in [REMOTE-NH] can be used. Two new Tunnel Types would be required for VXLAN and NVGRE.

This attribute will be carried with the E-VPN Ethernet A-D route. The IP address field of this attribute serves two functions:

- It indicates the tunnel endpoint destination IP address that must be used when load-balancing traffic associated with a given site (i.e. ESI).
- It is used to build the egress ACL for filtering multi-destination traffic on multi-homed Ethernet Segments. In this context, the IP address is the tunnel endpoint source address.

It is worth noting that for multi-homed Ethernet segments, the NVE will always advertise an Ethernet A-D route with the Remote-Next-Hop attribute, in addition to the MAC Advertisement routes. In this case, the NVEs which receive the routes derive the tunnel endpoint IP address for a given MAC address as follows:

- 1- The NVE identifies the Ethernet Segment Identifier (ESI) associated with the MAC address, as encoded in the MAC Advertisement route.
- 2- The NVE then sets the tunnel endpoint IP address for that MAC to the value encoded in the Remote-Next-Hop attribute of the Ethernet AD



route advertised for the ESI identified in step 1.

On the other hand, for single-homed Ethernet segments, the NVE will only advertise the MAC Advertisement routes. In this latter case, the tunnel endpoint IP address is derived from the BGP Next-Hop attribute associated with the MAC Advertisement route.

### 3.2.3 Support for Multicast

The E-VPN Inclusive Multicast BGP route can be used to discover the multicast 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

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.

### 3.2.4 Inter-AS Challenges

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.

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.

Even 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.

Using MPLS over GRE encapsulation addresses both of these concerns.

#### 4 Comparison between MPLSoGRE and VXLAN/NVGRE Encapsulation

The comparison between MPLSoGRE and VXLAN/NVGRE encapsulation depends on the required functionality on NVEs. If the hosts are single-homed to NVEs without any need to support redundancy group on NVEs, or if the hosts are multi-homed to two or more NVEs with active/standby redundancy but without the need for fast convergence upon a failure, then both MPLSoGRE and VXLAN/NVGRE do equally well with E-VPN control plane.

If we need to support active/standby multi-homing with fast convergence upon a failure or if we need to support active/active multi-homing, then MPLSoGRE encap can provide these additional functionality without any impact to E-VPN routes and procedures. Furthermore, it can provide complete support for inter-AS operation and complete set of E-VPN functions without impacting IP address assignment and management of the underlying network. However, VXLAN/NVGRE impacts E-VPN routes and procedures as well as the underlying data plane behavior as noted above. Furthermore, there are implications to IP address assignments, security, and inter-AS operations. It should be noted that the additional requirements on the data plane behavior as well as the above implications are the consequence of the functionality that need to be supported and

independent of the control-plane choice.

As noted previously, there are existing core switches that do not support ECMP by hashing the GRE key; however, vast majority of existing core switches support ECMP by hashing UDP header; therefore, VXLAN encapsulation can provide better ECMP functions for these existing switches. Thus, the choice for overlay encapsulation depends on needed functionality, inter-AS scenarios, security requirements, and the ECMP capabilities of the core switches.

## 5 Acknowledgement

The authors would like to thank John Mullooly and Dave Smith for providing value comments and feedbacks.

## 6 Security Considerations

## 7 IANA Considerations

## 8 References

### 8.1 Normative References

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

[REMOTE-NH] Van de Velde et al., "BGP Remote-Next-Hop", draft-vandeveldede-idr-remote-next-hop-01.txt, work in progress, July 2012.

### 8.2 Informative References

[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-12vpn-evpn-01.txt, work in progress, February, 2012.

[Problem-Statement] Narten et al., "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement-00, September 2012.

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

#### Authors' Addresses

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

Samer Salam  
Cisco  
595 Burrard Street  
Vancouver, BC V7X 1J1, Canada  
Email: [ssalam@cisco.com](mailto:ssalam@cisco.com)

Keyur Patel  
Cisco  
170 West Tasman Drive  
San Jose, CA 95134, US  
Email: [Keyupate@cisco.com](mailto:Keyupate@cisco.com)

Nabil Bitar  
Verizon Communications  
Email : [nabil.n.bitar@verizon.com](mailto:nabil.n.bitar@verizon.com)

Wim Henderickx  
Alcatel-Lucent  
Email: [wim.henderickx@alcatel-lucent.com](mailto:wim.henderickx@alcatel-lucent.com)

INTERNET-DRAFT  
Intended Status: Informational

Samer Salam  
Ali Sajassi  
Cisco

Sam Aldrin  
Huawei

John E. Drake  
Juniper Networks

Expires: April 18, 2013

October 15, 2012

E-VPN Operations, Administration and Maintenance  
Requirements and Framework

draft-salam-l2vpn-evpn-oam-req-frmwk-00

Abstract

This document specifies the requirements and reference framework for Ethernet VPN (E-VPN) Operations, Administration and Maintenance (OAM). The requirements cover the OAM aspects of E-VPN, PBB-EVPN and TRILL-EVPN. The framework defines the layered OAM model encompassing the E-VPN service layer, network layer and underlying PSN transport layer.

Status of this Memo

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

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

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

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

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

## Copyright and License Notice

Copyright (c) 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	Relationship to Other OAM Work	3
1.2	Specification of Requirements	4
1.3	Terminology	4
2	E-VPN OAM Framework	4
2.1	OAM Layering	4
2.2	E-VPN Service OAM	5
2.3	E-VPN Network OAM	5
2.4	Transport OAM for E-VPN	6
2.5	Link OAM	6
3	E-VPN OAM Requirements	6
3.1	Fault Management Requirements	6
3.1.1	Proactive Fault Management Functions	6
3.1.1.1	Fault Detection (Continuity Check)	6
3.1.1.2	Defect Indication	7
3.1.2	On-Demand Fault Management Functions	8
3.1.2.1	Connectivity Verification	8
3.1.2.2	Fault Isolation	9
3.2	Performance Management	9
3.2.1	Packet Loss	9
3.2.2	Packet Delay	9
4.	Security Considerations	10
5.	IANA Considerations	10
6.	References	10
6.1	Normative References	10
6.2	Informative References	10
	Authors' Addresses	11

## 1 Introduction

This document specifies the requirements and defines a reference framework for Ethernet VPN (E-VPN) Operations, Administration and Maintenance (OAM, [RFC6291]). In this context, we use the term E-VPN OAM to loosely refer to the OAM functions required for and/or applicable to [E-VPN], [PBB-EVPN] as well as [TRILL-EVPN].

E-VPN introduces an L2VPN solution for multipoint Ethernet services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reach-ability information over the core MPLS/IP network.

PBB-EVPN combines Provider Backbone Bridging (PBB) [802.1ah] with E-VPN in order to reduce the number of BGP MAC advertisement routes, provide client MAC address mobility using C-MAC aggregation and B-MAC sub-netting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes.

TRILL-EVPN provides a solution for interconnecting TRILL [TRILL] networks over an MPLS/IP network using E-VPN, with two key characteristics: C-MAC address transparency on the hand-off point and control-plane isolation among the interconnected TRILL networks.

This document focuses on the fault management and performance management aspects of E-VPN OAM.

### 1.1 Relationship to Other OAM Work

This document leverages concepts and draws upon elements defined and/or used in the following documents:

[RFC6136] specifies the requirements and a reference model for OAM as it relates to L2VPN services, pseudowires and associated Public Switched Network tunnels. This document focuses on VPLS and VPWS solutions and services.

[RFC4379] defines mechanisms for detecting data plane failures in MPLS LSPs, including procedures to check the correct operation of the data plane, as well as mechanisms to verify the data plane against the control plane.

[802.1Q] specifies the Ethernet Connectivity Fault Management (CFM) protocol, which defines the concepts of Maintenance Domains, Maintenance End Points, and Maintenance Intermediate Points.

[Y.1731] extends Connectivity Fault Management in the following

areas: it defines fault notification and alarm suppression functions for Ethernet. It also specifies mechanisms for Ethernet performance management, including loss, delay, jitter, and throughput measurement.

## 1.2 Specification of Requirements

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

## 1.3 Terminology

This document uses the following terminology defined in [RFC6136]:

MEP                    Maintenance End Point is responsible for origination and termination of OAM frames for a given MEG.

MIP                    Maintenance Intermediate Point is located between peer MEPs and can process and respond to certain OAM frames but does not initiate or terminate them.

Maintenance Domain   OAM Domain represents a region over which OAM frames can operate unobstructed.

## 2 E-VPN OAM Framework

### 2.1 OAM Layering

Multiple layers come into play for implementing an L2VPN service with the E-VPN family of solutions:

- The Service Layer runs end to end between the sites, or Ethernet Segments, that are being interconnected by the E-VPN solution. It can be either Ethernet (as in [E-VPN], [PBB-EVPN] and [SPB-EVPN]) or TRILL (as in [TRILL-EVPN]).

- The Network Layer extends in between the E-VPN PE nodes and is mostly transparent to the core nodes (except where Flow Entropy comes into play). It leverages MPLS for service (i.e. EVI) multiplexing and Split-Horizon functions.

- The Transport Layer is dictated by the networking technology of the PSN. It may be either based on MPLS LSPs or IP.

- The Link Layer is dependent upon the physical technology used. Ethernet is a popular choice for this layer, but other alternatives are deployed (e.g. POS, DWDM etc...).



This layering extends to the set of OAM protocols that are involved in the ongoing maintenance and diagnostics of E-VPN networks. The figure below depicts the OAM layering, and shows which devices have visibility into what OAM layer(s).

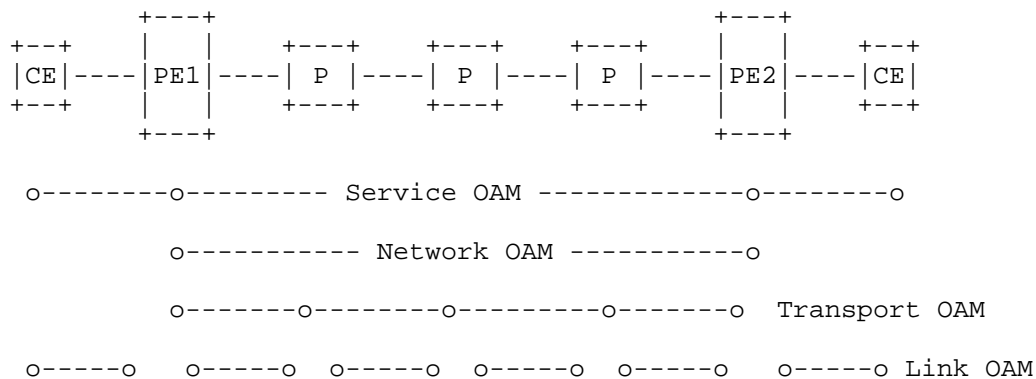


Figure 1: E-VPN OAM Layering

## 2.2 E-VPN Service OAM

The E-VPN Service OAM protocol depends on what service layer is being transported by the E-VPN solution. In case of [E-VPN] and [PBB-EVPN], the service OAM protocol is Ethernet Connectivity Fault Management (CFM) [802.1Q]. Whereas, in case of [TRILL-EVPN], the service OAM protocol is TRILL OAM [TRILL-OAM].

E-VPN service OAM is visible to the CEs and E-VPN PEs, but not to the core (P) nodes. This is because the PEs operate at the Ethernet MAC layer in [E-VPN][PBB-EVPN], or the TRILL RBridge layer in [TRILL-EVPN], whereas the P nodes do not.

The E-VPN PE should support both MEP and MIP functions for the associated service OAM protocol.

## 2.3 E-VPN Network OAM

E-VPN Network OAM is visible to the PE nodes only. This OAM layer is analogous to pseudowire OAM in the case of VPLS/VPWS. It provides capabilities to test connectivity for:

- a given unicast MAC address in a bridge-domain within an EVI (to verify unicast connectivity)

- a given Ethernet Segment in an EVI (to verify the correct operation of Aliasing)

- a given multicast group in a bridge-domain within an EVI (to verify multicast connectivity), including verification of the DF Election status and Split-Horizon filtering.

For the E-VPN network OAM mechanisms to be truly in-band, their messages must be encoded so that they exhibit identical entropy characteristics to data traffic.

## 2.4 Transport OAM for E-VPN

The transport OAM protocol depends on the nature of the underlying transport in the PSN. MPLS OAM mechanisms [RFC4379][RFC6425] as well as ICMP [RFC792] are applicable, depending on whether the PSN employs MPLS or IP transport, respectively.

## 2.5 Link OAM

Link OAM depends on the data link technology being used between the PE and P nodes. For e.g., if Ethernet links are employed, then Ethernet Link OAM [802.3] Clause 57 may be used.

## 3 E-VPN OAM Requirements

This section discusses the E-VPN OAM requirements pertaining to Fault Management and Performance Management. In a future revision of this document, we will identify the OAM layer(s) to which each of the requirements applies.

### 3.1 Fault Management Requirements

#### 3.1.1 Proactive Fault Management Functions

Proactive fault management functions are configured by the network operator to run periodically without a time bound, or are configured to trigger certain actions upon the occurrence of specific events.

##### 3.1.1.1 Fault Detection (Continuity Check)

Proactive fault detection is performed by periodically monitoring the reachability between service endpoints, i.e. MEPs in a given MA, through the exchange of Continuity Check messages. The reachability between any two arbitrary MEPs may be monitored for:

- a specified path taken by a particular user data flow. This enables per Flow monitoring of data paths. E-VPN OAM must support per user

flow fault detection.

- a representative path. This enables liveness check of the nodes hosting the MEPs but does not conclusively indicate liveness of the path(s) taken by user data traffic. This enables node failure detection but not path failure detection, through the use of a test flow. E-VPN OAM must support per test flow fault detection.

- all paths. For MPLS/IP networks with ECMP, monitoring of all unicast paths between MEPs may not be possible, since the per-hop ECMP hashing behavior may yield situations where it is impossible for a MEP to pick flow entropy characteristics that result in exercising the exhaustive set of ECMP paths. Monitoring of all ECMP paths between MEPs is not a requirement for E-VPN OAM.

The fact that MPLS/IP networks do not enforce congruency between unicast and multicast paths means that the proactive fault detection mechanisms for E-VPN must provide procedures to monitor the unicast paths independently of the multicast paths.

### 3.1.1.2 Defect Indication

E-VPN OAM MUST support event-driven defect indication upon the detection of a connectivity defect. Defect indications can be categorized into two types: forward and reverse defect indications.

#### 3.1.1.2.1 Forward Defect Indication

This is used to signal a failure that is detected by a lower layer OAM mechanism. Forward Defect indication is transmitted by a server MEP (i.e. an actual or virtual MEP) in a direction that is away from the direction of the failure (refer to Figure 2 below).

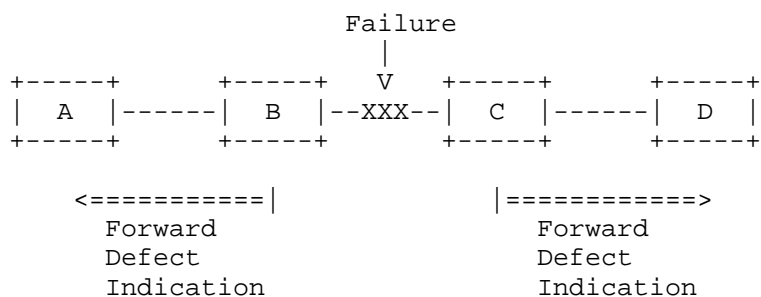


Figure 2: Forward Defect Indication

Forward defect indication may be used for alarm suppression and/or for purpose of inter-working with other layer OAM protocols. Alarm

suppression is useful when a transport/network level fault translates to multiple service or flow level faults. In such a scenario, it is enough to alert a network management station (NMS) of the single transport/network level fault in lieu of flooding that NMS with a multitude of Service or Flow granularity alarms.

### 3.1.1.2.2 Reverse Defect Indication (RDI)

RDI is used to signal that the advertising MEP has detected a loss of continuity (LoC) defect. RDI is transmitted in the direction of the failure (refer to Figure 3).

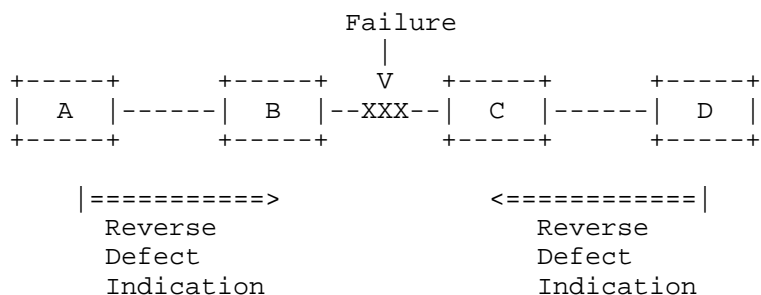


Figure 3: Reverse Defect Indication

RDI allows single-sided management, where the network operator can examine the state of a single MEP and deduce the overall health of a monitored service.

### 3.1.2 On-Demand Fault Management Functions

On-demand fault management functions are initiated manually by the network operator and continue for a time bound period. These functions enable the operator to run diagnostics to investigate a defect condition.

#### 3.1.2.1 Connectivity Verification

E-VPN OAM must support on-demand connectivity verification for unicast and multicast. The connectivity verification mechanism should provide a means for specifying and carrying in the messages:

- variable length payload/padding to test MTU related connectivity problems.
- test traffic patterns as defined in [RFC2544].

For multicast connectivity verification, E-VPN OAM must support

reporting on:

- DF Election Status
- Split Horizon Filtering Status

### 3.1.2.2 Fault Isolation

E-VPN OAM must support an on-demand connectivity fault localization function. This involves the capability to narrow down the locality of a fault to a particular port, link or node. The characteristic of forward/reverse path asymmetry, in MPLS/IP, renders fault isolation into a direction-sensitive operation. That is, given two PEs A and B, localization of connectivity faults between them requires running fault isolation procedures from PE A to PE B as well as from PE B to PE A.

## 3.2 Performance Management

Performance Management functions can be performed both proactively and on-demand. Proactive management involves a scheduling function, where the performance management probes can be triggered on a recurring basis. Since the basic performance management functions involved are the same, we make no distinction between proactive and on-demand functions in this section.

### 3.2.1 Packet Loss

E-VPN OAM must provide mechanisms for measuring packet loss for a given service.

Given that E-VPN provides inherent support for multipoint-to-multipoint connectivity, then packet loss cannot be accurately measured by means of counting user data packets. This is because user packets can be delivered to more RBridges or more ports than are necessary (e.g. due to broadcast, un-pruned multicast or unknown unicast flooding). As such, a statistical means of approximating packet loss rate is required. This can be achieved by sending "synthetic" OAM packets that are counted only by those ports (MEPs) that are required to receive them. This provides a statistical approximation of the number of data frames lost, even with multipoint-to-multipoint connectivity.

### 3.2.2 Packet Delay

E-VPN OAM must support measurement of one-way and two-way packet delay and delay variation (jitter).

#### 4. Security Considerations

E-VPN OAM must provide mechanisms for:

- Preventing denial of service attacks caused by exploitation of the OAM message channel.
- Optionally authenticate communicating endpoints (MEPs and MIPs)
- Preventing OAM packets from leaking outside of the E-VPN network or outside their corresponding Maintenance Domain. This can be done by having MEPs implement a filtering function based on the Maintenance Level associated with received OAM packets.

#### 5. IANA Considerations

None.

#### 6. References

##### 6.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6291] Andersson et al., BCP 161 "Guidelines for the Use of the "OAM" Acronym in the IETF", June 2011.
- [E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-01.txt, work in progress, July 2012.
- [PBB-EVPN] Sajassi et al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-03.txt, work in progress, June 2012.
- [TRILL-EVPN] Sajassi et al., "TRILL-EVPN", draft-ietf-l2vpn-trill-evpn-00.txt, work in progress, June 2012.

##### 6.2 Informative References

- [802.1Q] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks", 31 August 2011.
- [Y.1731] "ITU-T Recommendation Y.1731 (02/08) - OAM functions and mechanisms for Ethernet based networks", February 2008.
- [TRILL-OAM] Senevirathne et al., "Requirements for Operations,

Administration and Maintenance (OAM) in TRILL", draft-ietf-trill-oam-req-01.txt, work in progress, August 2012.

Authors' Addresses

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

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

Sam Aldrin  
Huawei Technologies  
2330 Central Express Way  
Santa Clara, CA 95051, USA  
Email: aldrin.ietf@gmail.com

John E. Drake  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089, USA  
Email: jdrake@juniper.net