

INTERNET-DRAFT  
Intended Status: Standard Track

Sami Boutros  
Ali Sajassi  
Samer Salam  
Cisco Systems

John Drake  
Juniper Networks

Jeff Tantsura  
Ericsson

Dirk Steinberg  
Steinberg Consulting

Thomas Beckhaus  
Deutsche Telecom

Expires: January 3, 2015

July 2, 2014

VPWS support in E-VPN  
draft-boutros-l2vpn-evpn-vpws-04.txt

#### Abstract

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

#### Status of this Memo

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

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

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

material or to cite them other than as "work in progress."

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

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

#### Copyright and License Notice

Copyright (c) 2014 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 . . . . .	4
1.2	Requirements . . . . .	4
2.	BGP Extensions . . . . .	5
3	Operation . . . . .	5
4	EVPN Comparison to PW Signaling . . . . .	7
5	ESI Bandwidth . . . . .	7
7	VPWS with multiple sites . . . . .	8
8	Acknowledgements . . . . .	8
9	Security Considerations . . . . .	8
10	IANA Considerations . . . . .	8
11	References . . . . .	8
11.1	Normative References . . . . .	8
11.2	Informative References . . . . .	8
	Authors' Addresses . . . . .	8

## 1 Introduction

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

[EVPN] has the ability to forward customer traffic to/from a given customer Attachment Circuit (AC), aka Ethernet Segment in EVPN terminology, without any MAC lookup. This capability is ideal in providing p2p services (aka VPWS services). [MEF] defines Ethernet Virtual Private Line (EVPL) service as p2p service between a pair of ACs (designated by VLANs) and Ethernet Private Line (EPL) service, in which all traffic flows are between a single pair of ESes. EVPL can be considered as a VPWS with only two ACs. In delivering an EVPL service, the traffic forwarding capability of EVPN based on the exchange of a pair of Ethernet AD routes is used; whereas, for more general VPWS, traffic forwarding capability of EVPN based on the exchange of a group of Ethernet AD routes (one Ethernet AD route per AC/segment) is used. In a VPWS service, the traffic from an originating Ethernet Segment can be forwarded only to a single destination Ethernet Segment; hence, no MAC lookup is needed and the MPLS label associated with the per-EVI Ethernet AD route can be used in forwarding user traffic to the destination AC.

Both services are supported by using the Per EVI Ethernet AD route which contains an Ethernet Segment Identifier, in which the customer ES is encoded, and an Ethernet Tag, in which the VPWS service instance identifier is encoded. I.e., for both EPL and EVPL services, a specific VPWS service instance is identified by a pair of Per EVI Ethernet AD routes which together identify the VPWS service instance endpoints and the VPWS service instance. In the control plane the VPWS service instance is identified using the VPWS service instance identifiers advertised by each PE and in the data plane the MPLS label advertised by one PE is used by the other PE to send it traffic for that VPWS service instance. As with the Ethernet Tag in standard EVPN, the VPWS service instance identifier has uniqueness within an EVPN instance. The Ethernet Segment identifier encoded in the per EVI Ethernet AD route is not used to identify the service, however it can be used for flow-based load-balancing and mass withdraw functions.

As with standard EVPN, the Per ES Ethernet AD route is used for fast convergence upon link or node failure and the Ethernet Segment route is used for auto-discovery of the PEs attached to a given multi-homed

CE and to synchronize state between them.

## 1.1 Terminology

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

MAC: Media Access Control

MPLS: Multi Protocol Label Switching.

OAM: Operations, Administration and Maintenance.

PE: Provide Edge Node.

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

EVI: EVPN Instance.

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

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

## 1.2 Requirements

1. EPL service access circuit maps to the whole Ethernet port.
2. EVPL service access circuits are VLANs on single or double tagged trunk ports. Each VLAN individually will be considered to be an endpoint for an EVPL service, without any direct dependency on any other VLANs on the trunk. Other VLANs on the same trunk could also be used for EVPL services, but could also be associated with other services.
3. If multiple VLANs on the same trunk are associated with EVPL services, the respective remote endpoints of these EVPLs could be dispersed across any number of PEs, i.e. different VLANs may lead to different destinations.
4. The VLAN tag on the access trunk only has PE-local significance.

The VLAN tag on the remote end could be different, and could also be double tagged when the other side is single tagged.

5. Also, multiple EVPL service VLANs on the same trunk could belong to the same EVPN instance (EVI), or they could belong to different EVIs. This should be purely an administrative choice of the network operator.

6. A given access trunk could have hundreds of EVPL services, and a given PE could have thousands of EVPLs configured. It must be possible to configure multiple EVPL services within the same EVI.

7. Local access circuits configured to belong to a given EVPN instance could also belong to different physical access trunks.

## 2. BGP Extensions

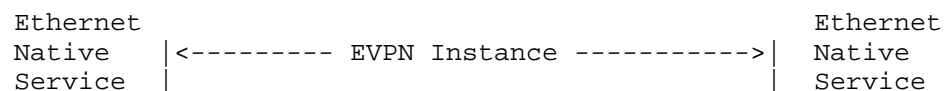
This document proposes the use of the Per EVI Ethernet AD route to signal VPWS services. The Ethernet Segment Identifier field is set to the customer ES and the Ethernet Tag field is set to the VPWS service instance identifier. For both EPL and EVPL services, for a given VPWS service instance the pair of PEs instantiating that VPWS service instance will each advertise a Per EVI Ethernet AD route with its VPWS service instance identifier and will each be configured with the other PE's VPWS service instance identifier. When each PE has received the other PE's

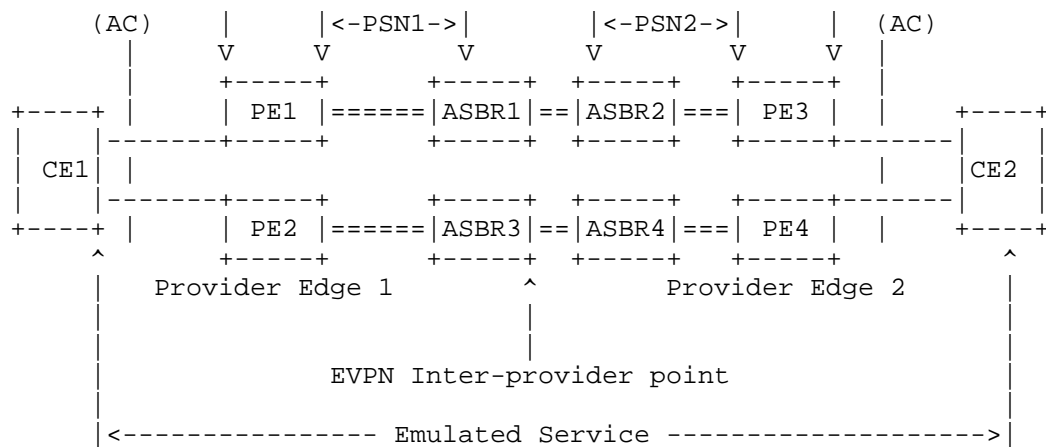
Per EVI Ethernet AD route the VPWS service instance is instantiated. It should be noted that the same VPWS service instance identifier may be configured on both PEs.

The Route-Target (RT) extended community with which the Per EVI Ethernet AD route is tagged identifies the EVPN instance in which the VPWS service instance is configured. It is the operator's choice as to how many and which VPWS service instances are configured in a given EVPN instance. However, a given EVPN instance MUST NOT be configured with both VPWS service instances and standard EVPN multi-point services.

## 3 Operation

The following figure shows an example of a P2P service deployed with EVPN.





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

All PEs and ASBRs are enabled for the EVPN SAFI and exchange Per EVI Ethernet AD routes, one route per VPWS service instance. For inter-AS option B, the ASBRs re-advertise these routes with Next Hop attribute set to their IP addresses. The link between the CE and the PE is either a C-tagged or S-tagged interface, as described in [802.1Q], that can carry a single VLAN tag or two nested VLAN tags and it is configured as a trunk with multiple VLANs, one per VPWS service instance. It should be noted that the VLAN ID used by the customer at either end of a VPWS service instance to identify that service instance may be different and EVPN doesn't perform that translation between the two values. Rather, this should be done by the Ethernet interface.

For single-homed CE, in an advertised Per EVI Ethernet AD route the ESI field is set to 0 and the Ethernet Tag field is set to the VPWS service instance identifier that identifies the EVPL or EPL service.

For a multi-homed CE, in an advertised Per EVI Ethernet AD route the ESI field is set to the CE's ESI and the Ethernet Tag field is set to the VPWS service instance identifier, which MUST have the same value on all PEs attached to that ES. This allows an ingress PE to perform flow-based load-balancing of traffic flows to all of the PEs attached to that ES. In all cases traffic follows the transport paths, which may be asymmetric.

The VPWS service instance identifier encoded in the Ethernet Tag field in an advertised Per EVI Ethernet AD route MUST either be

unique across all ASs, or an ASBR needs to perform a translation when the per EVI Ethernet AD route is re-advertised by the ASBR from one AS to the other AS.

Per ES EAD route can be used for mass withdraw to withdraw all per EVI EAD routes associated with the multi-home site on a given PE.

#### 4 EVPN Comparison to PW Signaling

In EVPN, service endpoint discovery and label signaling are done concurrently using BGP. Whereas, with VPWS based on [RFC4448], label signaling is done via LDP and service endpoint discovery is either through manual provisioning or through BGP.

In existing implementation of VPWS using pseudowires(PWs), redundancy is limited to single-active mode, while with EVPN implementation of VPWS both single-active and all-active redundancy modes can be supported.

In existing implementation with PWs, backup PWs are not used to carry traffic, while with EVPN, traffic can be load-balanced among different PEs multi-homed to a single CE.

Upon link or node failure, EVPN can trigger failover with the withdrawal of a single BGP route per EVPL service or multiple EVPL services, whereas with VPWS PW redundancy, the failover sequence requires exchange of two control plane messages: one message to deactivate the group of primary PWs and a second message to activate the group of backup PWs associated with the access link. Finally, EVPN may employ data plane local repair mechanisms not available in VPWS.

#### 5 ESI Bandwidth

The ESI Bandwidth will be encoded using the Link Bandwidth Extended community defined in [draft-ietf-idr-link-bandwidth] and associated with the Ethernet AD route used to realize the EVPL services.

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

In the case where PSN resources are not available, the PE receiving this attribute MUST re-send its local Ethernet AD routes for this EVPL service with the ESI Bandwidth = All FFs to declare that the

"PSN Resources Unavailable".

The scope of the ESI Bandwidth is limited to only one Autonomous System.

## 7 VPWS with multiple sites

The VPWS among multiple sites (full mesh of P2P connections - one per pair of sites) that can be setup automatically without any explicit provisioning of P2P connections among the sites is outside the scope of this document.

## 8 Acknowledgements

The authors would like to acknowledge Wen Lin contributions to this document.

## 9 Security Considerations

This document does not introduce any additional security constraints.

## 10 IANA Considerations

TBD.

## 11 References

### 11.1 Normative References

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

### 11.2 Informative References

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

[EVPN] A. Sajassi, R. Aggarwal et. al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt.

[PBB-EVPN] A. Sajassi et. al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-05.txt.

[draft-ietf-idr-link-bandwidth] P. Mohapatra, R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-06.txt

## Authors' Addresses



Sami Boutros  
Cisco  
Email: sboutros@cisco.com

Ali Sajassi  
Cisco  
Email: sajassi@cisco.com

Samer Salam  
Cisco  
Email: ssalam@cisco.com

John Drake  
Juniper Networks  
Email: jdrake@juniper.net

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

Dirk Steinberg  
Steinberg Consulting  
Email: dws@steinbergnet.net

Patrice Brissette  
Cisco  
Email: pbrisset@cisco.com

Thomas Beckhaus  
Deutsche Telecom  
Email:Thomas.Beckhaus@telekom.de>

INTERNET-DRAFT  
Intended Status: Informational

Sami Boutros  
Ali Sajassi  
Samer Salam  
Dennis Cai  
Samir Thoria  
Cisco Systems

Tapraj Singh  
John Drake  
Juniper Networks

Jeff Tantsura  
Ericsson

Expires: January 3, 2015

July 2, 2014

VXLAN DCI Using EVPN  
draft-boutros-l2vpn-vxlan-evpn-04.txt

#### Abstract

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

#### Status of this Memo

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

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

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

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

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

#### Copyright and License Notice

Copyright (c) 2014 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	Requirements	4
2.1	Control Plane Separation among VXLAN/NVGRE Networks	4
2.2	All-Active Multi-homing	5
2.3	Layer 2 Extension of VNIs/VSIDs over the MPLS/IP Network	5
2.4	Support for Integrated Routing and Bridging (IRB)	5
3	Solution Overview	5
3.1	Redundancy and All-Active Multi-homing	6
4	EVPN Routes	7
4.1	BGP MAC Advertisement Route	7
4.2	Ethernet Auto-Discovery Route	8
4.3	Per VPN Route Targets	8
4.4	Inclusive Multicast Route	8
4.5	Unicast Forwarding	8
4.6	Handling Multicast	9
4.6.2	Multicast Stitching with Per-VNI Load Balancing	9
5	NVGRE	10
6	Use Cases Overview	10
6.1	Homogeneous Network DCI interconnect Use cases	10
6.1.1	VNI Base Mode EVPN Service Use Case	10
6.1.2	VNI Bundle Service Use Case Scenario	11
6.1.3	VNI Translation Use Case	12
6.2	Heterogeneous Network DCI Use Cases Scenarios	12
6.2.1	VXLAN VLAN Interworking Over EVPN Use Case Scenario	12

7. Acknowledgements . . . . .	13
8. Security Considerations . . . . .	13
9. IANA Considerations . . . . .	13
10. References . . . . .	13
10.1 Normative References . . . . .	13
10.2 Informative References . . . . .	13
Authors' Addresses . . . . .	14

## 1 Introduction

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

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

### 1.1 Terminology

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

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

## 2. Requirements

### 2.1. Control Plane Separation among VXLAN/NVGRE Networks

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

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

trigger re-convergence in remote sites.

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

## 2.2 All-Active Multi-homing

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

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

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

## 2.4 Support for Integrated Routing and Bridging (IRB)

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

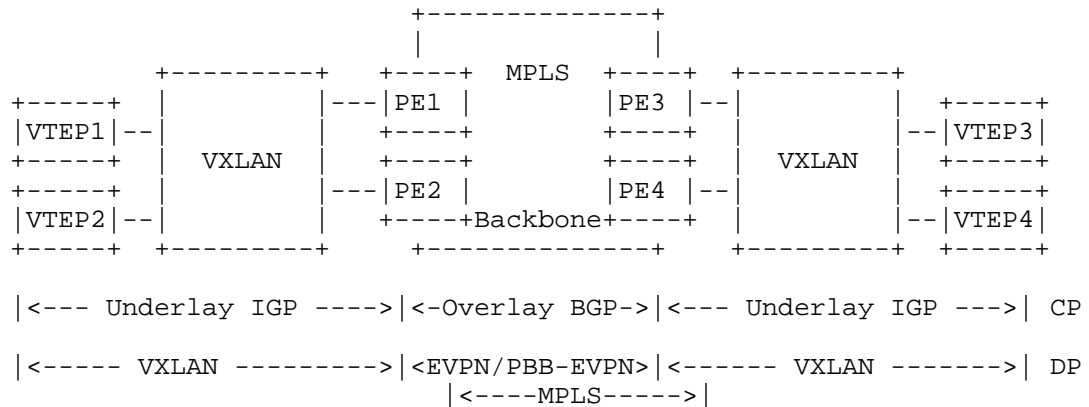
## 3. Solution Overview

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

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

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

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



Legend: CP = Control Plane View

DP = Data Plane View

Figure 1: Interconnecting VXLAN Networks with VXLAN-EVPN

### 3.1. Redundancy and All-Active Multi-homing

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

Since the context of this solution is VXLAN networks with data-plane learning paradigm, it is important for the multi-homing mechanism to ensure stability of the MAC forwarding tables at the NVEs, while supporting all-active forwarding at the PEs. For example, in Figure 1 above, if each PE uses a distinct IP address for its VTEP tunnel, then for a given VNI, when an NVE learns a host's MAC address against the originating VTEP source address, its MAC forwarding table will

keep flip-flopping among the VTEP addresses of the local PEs. This is because a flow associated with the same host MAC address can arrive at any of the PE devices. In order to ensure that there is no flip/flopping of MAC-to-VTEP address associations, an IP Anycast address MUST be used as the VTEP address on all PEs multi-homed to a given VXLAN network. The use of IP Anycast address has two advantages:

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

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

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

#### 4. EVPN Routes

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

##### 4.1. BGP MAC Advertisement Route

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

In case of EVPN, the Ethernet Tag ID of this route is set to zero for VNI-based mode, where there is one-to-one mapping between a VNI and



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

The Ethernet Tag field can be set to a normalized value that maps to the VNI, in VNI aware bundling services, this would make the VNI value of local significance in multiple Data centers. Data plane need to map to this normalized VNI value and have it on the IP VxLAN packets exchanged between the DCIs.

#### 4.2. Ethernet Auto-Discovery Route

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

#### 4.3. Per VPN Route Targets

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

#### 4.4 Inclusive Multicast Route

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

The Inclusive Multicast Route is encoded as follow:

- Ethernet Tag ID is set to zero for VNI-based mode and to VNI for VNI-aware bundle mode.
- Originating Router's IP Address is set to one of the PE's IP addresses.

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

Please see section 4.6 "Handling Multicast"

#### 4.5. Unicast Forwarding

Host MAC addresses will be learnt in data plane from the VXLAN

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

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

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

#### 4.6. Handling Multicast

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

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

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

When LSM is used, the trees are P2MP.

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

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

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

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

##### 4.6.2. Multicast Stitching with Per-VNI Load Balancing

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

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

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

## 5. NVGRE

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

## 6. Use Cases Overview

### 6.1. Homogeneous Network DCI interconnect Use cases This covers DCI interconnect of two or more VXLAN based Data center over MPLS enabled EVPN core.

#### 6.1.1. VNI Base Mode EVPN Service Use Case This use case handles the

EVPN service where there is one to one mapping between a VNI and an EVI. Ethernet TAG ID of EVPN BGP NLRI should be set to Zero. BD ID can be derived from the RT associated with the EVI/VNI.

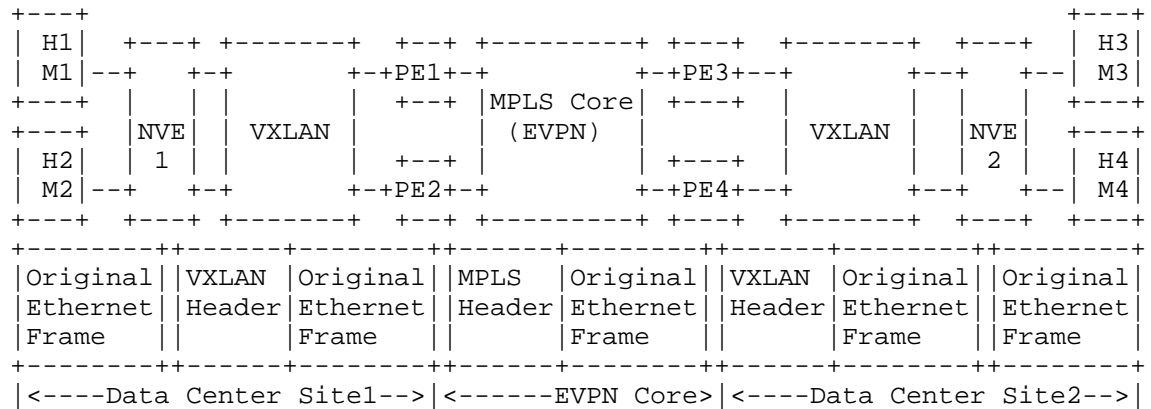


Figure 2 VNI Base Service Packet Flow.

VNI base Service(One VNI mapped to one EVI).

Hosts H1, H2, H3 and H4 are hosts and there associated MAC addresses are M1, M2, M3 and M4. PE1, PE2, PE3 and PE4 are the VXLAN-EVPN gateways. NVE1 and NVE2 are the originators of the VXLAN based network.

When host H1 in Data Center Site1 communicates with H3 in Data Center Site2, H1 forms a layer2 packet with source IP address as IP1 and Source MAC M1, Destination IP as IP3 and Destination MAC as M3(assuming that ARP resolution already happened). VNE1 learns Source MAC and lookup in bridge domain for the Destination MAC. Based on the MAC lookup, the frame needs to be sent to VXLAN network. VXLAN encapsulation is added to the original Ethernet frame and frame is sent over the VXLAN tunnel. Frames arrives at PE1. PE1(i.e. VXLAN gateway), identifies that frame is a VXLAN frame. The VXLAN header is de-capsulated and Destination MAC lookup is done in the bridge domain table of the EVI. Lookup of destination MAC results in the EVPN unicast NH. This NH will be used for identifying the labels (tunnel label and service label) to be added over the EVPN core. Similar processing is done on the other side of DCI.

#### 6.1.2. VNI Bundle Service Use Case Scenario

In the case of VNI-aware bundle service mode, there are multiple VNIs are mapped to one EVI. The Ethernet TAG ID must be set to the VNI ID in the EVPN BGP NLRI. MPLS label allocation in this use case

scenario can be done either per EVI or per EVI, VNI ID basis. If MPLS label allocation is done per EVI basis, then in data path there is a need to push a VLAN TAG for identifying bridge-domain at egress PE so that Destination MAC address lookup can be done on the bridge domain.

### 6.1.3. VNI Translation Use Case

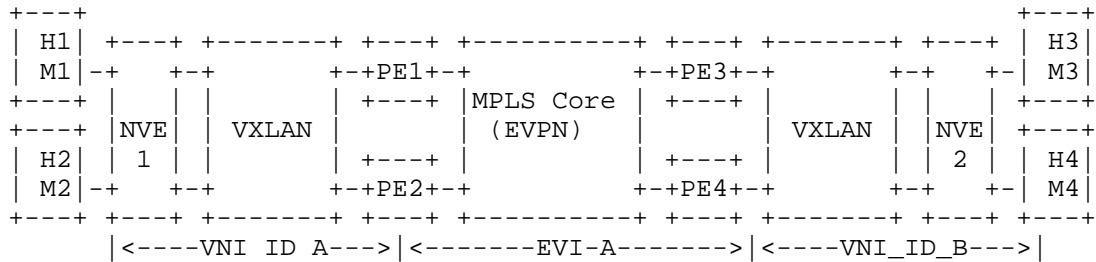


Figure 3 VNI Translation Use Case Scenarios.

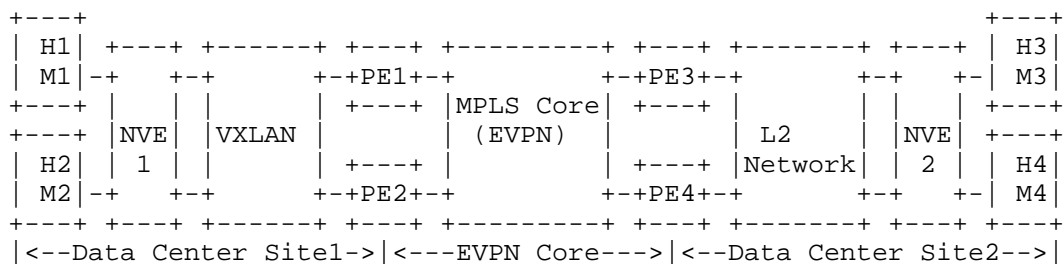
There are two or more Data Center sites. These Data Center sites might use different VNI ID for same service. For example, Service A usage "VNI\_ID\_A" at data center site1 and "VNI\_ID\_B" for same service in data center site 2. VNI ID A is terminated at ingress EVPN PE and VNI ID B is encapsulated at the egress EVPN PE.

## 6.2. Heterogeneous Network DCI Use Cases Scenarios

Data Center sites are upgraded slowly; so heterogeneous network DCI solution is required from the perspective of migration approach from traditional data center to VXLAN based data center. For Example Data Center Site1 is upgrade to VXLAN but Data Center Site 2 and 3 are still layer2/VLAN based data centers. For these use cases, it is required to provide VXLAN VLAN interworking over EVPN core.

### 6.2.1. VXLAN VLAN Interworking Over EVPN Use Case Scenario

The new data center site is VXLAN based data center site. But the older data center sites are still based on the VLAN.



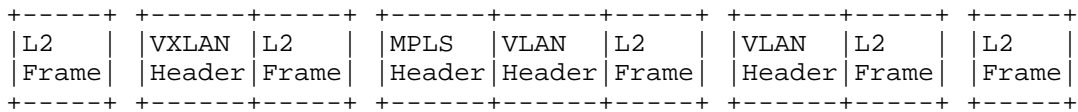


Figure 5 VXLAN VLAN interworking over EVPN Use Case

If a service that are represented by VXLAN on one site of data center and via VLAN at different data center sites, then it is a recommended to model the service as a VNI base EVPN service. The BGP NLRI's will always advertise VLAN ID TAG as '0' in BGP routes. The advantage with this approach is that there is no requirement to do the VNI normalization at EVPN core. VNI ID A is terminated at ingress EVPN PE and "VLAN ID B" is encapsulated at the egress EVPN PE.

## 7. Acknowledgements

The authors would like to acknowledge Wen Lin contributions to this document.

## 8. Security Considerations

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

## 9. IANA Considerations

TBD

## 10. References

### 10.1 Normative References

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

### 10.2 Informative References

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

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

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

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

#### Authors' Addresses

Sami Boutros  
Cisco Systems

EMail: sboutros@cisco.com

Ali Sajassi  
Cisco Systems

EMail: sajassi@cisco.com

Samer Salam  
Cisco Systems  
EMail: ssalam@cisco.com

Dennis Cai  
Cisco Systems  
EMail: dcai@cisco.com

Tapraj Singh  
Juniper Networks  
Email: tsingh@juniper.net

John Drake  
Juniper Networks  
Email: jdrake@juniper.net

Samir Thoria  
Cisco  
EMail: sthoria@cisco.com

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

INTERNET-DRAFT  
Intended Status: Standards track  
Expires: January 4, 2015

Ning So  
Vinci Systems  
Jim Guichard  
Cisco  
Wen Wang  
CenturyLink  
Manuel Paul  
Deutsche Telekom  
Wim Henderichx  
Alcatel-Lucent

Luyuan Fang, Ed.  
Microsoft  
David Ward  
Rex Fernando  
Cisco  
Maria Napierala  
AT&T  
Nabil Bitar  
Verizon  
Dhananjaya Rao  
Cisco  
Bruno Rijsman  
Juniper

July 4, 2014

BGP/MPLS VPN Virtual PE  
draft-fang-l3vpn-virtual-pe-05

Abstract

This document describes the architecture solutions for BGP/MPLS L3 and L2 Virtual Private Networks (VPNs) with virtual Provider Edge (vPE) routers. It provides a functional description of the vPE control, forwarding, and management. The proposed vPE solutions support both the Software Defined Networks (SDN) approach which allows physical decoupling of the control and the forwarding, and the traditional distributed routing approach. A vPE can reside in any network or compute devices, such as a server as co-resident with the application virtual machines (VMs), or a Top-of-Rack (ToR) switch in a Data Center (DC) network.

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) 2014 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
1.2	Requirements . . . . .	5
2.	Virtual PE Architecture . . . . .	6
2.1	Virtual PE definitions . . . . .	6
2.2	vPE Architecture and Design options . . . . .	8
2.2.1	vPE-F host location . . . . .	8
2.2.2	vPE control plane topology . . . . .	8
2.2.3	Data Center orchestration models . . . . .	8
2.3	vPE Architecture reference models . . . . .	8
2.3.1	vPE-F in an end-device and vPE-C in the controller . . . . .	8
2.3.2	vPE-F and vPE-C on the same end-device . . . . .	10
2.3.3	vPE-F and vPE-C are on the ToR . . . . .	11
2.3.4	vPE-F on the ToR and vPE-C on the controller . . . . .	12
2.3.5	The server view of a vPE . . . . .	12
3.	Control Plane . . . . .	13
3.1	vPE Control Plane (vPE-C) . . . . .	13
3.1.1	The SDN approach . . . . .	13
3.1.2	Distributed control plane . . . . .	14
3.3	Use of router reflector . . . . .	14
3.4	Use of Constrained Route Distribution [RFC4684] . . . . .	14
4.	Forwarding Plane . . . . .	14
4.1	Virtual Interface . . . . .	14
4.2	Virtual Provider Edge Forwarder (vPE-F) . . . . .	15

4.3 Encapsulation . . . . .	15
4.4 Optimal forwarding . . . . .	15
4.5 Routing and Bridging Services . . . . .	16
5. Addressing . . . . .	17
5.1 IPv4 and IPv6 support . . . . .	17
5.2 Address space separation . . . . .	17
6.0 Inter-connection considerations . . . . .	17
7. Management, Control, and Orchestration . . . . .	18
7.1 Assumptions . . . . .	18
7.2 Management/Orchestration system interfaces . . . . .	19
7.3 Service VM Management . . . . .	19
7.4 Orchestration and MPLS VPN inter-provisioning . . . . .	19
7.4.1 vPE Push model . . . . .	20
7.4.2 vPE Pull model . . . . .	21
8. Security Considerations . . . . .	21
9. IANA Considerations . . . . .	22
10. Acknowledgments . . . . .	22
11. References . . . . .	22
11.1 Normative References . . . . .	22
11.2 Informative References . . . . .	23
Authors' Addresses . . . . .	24

## 1 Introduction

Network virtualization enables multiple isolated individual networks over a shared common network infrastructure. BGP/MPLS IP Virtual Private Networks (IP VPNs) [RFC4364] have been widely deployed to provide network based Layer 3 VPNs solutions. [RFC4364] provides routing isolation among different customer VPNs and allow address overlap among these VPNs through the implementation of per VPN Virtual Routing and Forwarding instances (VRFs) at a Service Provider Edge (PE) routers, while forwarding customer traffic over a common IP/MPLS network. For L2 VPN, a similar technology is being defined in [I-D.ietf-l2vpn-evpn] on the basis of BGP/MPLS, to provide switching isolation and allow MAC address overlap.

With the advent of compute capabilities and the proliferation of virtualization in Data Center servers, multi-tenant Data Centers are becoming the norm. As applications and appliances are increasingly being virtualized, support for virtual edge devices, such as virtual L3/L2 VPN PE routers, becomes feasible and desirable for Service Providers who want to extend their existing MPLS VPN deployments into Data Centers to provide end-to-end Virtual Private Cloud (VPC) services. Virtual PE work is also one of early effort for Network Functions Virtualization (NFV). In general, scalability, agility, and cost efficiency are primary motivations for vPE solutions.

The virtual Provider Edge (vPE) solution described in this document allows for the extension of the PE functionality of L3/L2 VPN to an end device, such as a server where the applications reside, or to a first hop routing/switching device, such as a Top of the Rack (ToR) switch in a DC.

The vPE solutions support both the Software Defined Networks (SDN) approach, which allows physical decoupling of the control and the forwarding, and the traditional distributed routing approach.

### 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].

Term	Definition
ASBR	Autonomous System Border Router
BGP	Border Gateway Protocol
CE	Customer Edge
Forwarder	IP VPN forwarding function

GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager
I2RS	Interface to Routing Systems
LDP	Label Distribution Protocol
MP-BGP	Multi-Protocol Border Gateway Protocol
MPLS	Multi-Protocol Label Switching
PCEF	Policy Charging and Enforcement Function
QoS	Quality of Service
RR	Route Reflector
RT	Route Target
RTC	RT Constraint
SDN	Software Defined Networks
ToR	Top-of-Rack switch
VI	Virtual Interface
vCE	virtual Customer Edge Router
VM	Virtual Machine
vPC	virtual Private Cloud
vPE	virtual Provider Edge Router
vPE-C	virtual Provider Edge Control plane
vPE-F	virtual Provider Edge Forwarder
VPN	Virtual Private Network
vRR	virtual Route Reflector
WAN	Wide Area Network

End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside.

## 1.2 Requirements

The following are key requirements for vPE solutions.

- 1) MUST support end device multi-tenancy, per tenant routing isolation and traffic separation.
- 2) MUST support large scale MPLS VPNs in the Data Center, upto tens of thousands of end devices and millions of VMs in the single Data Center.
- 3) MUST support end-to-end MPLS VPN connectivity, e.g. MPLS VPN can start from a DC end device, connect to a corresponding MPLS VPN in the WAN, and terminate in another Data Center end device.
- 4) MUST allow physical decoupling of MPLS VPN PE control and forwarding for network virtualization and abstraction.
- 5) MUST support the control plane with both SDN controller approach, and the traditional distributed control plane approach with MP-BGP protocol.

- 6) MUST support VM mobility.
- 7) MUST support orchestration/auto-provisioning deployment model.
- 8) SHOULD be capable to support service chaining as part of the solution [I-D.rfernando-l3vpn-service-chaining], [I-D.bitar-i2rs-service-chaining].

The architecture and protocols defined in BGP/MPLS IP VPN [RFC4364] and BGP/MPLS EVPN [I-D.ietf-l2vpn-evpn] provide the foundation for vPE extension. Certain protocol extensions may be needed to support the virtual PE solutions.

## 2. Virtual PE Architecture

### 2.1 Virtual PE definitions

As defined in [RFC4364] and [I-D.ietf-l2vpn-evpn], an MPLS VPN is created by applying policies to form a subset of sites among all sites connected to the backbone networks. It is a collection of "sites". A site can be considered as a set of IP/ETH systems maintaining IP/ETH inter-connectivity without direct connecting through the backbone. The typical use of L3/L2 VPN has been to inter-connect different sites of an Enterprise networks through a Service Provider's BGP MPLS VPNs in the WAN.

A virtual PE (vPE) is a BGP/MPLS L3/L2 VPN PE software instance which may reside in any network or computing devices. The control and forwarding components of the vPE can be decoupled, they may reside in the same physical device, or in different physical devices.

A virtualized Provider Edge Forwarder (vPE-F) is the forwarding element of a vPE. vPE-F can reside in an end device, such as a server in a Data Center where multiple application Virtual Machines (VMs) are supported, or a Top-of-Rack switch (ToR) which is the first hop switch from the Data Center edge. When a vPE-F is residing in a server, its connection to a co-resident VM can be viewed as similar to the PE-CE relationship in the regular BGP L3/L2 VPNs, but without routing protocols or static routing between the virtual PE and end-host because the connection is internal to the device.

The vPE Control plane (vPE-C) is the control element of a vPE. When using the approach where control plane is decoupled from the physical topology, the vPE-F may be in a server and co-resident with application VMs, while one vPE-C can be in a separate device, such as an SDN Controller where control plane elements and orchestration functions are located. Alternatively, the vPE-C can reside in the same physical device as the vPE-F. In this case, it is similar to the

traditional implementation of VPN PEs where, distributed MP-BGP is used for L3/L2 VPN information exchange, though the vPE is not a dedicated physical entity as it is in a physical PE implementation.

## 2.2 vPE Architecture and Design options

### 2.2.1 vPE-F host location

Option 1a. vPE-F is on an end device as co-resident with application VMs. For example, the vPE-F is on a server in a Data Center.

Option 1b. vPE-F forwarder is on a ToR or other first hop devices in a DC, not as co-resident with the application VMs.

Option 1c. vPE-F is on any network or compute devices in any types of networks.

### 2.2.2 vPE control plane topology

Option 2a. vPE control plane is physically decoupled from the vPE-F. The control plane may be located in a controller in a separate device (a stand alone device or can be in the gateway as well) from the vPE forwarding plane.

Option 2b. vPE control plane is supported through dynamic routing protocols and located in the same physical device as the vPE-F.

### 2.2.3 Data Center orchestration models

Option 3a. Push model: It is a top down approach, push IP VPN provisioning state from a network management system or other centrally controlled provisioning system to the IP VPN network elements.

Option 3b. Pull model: It is a bottom-up approach, pull state information from network elements to network management/AAA based upon data plane or control plane activity.

## 2.3 vPE Architecture reference models

### 2.3.1 vPE-F in an end-device and vPE-C in the controller

Figure 1 illustrates the reference model for a vPE solution with the vPE-F in the end device co-resident with applications VMs, while the vPE-C is physically decoupled and residing on a controller.

The Data Center is connected to the IP/MPLS core via the Gateways/ASBRs. The MPLS VPN, e.g. VPN RED, has a single termination point within the DC at one of the VPE-F, and is inter-connected in the WAN to other member sites which belong to the same client, and the remote ends of VPN RED can be a PE which has VPN RED attached to it, or another vPE in a different Data Center.

Note that the DC fabrics/intermediate underlay devices in the DC do not participate IP VPNs, their function is the same as provider backbone routers in the IP/MPLS back bone and they do not maintain the VPN states, nor they are VPN aware.

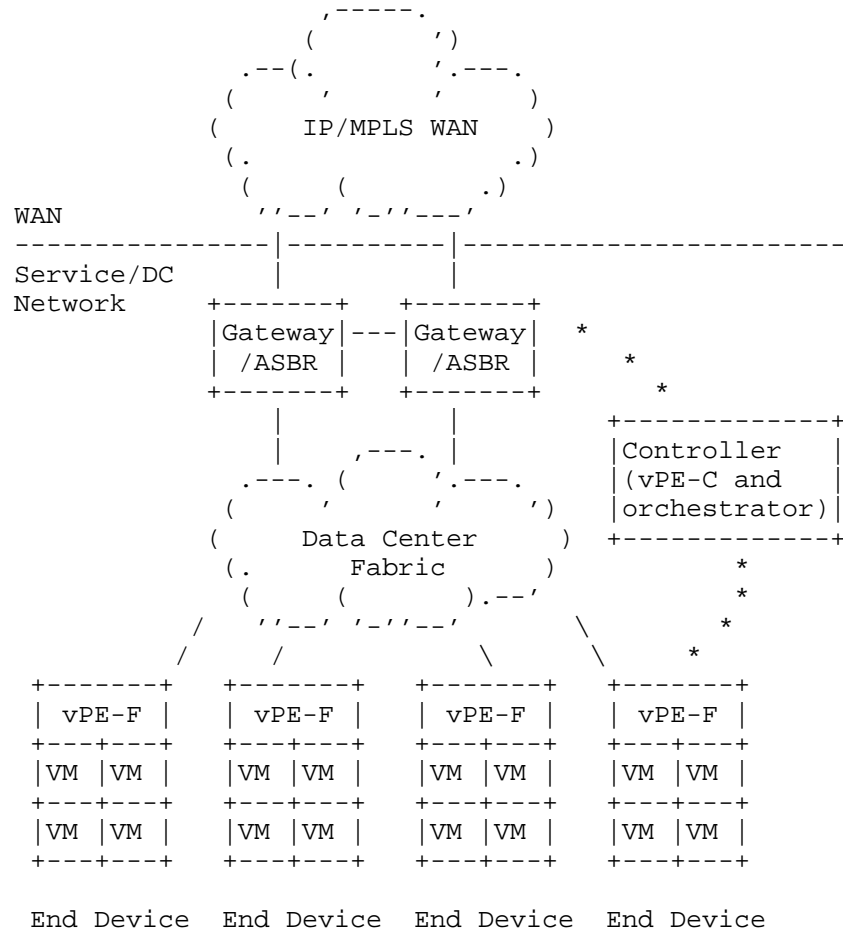


Figure 1. Virtualized Data Center with vPE at the end device and vPE-C and vPE-F physically decoupled

Note:

- \*\*\* represents Controller logical connections to the all Gateway/ASBRs and to all vPE-F.
- ToR is assumed included in the Data Center cloud.



## 2.3.2 vPE-F and vPE-C on the same end-device

In this option, vPE-F and vPE-C functionality are both resident in the end-device. The vPE functions the same as it is in a physical PE. MP-BGP is used for the VPN control plane. Virtual or physical Route Reflectors (RR) (not shown in the diagram) can be used to assist scaling.

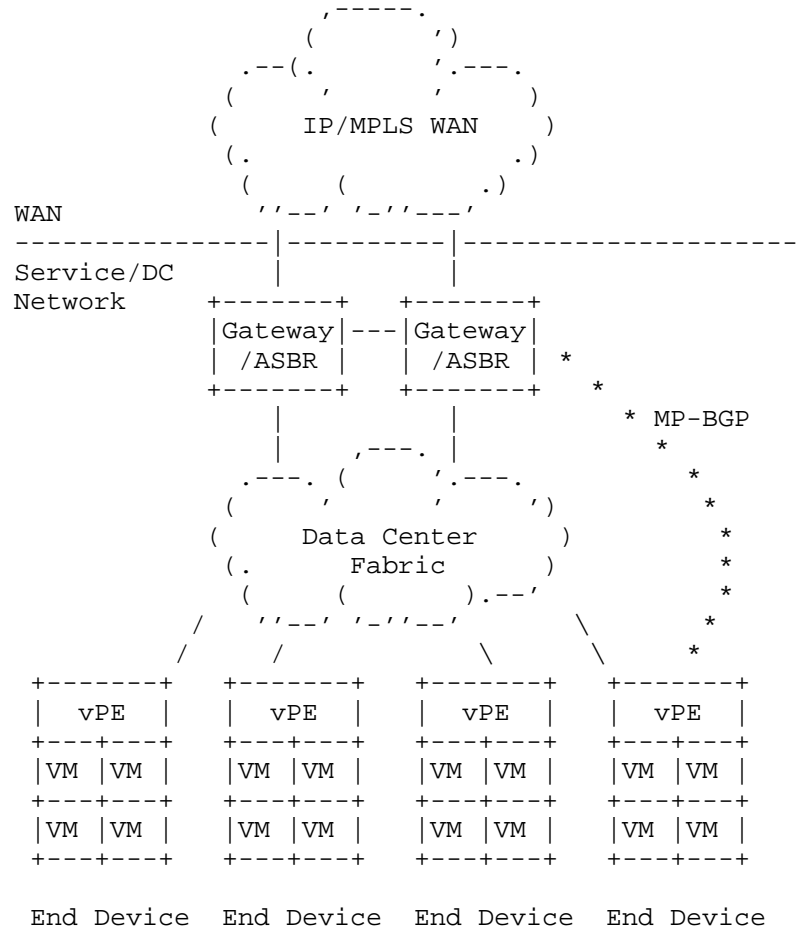


Figure 2. Virtualized Data Center with vPE at the end device, VPN control signal uses MP-BGP

Note:

a) \*\*\* represents the logical connections using MP-BGP among the Gateway/ASBRs and to the vPEs on the end devices.

b) ToR is assumed included in the Data Center cloud.

### 2.3.3 vPE-F and vPE-C are on the ToR

In this option, vPE functionality is the same as a physical PE. MP-BGP is used for the VPN control plane. Virtual or physical Route Reflector (RR) (not shown in the diagram) can be used to assist scaling.

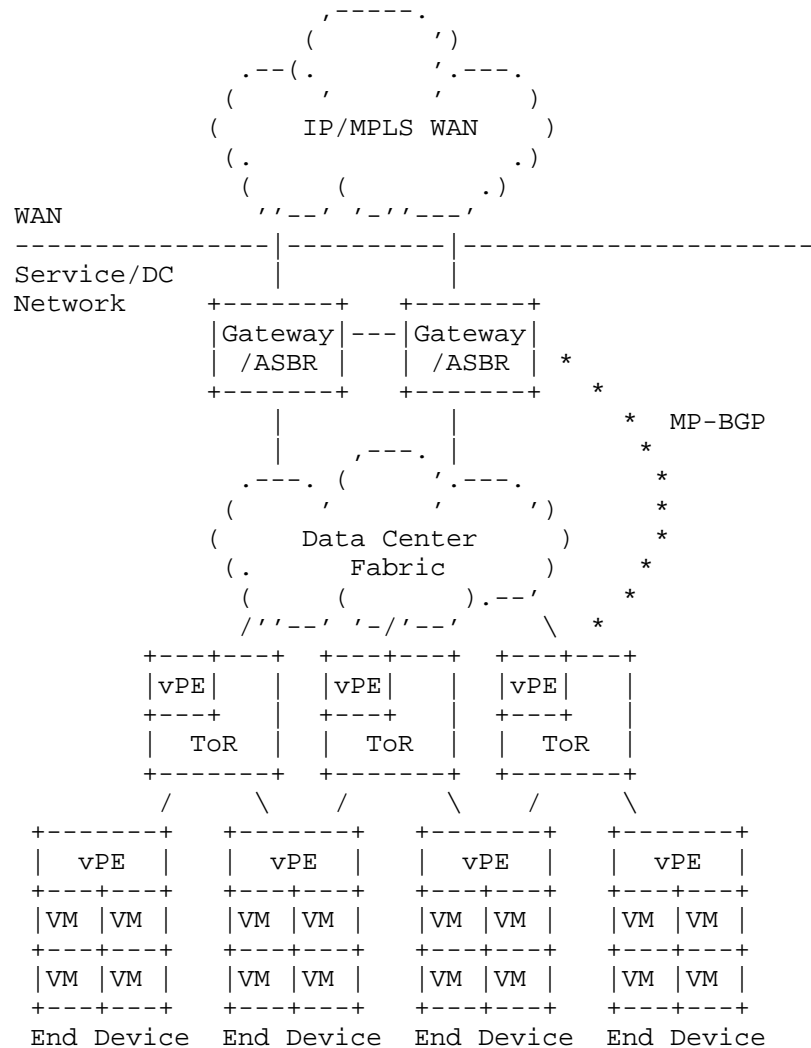


Figure 3. Virtualized Data Center with vPE at the ToP, VPN control signal uses MP-BGP

Note: \*\*\* represents the logical connections using MP-BGP among the Gateway/ASBRs and to the vPEs on the ToRs.

#### 2.3.4 vPE-F on the ToR and vPE-C on the controller

In this option, the L3/L2 VPN termination is at the ToR, but the control plane decoupled from the data plane and resided in a controller, which can be on a stand alone device, or can be placed at the Gateway/ASBR.

#### 2.3.5 The server view of a vPE

An end device shown in Figure 4 is a virtualized server that hosts multiple VMs. The virtual PE is co-resident in the server with application VMs. The vPE supports multiple VRFs, VRF Red, VRF Grn, VRF Yel, VRF Blu, etc. Each application VM is associated to a particular VRF as a member of the particular VPN. For example, VM1 is associated to VRF Red, VM2 and VM47 are associated to VRF Grn, etc. Routing/switching isolation applies between VPNs for multi-tenancy support. For example, VM1 and VM2 cannot communicate directly in a simple intranet VPN topology as shown in the configuration.

The vPE connectivity relationship between vPE and the application VM is similar to the PE-to-CE relationship in regular BGP VPNs. However, as the vPE and end-host functions are co-resident in the same server, the connection between them is an internal implementation of the server.

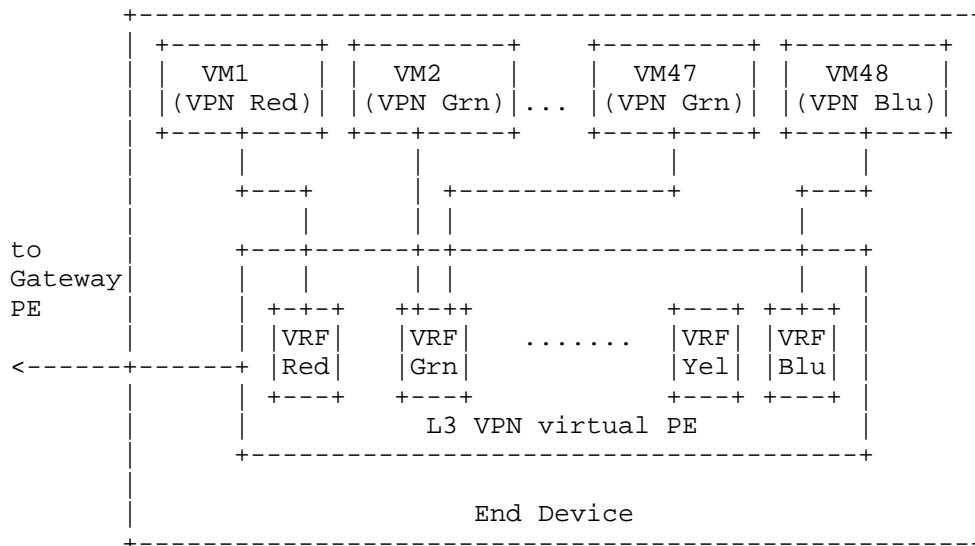


Figure 4. Server View of vPE to VM relationship

An application VM may send packets to a vPE forwarder that need to be bridged, either locally to another VM, or to a remote destination. In this case, the vPE contains a virtual bridge instance to which the application VMs (CEs) are attached.

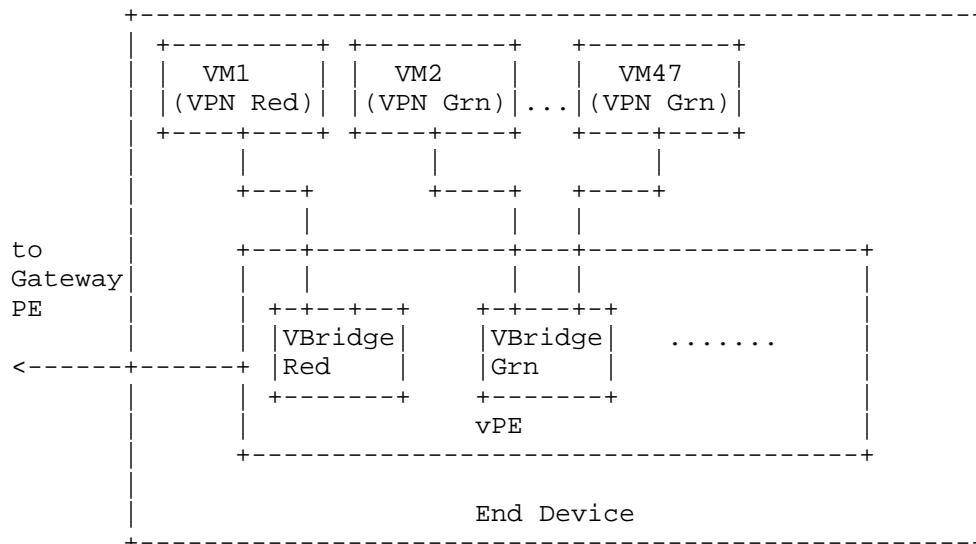


Figure 4. Bridging Service at vPE

### 3. Control Plane

#### 3.1 vPE Control Plane (vPE-C)

##### 3.1.1 The SDN approach

This approach is appropriate when the vPE control and data planes are physically decoupled. The control plane directing the data flow may reside elsewhere, e.g. in a SDN controller. This approach requires a standard interface to the routing system. The Interface to Routing System (I2RS) is work in progress in IETF as described in [I-D.ietf-i2rs-architecture], [I-D.ietf-i2rs-problem-statement].

Although MP-BGP is often the de facto preferred choice between vPE and gateway-PE/ASBR, the use of extensible signaling messaging protocols MAY often be more practical in a Data Center environment. One such proposal that uses this approach is detailed in [I-D.ietf-l3vpn-end-system].

### 3.1.2 Distributed control plane

In the distributed control plane approach, the vPE participates in the overlay L3/L2 VPN control protocol: MP-BGP [RFC4364].

When the vPE function is on a ToR, it participates the underlay routing through IGP protocols (ISIS or OSPF) or BGP.

When the vPE function is on a server, it functions as a host attached to a server.

### 3.3 Use of router reflector

Modern Data Centers can be very large in scale. For example, the number of VPNs routes in a very large DC can surpass the scale of those in a Service Provider backbone VPN networks. There may be tens of thousands of end devices in a single DC.

Use of Router Reflector (RR) is necessary in large-scale IP VPN networks to avoid a full iBGP mesh among all vPEs and PEs. The VPN routes can be partitioned to a set of RRs, the partitioning techniques are detailed in [RFC4364] and [I-D.ietf-l2vpn-evpn].

When a RR software instance is residing in a physical device, e.g., a server, which is partitioned to support multi-functions and application VMs, the RR becomes a virtualized RR (vRR). Since RR performs control functions only, a dedicated or virtualized server with large scale of computing power and memory can be a good candidate as host of vRRs. The vRR can also reside in a Gateway PE/ASBR, or in an end device.

### 3.4 Use of Constrained Route Distribution [RFC4684]

The Constrained Route Distribution [RFC4684] is a powerful tool for selective VPN route distribution. With RTC, only the BGP receivers (e.g, PE/vPE/RR/vRR/ASBRs, etc.) with the particular IP VPNs attached will receive the route update for the corresponding VPNs. It is critical to use constrained route distribution to support large-scale IP VPN developments.

## 4. Forwarding Plane

### 4.1 Virtual Interface

A Virtual Interface (VI) is an interface within an end device that is used for connection of the vPE to the application VMs in the same end device. Such application VMs are treated as CEs in the regular VPN's view.

#### 4.2 Virtual Provider Edge Forwarder (vPE-F)

The Virtual Provider Edge Forwarder (vPE-F) is the forwarding component of a vPE where the tenant identifiers (for example, MPLS VPN labels) are pushed/popped.

The vPE-F location options include:

- 1) Within the end device where the virtual interface and application VMs are located.
- 2) In an external device such as a Top of the Rack switch (ToR) in a DC into which the end device connects.

Multiple factors should be considered for the location of the vPE-F, including device capabilities, overall solution economics, QoS/firewall/NAT placement, optimal forwarding, latency and performance, operational impact, etc. There are design tradeoffs, it is worth the effort to study the traffic pattern and forwarding looking trend in your own unique Data Center as part of the exercise.

#### 4.3 Encapsulation

BGP/MPLS VPNs can be tunneled through the network as overlays using MPLS-based or IP-based encapsulation.

In the case of MPLS-based encapsulation, most existing core deployments use distributed protocols such as Label Distribution Protocol (LDP), [RFC3032][RFC5036], or RSVP-TE [RFC3209].

Due to its maturity, scalability, and header efficiency, MPLS Label Stacking is gaining traction by service providers, and large-scale cloud providers in particular, as the unified forwarding mechanism of choice.

With the emergence of the SDN paradigm, label distribution may be achieved through SDN controllers, or via a combination of centralized control and distributed protocols.

In the case of IP-based encapsulation, MPLS VPN packets are encapsulated in IP or Generic Routing Encapsulation (GRE), [RFC4023], [RFC4797]. IP-based encapsulation has not been extensively deployed for BGP/MPLS VPN in the core; however it is considered as one of the tunneling options for carrying MPLS VPN overlays in the data center. Note that when IP encapsulation is used, the associated security properties must be analyzed carefully.

#### 4.4 Optimal forwarding

Many large cloud service providers have reported the DC traffic is now dominated by East-West across subnet traffic (between the end device hosting different applications in different subnets) rather than North-South traffic (going in/out of the Data Center and to/from the WAN) or switched traffic within subnets. This is the primary reason that newer DC design has moved away from traditional Layer-2 design to Layer-3, especially for the overlay networks.

When forwarding the traffic within the same VPN, the vPE SHOULD be capable to provide direct communication among the VMs/application senders/receivers without the need of going through Gateway devices. If the senders and the receivers are on the same end device, the traffic SHOULD NOT need to leave the device. If they are on different end devices, optimal routing SHOULD be applied.

Extranet MPLS VPN techniques can be used for multiple VPNs access without the need of Gateway facilitation. This is done through the use of VPN policy control mechanisms.

In addition, ECMP is a built in IP mechanism for load sharing. Optimal use of available bandwidth can be achieved by virtue of using ECMP in the underlay, as long as the encapsulation includes certain entropy in the header, VXLAN is such an example.

#### 4.5 Routing and Bridging Services

A VPN forwarder (vPE-F) may support both IP forwarding as well as Layer 2 bridging for traffic from attached end hosts. This traffic may be between end hosts attached to the same VPN forwarder or to different VPN forwarders.

In both cases, forwarding at a VPN forwarder takes place based on the IP or MAC entries provisioned by the vPE controller.

When the vPE is providing Layer 3 service to the attached CEs, the VPN forwarder has a VPN VRF instance with IP routes installed for both locally attached end-hosts and ones reachable via other VPN forwarders. The vPE may perform IP routing for all IP packets in this mode.

When the vPE provides Layer 2 service to the attached end-hosts, the VPN forwarder has an E-VPN instance with appropriate MAC entries.

The vPE may support an Integrated Routing and Bridging service, in which case the relevant VPN forwarders will have both MAC and IP table entries installed, and will appropriately route or switch incoming packets.

The vPE controller performs the necessary provisioning functions to support various services, as defined by an user.

## 5. Addressing

### 5.1 IPv4 and IPv6 support

IPv4 and IPv6 MUST be supported in the vPE solution.

This may present a challenge for older devices, but this normally is not an issue for the newer generation of forwarding devices and servers. Note that a server is replaced much more frequently than a network router/switch, and newer equipment SHOULD be capable of IPv6 support.

### 5.2 Address space separation

The addresses used for the IP VPN overlay in a DC, SHOULD be taken from separate address blocks outside the ones used for the underlay infrastructure of the DC. This practice is to protect the DC infrastructure from being attacked if the attacker gains access to the tenant VPNs.

Similarity, the addresses used for the DC SHOULD be separated from the WAN backbone addresses space.

### 6.0 Inter-connection considerations

The inter-connection considerations in this section are focused on intra-DC inter-connections.

There are deployment scenarios where BGP/MPLS IP VPN may not be supported in every segment of the networks to provide end-to-end IP VPN connectivity. A vPE may be reachable only via an intermediate inter-connecting network; interconnection may be needed in these cases.

When multiple technologies are employed in the solution, a clear demarcation should be preserved at the inter-connecting points. The problems encountered in one domain SHOULD NOT impact other domains.

From an IP VPN point of view: An IP VPN vPE that implements [RFC4364] is a component of the IP VPN network only. An IP VPN VRF on a physical PE or vPE contains IP routes only, including routes learnt over the locally attached network.

The IP VPN vPE should ideally be located as close to the "customer" edge devices as possible. When this is not possible, simple existing



"IP VPN CE connectivity" mechanisms should be used, such as static, or direct VM attachments such as described in the vCE [I-D.fang-l3vpn-virtual-ce] option below.

Consider the following scenarios when BGP MPLS VPN technology is considered as whole or partial deployment:

Scenario 1: All VPN sites (CEs/VMs) support IP connectivity. The most suited BGP solution is to use IP VPNs [RFC4364] for all sites with PE and/or vPE solutions.

Scenario 2: Legacy Layer 2 connectivity must be supported in certain sites/CEs/VMs, and the rest of the sites/CEs/VMs need only Layer 3 connectivity.

One can consider using a combined vPE and vCE [I-D.fang-l3vpn-virtual-ce] solution to solve the problem. Use IP VPN for all sites with IP connectivity, and a physical or virtual CE (vCE, may reside on the end device) to aggregate the Layer 2 sites which for example, are in a single container in a Data Center. The CE/vCE can be considered as inter-connecting points, where the Layer 2 network is terminated and the corresponding routes for connectivity of the L2 network are inserted into IP VPN VRFs. The Layer 2 aspect is transparent to the L3VPN in this case.

Reducing operation complicity and maintaining the robustness of the solution are the primary reasons for the recommendations.

The interconnection of MPLS VPN in the data center and the MPLS core through ASBR using existing inter-AS options is discussed in detail in [I-D.fang-l3vpn-data-center-interconnect].

## 7. Management, Control, and Orchestration

### 7.1 Assumptions

The discussion in this section is based on the following set of assumptions:

- The WAN and the inter-connecting Data Center, MAY be under control of separate administrative domains
- WAN Gateways/ASBRs/PEs are provisioned by existing WAN provisioning systems
- If a single Gateway/ASBR/PE connecting to the WAN on one side, and connecting to the Data Center network on the other side, then this Gateway/ASBR/PE is the demarcation point between the two networks.

- vPEs and VMs are provisioned by Data Center Orchestration systems.
- Managing IP VPNs in the WAN is not within the scope of this document except the inter-connection points.

## 7.2 Management/Orchestration system interfaces

The Management/Orchestration system CAN be used to communicate with both the DC Gateway/ASBR, and the end devices.

The Management/Orchestration system MUST support standard, programmatic interface for full-duplex, streaming state transfer in and out of the routing system at the Gateway.

The programmatic interface is currently under definition in IETF Interface to Routing Systems (I2RS)) initiative. [I-D.ietf-i2rs-architecture], and [I-D.ietf-i2rs-problem-statement].

Standard data modeling languages will be defined/identified in I2RS. YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF) [RFC6020] is a promising candidate currently under investigation.

To support remote access between applications running on an end device (e.g., a server) and routers in the network (e.g. the DC Gateway), a standard mechanism is expected to be identified and defined in I2RS to provide the transfer syntax, as defined by a protocol, for communication between the application and the network/routing systems. The protocol(s) SHOULD be lightweight and familiar by the computing communities. Candidate examples include ReSTful web services, JSON [RFC7159], NETCONF [RFC6241], XMPP [RFC6120], and XML. [I-D.ietf-i2rs-architecture].

## 7.3 Service VM Management

Service VM Management SHOULD be hypervisor agnostic, e.g. On demand service VMs turning-up SHOULD be supported.

## 7.4 Orchestration and MPLS VPN inter-provisioning

The orchestration system

- 1) MUST support MPLS VPN service activation in virtualized DC.
- 2) MUST support automated cross-provisioning accounting correlation between the WAN MPLS VPN and Data Center for the same tenant.
- 3) MUST support automated cross provisioning state correlation

between WAN MPLS VPN and Data Center for the same tenant

There are two primary approaches for IP VPN provisioning - push and pull, both CAN be used for provisioning/orchestration.

#### 7.4.1 vPE Push model

Push model: push IP VPN provisioning from management/orchestration systems to the IP VPN network elements.

This approach supports service activation and it is commonly used in existing MPLS VPN Enterprise deployments. When extending existing WAN IP VPN solutions into the a Data Center, it MUST support off-line accounting correlation between the WAN MPLS VPN and the cloud/DC MPLS VPN for the tenant. The systems SHOULD be able to bind interface accounting to particular tenant. It MAY requires offline state correlation as well, for example, binding of interface state to tenant.

Provisioning the vPE solution:

##### 1) Provisioning process

- a. The WAN provisioning system periodically provides to the DC orchestration system the VPN tenant and RT context.
- b. DC orchestration system configures vPE on a per request basis

##### 2) Auto state correlation

##### 3) Inter-connection options:

Inter-AS options defined in [RFC4364] may or may not be sufficient for a given inter-connection scenario. BGP IP VPN inter-connection with the Data Center is discussed in [I-D.fang-l3vpn-data-center-interconnect].

This model requires offline accounting correlation

##### 1) Cloud/DC orchestration configures vPE

2) Orchestration initiates WAN IP VPN provisioning; passes connection IDs (e.g., of VLAN/VXLAN) and tenant context to WAN IP VPN provisioning systems.

3) WAN MPLS VPN provisioning system provisions PE VRF and policies as in typical Enterprise IP VPN provisioning processes.

4) Cloud/DC Orchestration system or WAN IP VPN provisioning system

MUST have the knowledge of the connection topology between the DC and WAN, including the particular interfaces on core router and connecting interfaces on the DC PE and/or vPE.

In short, this approach requires off-line accounting correlation and state correlation, and requires per WAN Service Provider integration.

Dynamic BGP sessions between PE/vPE and vCE MAY be used to automate the PE provisioning in the PE-vCE model, that will remove the needs for PE configuration. Caution: This is only under the assumption that the DC provisioning system is trusted and can support dynamic establishment of PE-vCE BGP neighbor relationships, for example, the WAN network and the cloud/DC belong to the same Service Provider.

#### 7.4.2 vPE Pull model

Pull model: pull from network elements to network management/AAA based upon data plane or control plane activity. It supports service activation. This approach is often used in broadband deployments. Dynamic accounting correlation and dynamic state correlation are supported. For example, session based accounting is implicitly includes tenant context state correlation, as well as session-based state that implicitly includes tenant context. Note that the pull model is less common for vPE deployment solutions.

Provisioning process:

- 1) Cloud/DC orchestration configures vPE
- 2) Orchestration primes WAN MPLS VPN provisioning/AAA for new service, passes connection IDs (e.g., VLAN/VXLAN) and tenant context.
- 3) Cloud/DC ASBR detects new VLAN and sends Radius Access-Request (or Diameter Base Protocol request message [RFC6733]).
- 4) Radius Access-Accept (or Diameter Answer) with VRF and other policies

Auto accounting correlation and auto state correlation is supported.

## 8. Security Considerations

As vPE is an extended BGP/MPLS VPN solution, security threats and defense techniques described in RFC 4111 [RFC4111] generally apply.

When the SDN approach is used, the protocols between the vPE agent and the vPE-C in the controller MUST be mutually authenticated. Given the potentially very large scale and the dynamic nature in the cloud/DC environment, the choice of key management mechanisms need to be further studied.

VMs in the servers can belong to different tenants with different characteristics depending on the application. Classification of the VMs must be done through the orchestration system and appropriate security policies must be applied based on such classification before turning on the services.

## 9. IANA Considerations

None.

## 10. Acknowledgments

The authors would like to thank Daniel Voyer for his review and comments.

## 11. References

### 11.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3209] Awduche, D., et al., "RSVP-TE: Extension to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route

Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.

- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.
- [RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, October 2010.
- [RFC6120] Saint-Andre, P., "Extensible Messaging and Presence Protocol (XMPP): Core", RFC 6120, March 2011.
- [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, June 2011.
- [RFC6733] Fajardo, V., Ed., Arkko, J., Loughney, J., and G. Zorn, Ed., "Diameter Base Protocol", RFC 6733, October 2012.

## 11.2 Informative References

- [RFC4111] Fang, L., Ed., "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4111, July 2005.
- [RFC7159] Bray, T., "The JavaScript Object Notation (JSON) Data Interchange Format", RFC 7159, March 2014.
- [RFC4797] Rekhter, Y., Bonica, R., and E. Rosen, "Use of Provider Edge to Provider Edge (PE-PE) Generic Routing Encapsulation (GRE) or IP in BGP/MPLS IP Virtual Private Networks", RFC 4797, January 2007.
- [I-D.ietf-l3vpn-end-system] Marques, P., Fang, L., Pan, P., Shukla, A., Napierala, M., Bitar, N., "BGP-signaled end-system IP/VPNs", draft-ietf-l3vpn-end-system, work in progress.
- [I-D.rfernando-l3vpn-service-chaining] Fernando, R., Rao, D., Fang, L., Napierala, M., So, N., draft-rfernando-l3vpn-service-chaining, work in progress.
- [I-D.fang-l3vpn-virtual-ce] Fang, L., Evans, J., Ward, D., Fernando, R., Mullooly, J., So, N., Bitar, N., Napierala, M., "BGP

IP VPN Virtual PE", draft-fang-l3vpn-virtual-ce, work in progress.

[I-D.ietf-i2rs-architecture] Atlas, A., Halpern, J., Hares, S., Ward, D., and Nadeau, T., "An Architecture for the Interface to the Routing System", draft-ietf-i2rs-architecture, work in progress.

[I-D.ietf-i2rs-problem-statement] Atlas, A., Nadeau, T., and Ward, D., "Interface to the Routing System Problem Statement", draft-ietf-i2rs-problem-statement, work in progress.

[I-D.bitar-i2rs-service-chaining] Bitar, N., Geron, G., Fang, L., Krishnan, R., Leymann, N., Shah, H., Chakrabarti, S., Haddad, W., draft-bitar-i2rs-service-chaining, work in progress.

[I-D.fang-l3vpn-data-center-interconnect] Fang, L., Fernando, R., Rao, D., Boutros, S., "BGP IP VPN Data Center Interconnect", draft-fang-l3vpn-data-center-interconnect, work in progress.

[I-D.ietf-l2vpn-evpn] Sajassi, A., et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn, work in progress.

#### Authors' Addresses

Luyuan Fang  
Microsoft  
5600 148th Ave NE  
Redmond, WA 98052  
Email: lufang@microsoft.com

David Ward  
Cisco  
170 W Tasman Dr  
San Jose, CA 95134  
Email: wardd@cisco.com

Rex Fernando  
Cisco  
170 W Tasman Dr  
San Jose, CA  
Email: rex@cisco.com

Maria Napierala  
AT&T  
200 Laurel Avenue  
Middletown, NJ 07748  
Email: mnapierala@att.com

Nabil Bitar  
Verizon  
40 Sylvan Road  
Waltham, MA 02145  
Email: nabil.bitar@verizon.com

Dhananjaya Rao  
Cisco  
170 W Tasman Dr  
San Jose, CA  
Email: dhrao@cisco.com

Bruno Rijsman  
Juniper Networks  
10 Technology Park Drive  
Westford, MA 01886  
Email: brijsman@juniper.net

Ning So  
Vinci Systems  
Plano, TX 75082, USA  
Email: ning.so@vinci-systems.com

Jim Guichard  
Cisco  
Boxborough, MA 01719  
Email: jguichar@cisco.com

Wen Wang  
CenturyLink  
2355 Dulles Corner Blvd.  
Herndon, VA 20171  
Email: Wen.Wang@CenturyLink.com

Manuel Paul  
Deutsche Telekom  
Winterfeldtstr. 21-27  
10781 Berlin, Germany  
Email: manuel.paul@telekom.de

Wim Henderichx  
Alcatel-Lucent



Email: [wim.henderichx@alcatel-lucent.com](mailto:wim.henderichx@alcatel-lucent.com)

L2VPN

Internet Draft

Intended status: Informational

Expires: January 2015

Weiguo Hao  
Liang Xia  
Shunwan Zhuang  
Huawei  
Vic Liu  
China Mobile  
July 4, 2014

Inter-AS Option B between NVO3 and MPLS EVPN network  
draft-hao-l2vpn-inter-nvo3-evpn-00.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document.

## Abstract

This draft describes option-B inter-as connection between NVO3 network and MPLS EVPN network. Comparing to traditional MPLS EVPN Option-B inter-as connection, this draft provides enhancement for heterogeneous network multi-as connection, the control plane and data plane procedures in NVO3 network are described.

## Table of Contents

1. Introduction .....	2
2. Conventions used in this document.....	3
3. Reference model .....	5
4. Option-A inter-as solution overview.....	6
5. Inter-as option-B routing distribution process.....	7
5.1. Ethernet Tag ID conversion on ASBR.....	7
5.2. Ethernet Auto-Discovery Route process.....	8
5.2.1. Optimized MPLS Label solution on ASBR.....	9
5.3. Ethernet Segment Route process.....	10
5.4. Inclusive Multicast Ethernet Tag Route process.....	10
5.5. MAC/IP advertisement route process .....	10
6. Inter-as option-B data plane procedures .....	12
6.1. Internal DC to external DC direction .....	12
6.2. External DC to internal DC direction .....	12
7. Inter-as option-B solution between PBB-EVPN network and NVO3 network .....	13
8. Security Considerations.....	13
9. IANA Considerations .....	13
10. References .....	13
10.1. Normative References.....	13
10.2. Informative References.....	13
11. Acknowledgments .....	14

## 1. Introduction

In cloud computing era, multi-tenancy has become a core requirement for data centers. Since NVO3 can satisfy multi-tenancy key requirements, this technology is being deployed in an increasing number of cloud data center network. NVO3 focuses on the construction of overlay networks that operate over an IP (L3) underlay transport network. It can provide layer 2 bridging and

layer 3 IP service for each tenant. VXLAN and NVGRE are two typical NVO3 technologies. NVO3 overlay network can be controlled through centralized NVE-NVA architecture or through distributed BGP VPN protocol.

NVO3 has good scaling properties from relatively small networks to networks with several million tenant systems (TSs) and hundreds of thousands of virtual networks within a single administrative domain. In NVO3 network, 24-bit VN ID is used to identify different virtual networks, theoretically 16M virtual networks can be supported in a data center. In a data center network, each tenant may include one or more layer 2 virtual network and in normal cases each tenant corresponds to one routing domain (RD). Normally each layer 2 virtual network corresponds to one or more subnets.

To provide cloud service to external data center client, data center networks should be connected with WAN networks. BGP MPLS based Ethernet VPNs(EVPN)[EVPN] is being deployed in an increasing number of WAN networks. If EVPN CEs in external DC and TSs in internal DC belong to same subnet of same tenant, they are in same broadcast domain and can freely layer 2 communicate with each other in the broadcast domain.

Normally internal data center and external EVPN network belongs to different autonomous system(AS). This requires the setting up of inter-as connections at Autonomous System Border Routers(ASBRs) between NVO3 network and external EVPN network.

Currently, a typical connection mechanism between a data center network and an MPLS EVPN network is similar to Inter-AS Option-A of RFC4364, but it has scalability issue if there is huge number of tenants in data center networks. To overcome the issue, inter-as Option-B between NVO3 network and BGP MPLS EVPN network is proposed in this draft.

## 2. Conventions used in this document

EVI - An EVPN instance spanning across the PEs participating in that EVPN.

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

Network Virtualization Edge (NVE) - An NVE is the network entity that sits at the edge of an underlay network and implements network virtualization functions.

Tenant System - A physical or virtual system that can play the role of a host, or a forwarding element such as a router, switch, firewall, etc. It belongs to a single tenant and connects to one or more VNs of that tenant.

VN - A VN is a logical abstraction of a physical network that provides L2 network services to a set of Tenant Systems.

RD - Route Distinguisher. RDs are used to maintain uniqueness among identical routes in different MAC-VRFs, The route distinguisher is an 8-octet field prefixed to the customer's MAC address. The resulting 12-octet field is a unique "VPN-MAC" address.

RT - Route targets. It is used to control the import and export of routes between different MAC-VRFs.

## 3. Reference model

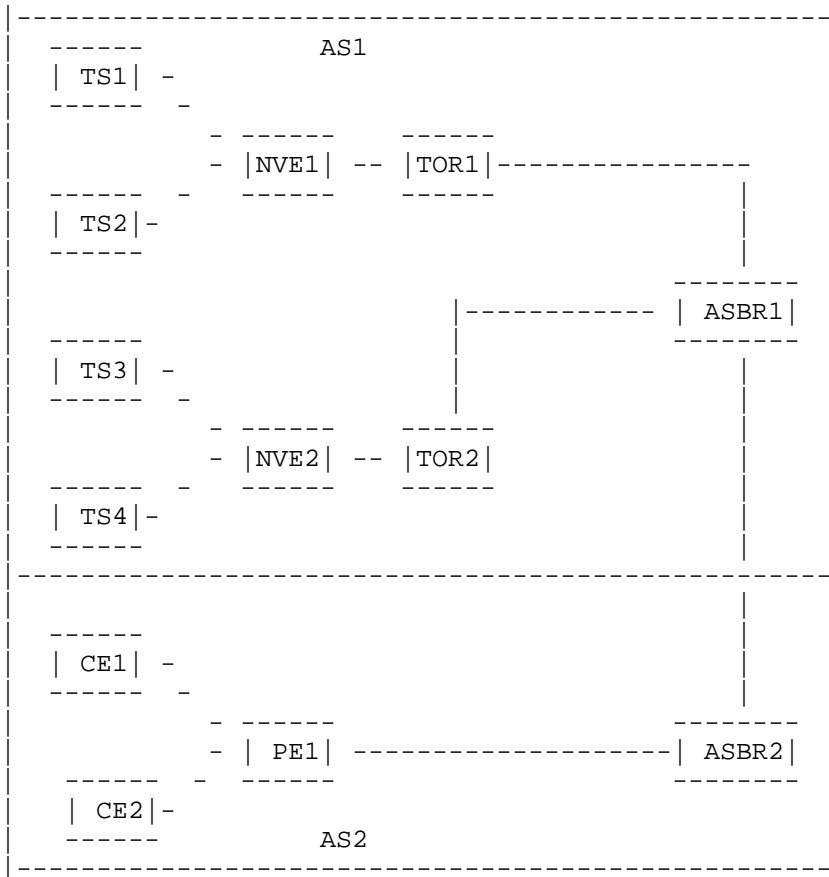


Figure 1 Reference model

Figure 1 shows an arbitrary Multi-AS VPN interconnectivity scenario between NVO3 network and MPLS EVPN network. NVE1, NVE2, and ASBR1 forms NVO3 overlay network in internal DC. TS1 and TS2 connect to NVE1, TS3 and TS4 connect to NVE2. PE1 and ASBR2 forms MPLS EVPN network in external DC. CE1 and CE2 connect to PE1. The NVO3 network belongs to AS 1, the MPLS EVPN network belongs to AS 2.

There are two tenants of tenant 1 and tenant 2 in NVO3 network. MAC-VRF 1 and MAC-VRF 2 are created on NVE1 and NVE2 for these two tenants. TSs(TS1 and TS3) in tenant 1 and CE(CE1) in VPN-Red belongs to same subnet and broadcast domain, TSs(TS2 and TS4) in tenant 2

and CE(CE2) in VPN-Green belongs to same subnet and broadcast domain. The TSs and CEs in same broadcast domain can freely layer 2 communicate with each other.

The TS and CE information in above figure 1 are as follows:

TS	Tenant	IP Address	MAC	VN ID
TS1	1	10.1.1.2	MAC1	10
TS2	2	20.1.1.2	MAC2	20
TS3	1	10.1.1.3	MAC3	10
TS4	2	20.1.1.3	MAC4	20

Table 1 TS information in NVO3 network

CE	Route Distinguisher	Route Target	IP Address	MAC
CE1	VPN-Red1	1:1	10.1.1.4	MAC5
CE2	VPN-Green1	2:2	20.1.1.4	MAC6

Table 2 CE information in MPLS/IP VPN network

#### 4. Option-A inter-as solution overview

In Option-A inter-as solution, peering ASBRs are connected by multiple sub-interfaces, each ASBR acts as a PE, and thinks that the other ASBR is a CE. EVPN instances are configured at AS border routers (ASBR1 and ASBR2) so that each ASBRs associate each such sub-interface with a EVPN instance. It requires MAC look up on ASBRs. MAC address propagation for each EVPN instance between ASBR1 and ASBR2 relies on data plane learning mechanism. In the data-plane, VLANs are used for VPN traffic separation.

Option-A inter-as solution has following issues:

1. Up to 16 million (16M) sub-interfaces need to exist between the ASBRs, if there are 16M VN in NVO3 network.
2. UP to 16M EVPN instances need to be supported on border routers.

3. Several million MAC routing entries need to be supported on border routers.

Inter-as option B between NVO3 network and MPLS EVPN network can be used to address these issues. Due to it is for multi-as interconnection between heterogeneous networks, so there are some differences from traditional homogenous EVPN Inter-AS Option-B.

## 5. Inter-as option-B routing distribution process

In option-B inter-as solution, an EBGP session is used to distribute labeled EVPN NLRI between the ASBRs. The advantage of this option is that it's more scalable, as there is no need to have one sub-interface per VPN between ASBRs.

There are four Route Types in EVPN NLRI, they are Ethernet Segment Route, Ethernet Auto-Discovery Route, Inclusive Multicast Ethernet Tag Route and MAC/IP advertisement route which are used for Designated Forwarder Election, fast Convergence and aliasing or backup-path, multicast traffic handling and unicast traffic handling respectively.

For inter-as option-B interconnection between EVPN and NVO3 network, When a ASBR receives BGP update message carrying the routes from peer PEs or NVEs in local AS, it should re-constructs these message and advertises it to peer ASBR, the Next Hop field of the MP\_REACH\_NLRI attribute should be set to a routable IP address of the ASBR. When the peer ASBR receives the message, the ASBR also should re-construct the message and advertise it to peer PEs or NVEs in its local AS, the Next Hop field of the MP\_REACH\_NLRI attribute also should be set to a routable IP address of the ASBR.

In NVO3 network, there are two options for mapping the VNI to an EVI [EVPN-OVERLAY], one is single subnet per EVI, and another one is multiple subnets per EVI.

In the following subsection, a detail explanation will be given on how to re-construct EVPN update message and how to generate incoming and outgoing forwarding table on ASBR.

### 5.1. Ethernet Tag ID conversion on ASBR

A broadcast domain can be identified by different Ethernet Tag ID in NVO3 and MPLS EVPN network. The Ethernet Tag ID mapping relationship between NVO3 and MPLS EVPN network should be configured on each ASBR in beforehand. For example, VLAN 10 in EVPN network and VN 100 in



NVO3 network belong to same broadcast domain, NVO3 network uses 100 as Ethernet Tag ID, EVPN network uses 10 as Ethernet Tag ID.

When a ASBR receives BGP update message carrying EVPN NLRI from peer ASBR, it should replace Ethernet Tag ID field with local corresponding value and then advertise the message to peer PEs or NVEs in its local AS.

## 5.2. Ethernet Auto-Discovery Route process

There are two Ethernet A-D route types, one is per ES route, and another one is per EVI. The "ESI Label Extended Community" MUST be included in the route, it is to indicate ES's redundancy mode and to advertise ESI Label for split-horizon filtering.

When an ASBR in NVO3 network receives Ethernet A-D per ES route, the ASBR learns a ES and multi-homed NVEs correspondence, the ES's redundancy mode. If "Single-Active" flag in "ESI Label Extended Community" is set, the ES is operating in Single-Active redundancy mode. Otherwise, it is operating in All-Active redundancy mode. The Ethernet A-D per EVI route can be used for Aliasing and Backup-Path, aliasing is used for all-active mode, backup-path is for single-active mode.

When an ASBR in NVO3 network receives Ethernet A-D per EVI route, the ASBR should allocate new MPLS Label and advertises it to all peer ASBRs in Ethernet Auto-Discovery Route MPLS Label field. In NVO3 network, the MPLS Label allocation principle is: If ESI is 0, MPLS label is allocated per NVE per VN(This is single-homed case). Otherwise, MPLS label is allocated per ESI per VN(This is multi-homed case). MPLS VPN Label and <remote NVE,VN ID> correspondence is used to generate incoming forwarding table on each ASBR, traffic forwarding from external to internal DC direction relies on the incoming forwarding table.

In multi-homed scenario, when an ESI occurs link failure and lost connection with a NVE, the NVE should trigger ASBR in its local AS to mass update its local forwarding table by Ethernet A-D per ES route. This is called fast convergence procedure. The ASBR doesn't need re-allocate MPLS Label for each VN on the ESI and advertise to peer AS, i.e., fast convergence process is restricted to local AS, the Ethernet A-D route per ES doesn't need to be transmitted to peer AS.

For aliasing and backup-path procedures, these procedures also don't need cross different AS domain, they are only restricted in local AS,

each ASBR in local AS needs to process Ethernet A-D per EVI route from PEs or NVEs in local AS for these procedures.

In aliasing case, when a ASBR in NVO3 network receives traffic data from external DC to external DC, the traffic will be forwarded to all-active remote NVEs in load balancing mode. For each aliasing ES and VN, there is a corresponding incoming forwarding table item which includes one MPLS Label and multiple <NVE,VN ID> pairs on ASBR, the NVE is a member of remote multi-homed NVEs attaching the aliasing ES.

In backup-path case, for each backup-path ES and VN, there is a corresponding incoming forwarding table item which includes one MPLS Label and one <NVE,VN ID> on ASBR, the NVE is primary NVE that advertises the MAC/IP advertisement route in the VN. When a ASBR receives first MAC/IP advertisement route from remote primary NVE, it will know the primary NVE and generate the incoming forwarding table item.

#### 5.2.1. Optimized MPLS Label solution on ASBR

MPLS Label consumption on ASBR is high through the above per ESI per VN solution, the optimized allocation solution is provided as follows:

If multiple ESIs are operating in all-active mode and attached to the exact same set of NVEs, then these ESIs can share same MPLS Label for same VN to save MPLS Label space on ASBR.

If multiple ESIs are operating in single-active mode and attached to the exact same set of NVEs, primary NVEs for these ESI are same NVE, then the VNs on these ESIs can share same MPLS Label for same VN to save MPLS VPN Label space on ASBR.

In this case, if a ESI occurs link failure and lost connection with a NVE, the NVE advertises Ethernet Auto-Discovery Route per ES to each ASBR in its local AS, the ASBRs knows that the ESI is attached to a different set of NVEs, it should re-allocate new MPLS Labels for each VN on the ESI, mass update its incoming forwarding table, then advertise these MPLS Labels using Ethernet Auto-Discovery route per EVI to peer ASBR.

When peer ASBR receives the Ethernet Auto-Discovery route per EVI, it allocates new MPLS Label and replaces the value in Ethernet Auto-Discovery Route MPLS Label field, then advertises it to all peer PEs.

Remote PEs in peer AS should update all its MAC entries with the new MPLS Label associated with the ESI and EVI.

### 5.3. Ethernet Segment Route process

Due to this route is used for DF election and multi-homed PE or NVE devices won't straddle between MPLS EVPN and NVO3 network, so when a ASBR receives BGP update message carrying the route from peer PE or NVE in its own AS, it just removes it from the message, the route don't need to be transmitted to peer AS.

### 5.4. Inclusive Multicast Ethernet Tag Route process

Similar to regular EVPN inter-as solution, when a ASBR receives from one of its IBGP neighbors a BGP Update message that carries the route, it re-advertises it to peer ASBRs and these peer ASBRs re-advertise it to peer PEs or NVEs in its local AS. The re-advertised routes MUST be the same as the original ones, except for the PMSI Tunnel attribute in Inclusive Multicast Ethernet Tag Route and Ethernet Tag ID. If ingress replication is used between ASBRs, the Tunnel Type in PMSI Tunnel attribute is set to Ingress Replication, the Leaf Information Required flag is set to 1, the PMSI Tunnel attribute carries no MPLS labels.

### 5.5. MAC/IP advertisement route process

Because the ASBR in NVO3 network has already assigned MPLS Label for each ESI(or NVE in single-homed case) and each VN when it received Ethernet Auto-Discovery Route from remote NVEs in its local AS, so the ASBR receives first MAC/IP advertisement route from a <ES,VN>, it will search the already assigned MPLS Label for the <ES,VN>, generate a incoming forwarding item, fuel MPLS Label field in the MAC/IP advertisement route, and then send it to peer ASBR. The incoming forwarding table is used for traffic forwarding from external DC to internal DC direction.

In above figure 1, all TSs are single-homed to a NVE, MPLS VPN Label is assigned per NVE per VN, the incoming forwarding table on ASBR in NVO3 network is as follows:

MPLS VPN Label	NVE + VN ID
1000	NVE1 + 10
2000	NVE1 + 20
1001	NVE2 + 10
2001	NVE2 + 20

Incoming forwarding table

When ASBR1 in NVO3 network receives from EBGp neighbors ASBR2 a BGP Update message that carries MAC/IP advertisement route, it should allocate VN ID per MPLS VPN Label, generate outgoing forwarding table, and then advertises it to peer NVEs in its local AS.

In above figure 1, ASBR1 allocates VN ID for each VPN Label receiving from ASBR2, and then ASBR2 advertises the MAC/IP advertisement route with new allocated VN ID to each NVE (NVE1 and NVE2). The role of the VN ID is similar to the role of In VPN Label in ASBR1, it has local significance on ASBR1, each VN ID corresponds to each MPLS VPN Label on ASBR2; The VN ID space should be assigned in beforehand and should be orthogonal to the VN ID space for tenant identification(for example, assuming ASBR1 has local connecting TSs of tenant 1 to 100, VN ID 1 to 100 are allocated for these tenants, other VN ID other than 1 to 100 can be allocated for outgoing forwarding table purpose). The allocated VN ID and its corresponding out VPN Label forms an outgoing forwarding table which is used to forward NVO3 traffic from internal DC to external DC. The outgoing forwarding table on ASBR1 is as follows:

VN ID	Out VPN Label
10000	3000
10001	4000

Outgoing forwarding table

## 6. Inter-as option-B data plane procedures

This section describes the step by step procedures of data forward for either: a) internal DC to external DC data flows, or b) the external DC to internal DC data flows.

### 6.1. Internal DC to external DC direction

1. TS1 sends traffic to NVE1, the destination MAC is CE1's MAC address of MAC5.
2. NVE1 looks up MAC-VRF 1's MAC forwarding table, then it gets NVO3 tunnel encapsulation information. The destination outer address is ASBR1's IP address, VN ID is 10000. NVE1 performs NVO3 encapsulation and sends the traffic to ASBR1.
3. ASBR1 decapsulates NVO3 encapsulation and gets VN ID 10000. Then it looks up outgoing forwarding table based on the VN ID and gets MPLS VPN label 3000. Finally it pushes MPLS VPN label for the IP traffic and sends it to ASBR2.
4. ASBR2 swaps VPN Label, then sends the traffic to PE1 through IGP tunnel.
5. PE1 terminates IGP tunnel, pops MPLS VPN label 3000, looks up local MAC-VRF 1, and then forwards the traffic to CE1.

### 6.2. External DC to internal DC direction

1. CE1 sends traffic to PE1, destination MAC is TS1's MAC address of MAC1.
2. PE1 looks up local MAC forwarding table in VPN-RED, pushes MPLS VPN label 1000, then searches IGP tunnel, then the PE sends the traffic to ASBR2 through IGP tunnel.
3. ASBR2 terminates IGP tunnel, swaps MPLS VPN label, then sends the traffic to ASBR1.
4. ASBR1 looks up incoming forwarding table and gets NVO3 encapsulation, then performs NVO3 encapsulation and sends the traffic to NVE1. The destination outer IP is NVE1's IP, VN ID is 10.
5. NVE1 decapsulates NVO3 encapsulation, gets local EVPN instance 1 relying on VN ID 10, looks up local MAC-VRF 1, then sends the traffic to TS1.

## 7. Inter-as option-B solution between PBB-EVPN network and NVO3 network

For the further study.

## 8. Security Considerations

Similar to the security considerations for inter-as Option-B in [RFC4364] the appropriate trust relationship must exist between NVO3 network and MPLS EVPN network. EVPN routes in NVO3 network should neither be distributed to nor accepted from the public Internet, or from any BGP peers that are not trusted. For other general VPN Security Considerations, see [RFC4364].

## 9. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

## 10. References

### 10.1. Normative References

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

### 10.2. Informative References

- [1] [RFC4364] E. Rosen, Y. Rekhter, " BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [2] [EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-02.txt, work in progress, February, 2012.
- [3] [NVA] D.Black, etc, "An Architecture for Overlay Networks (NVO3)", draft-ietf-nvo3-arch-01, February 14, 2014
- [4] [EVPN-OVERLAY] A. Sajassi, etc, "'A Network Virtualization Overlay Solution using EVPN'", draft-sd-l2vpn-evpn-overlay-03, June, 2014
- [5] [NOV3-FRWK] Lasserre et al., "Framework for DC Network Virtualization", draft-ietf-nvo3-framework-01.txt, work in progress, October 2012.
- [6] [NVGRE] Sridhavan, M., et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-01.txt, July 8, 2012.

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

## 11. Acknowledgments

Authors like to thank Junlin Zhang for his valuable inputs.

### Authors' Addresses

Weiguo Hao  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012  
China  
Email: haoweiguo@huawei.com

Liang Xia (Frank)  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012  
China  
Email: frank.xialiang@huawei.com

Shunwan Zhuang  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China  
Email: zhuangshunwan@huawei.com

Vic Liu  
China Mobile  
32 Xuanwumen West Ave, Beijing, China  
Email: liuzhiheng@chinamobile.com





Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 4, 2015

K. Patel  
S. Boutros  
J. Liste  
Cisco Systems  
B. Wen  
Comcast  
J. Rabadan  
Alcatel-Lucent  
July 3, 2014

Extensions to BGP Signaled Pseudowires to support Flow-Aware Transport  
Labels  
draft-keyupate-l2vpn-fat-pw-bgp-01.txt

Abstract

[RFC6391] describes a mechanism that uses an additional label (Flow Label) in the MPLS label stack that allows Label Switch Routers to balance flows within Pseudowires at a finer granularity than the individual Pseudowires across the Equal Cost Multiple Paths (ECMPs) that exists within the Packet Switched Network (PSN).

Furthermore, [RFC6391] defines the LDP protocol extensions required to synchronize the flow label states between the ingress and egress PEs when using the signaling procedures defined in the [RFC4447].

This draft defines protocol extensions required to synchronize flow label states among PEs when using the BGP-based signaling procedures defined in [RFC4761]. These protocol extensions are equally applicable to point-to-point L2VPNs defined in [RFC6624].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 4, 2015.

## Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

## Table of Contents

1. Introduction . . . . .	2
1.1. Requirements Language . . . . .	3
2. Modifications to Layer 2 Info Extended Community . . . . .	3
3. Signaling the Presence of the Flow Label . . . . .	5
4. Acknowledgements . . . . .	6
5. Contributors . . . . .	6
6. IANA Considerations . . . . .	6
7. Security Considerations . . . . .	6
8. References . . . . .	6
8.1. Normative References . . . . .	6
8.2. Informative References . . . . .	7
Authors' Addresses . . . . .	7

## 1. Introduction

A pseudowire (PW)[RFC3985] is normally transported over one single network path, even if multiple Equal Cost Multiple Paths (ECMPs) exist between the ingress and egress PW provider edge (PE) equipment. This is required to preserve the characteristics of the emulated

service. The use of a single path to preserve the packet delivery order remains the default mode of operation of a PW and is described in [RFC4385], [RFC4928].

Using the principles defined in [RFC6391], this draft augments the BGP-signaling procedures of [RFC4761] and [RFC6624] to allow an OPTIONAL mode that may be employed when the use of ECMPs is known to be beneficial to the operation of the PW.

High bandwidth Ethernet-based services are a prime example that benefits from the ability to load-balance flows in a PW over multiple PSN paths. In general, load-balancing is applicable when the PW attachment circuit bandwidth and PSN core link bandwidth are of same order of magnitude.

To achieve the load-balancing goal, [RFC6391] introduces the notion of an additional Label Stack Entry (LSE) (Flow label) located at the bottom of the stack (right after PW LSE). Label Switching Routers (LSRs) commonly generate a hash of the label stack in order to discriminate and distribute flows over available ECMPs. The presence of the Flow label (closely associated to a flow determined by the ingress PE) will normally provide the greatest entropy.

Furthermore, following the procedures for Inter-AS scenarios described in [RFC4761] section 3.4, the Flow label should never be handled by the ASBRs, only the terminating PEs on each AS will be responsible for popping or pushing this label. This is equally applicable to Method B [section 3.4.2] of [RFC4761] where ASBRs are responsible for swapping the PW label as traffic traverses from ASBR to PE and ASBR to ASBR directions. Therefore, the Flow label will remain untouched across AS boundaries.

#### 1.1. Requirements Language

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

#### 2. Modifications to Layer 2 Info Extended Community

The Layer 2 Info Extended Community is used to signal control information about the pseudowires to be setup. The extended community format is described in [RFC4761]. The format of this extended community is described as:

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

#### Layer 2 Info Extended Community

##### Control Flags:

This field contains bit flags relating to the control information about pseudowires. This field is augmented with a definition of 2 new flags field.

0 1 2 3 4 5 6 7	
Z Z Z T R Z C S	(Z = MUST Be Zero)

##### Control Flags Bit Vector

With Reference to the Control Flags Bit Vector, the following bits in the Control Flags are defined; the remaining bits, designated Z, MUST be set to zero when sending and MUST be ignored when receiving this Extended Community.

- Z Must be set to Zero.
- T When the bit value is 1, the PE is requesting the ability to send a Pseudowire packet that includes a flow label. When the bit value is 0, the PE is indicating that it will not send a Pseudowire packet containing a flow label.
- R When the bit value is 1, the PE is able to receive a Pseudowire packet with a flow label present. When the bit value is 0, the PE is unable to receive a Pseudowire packet with the flow label present.
- C Defined in [RFC4761].
- S Defined in [RFC4761].

### 3. Signaling the Presence of the Flow Label

As part of the Pseudowire signaling procedures described in [RFC4761], a Layer 2 Info Extended Community is advertised in the VPLS BGP NLRI. This draft recommends that the Control Flags field of this extended community be used to synchronize the flow label states amongst PEs for a given L2VPN.

A PE that wishes to send a flow label in a Pseudowire packet MUST include in its VPLS BGP NLRI a Layer 2 Info Extended Community using Control Flags field with T = 1.

A PE that is willing to receive a flow label in a Pseudowire packet MUST include in its VPLS BGP NLRI a Layer 2 Info Extended Community using Control Flags field with R = 1.

A PE that receives a VPLS BGP NLRI containing a Layer 2 Info Extended Community with R = 0 MUST NOT include a flow label in the Pseudowire packet.

Therefore, a PE sending a Control Flags field with T = 1 and receiving a Control Flags field with R = 1 MUST include a flow label in the Pseudowire packet. Under all other combinations, a PE MUST NOT include a flow label in the Pseudowire packet.

A PE MAY support the configuration of the flow label (T and R bits) on a per-service (e.g. VPLS VFI) basis. Furthermore, it is also possible that on a given service, PEs may not share the same flow label settings. The presence of a flow label is therefore determined on a per-peer basis and according to the local and remote T and R bit values. For example, a PE part of a VPLS and with a local T = 1, must only transmit traffic with a flow label to those peers that signaled R = 1. And if the same PE has local R = 1, it must only expect to receive traffic with a flow label from peers with T = 1. Any other traffic MUST not have a flow label.

Modification of flow label settings may impact traffic over a PW as these could trigger changes in the PEs data-plane programming (i.e. imposition / disposition of flow label). This is an implementation specific behavior and outside the scope of this draft

The signaling procedures in [RFC4761] state that the unspecified bits in the Control Flags field (bits 0-5) MUST be set to zero when sending and MUST be ignored when receiving. The signaling procedure described here is therefore backwards compatible with existing implementations. A PE not supporting the extensions described in this draft will always advertise a value of ZERO in the position assigned by this draft to the R bit and therefore a flow label will

never be included in a packet sent to it by one of its peers. Similarly, it will always advertise a value of ZERO in the position assigned by this draft to the T bit and therefore a peer will know that a flow label will never be included in a packet sent by it.

Note that what is signaled is the desire to include the flow LSE in the label stack. The value of the flow label is a local matter for the ingress PE, and the label value itself is not signaled.

#### 4. Acknowledgements

The authors would like to thank Bertrand Duvivier and John Drake for their review and comments.

#### 5. Contributors

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

Eric Lent

John Brzozowski

Steven Cotter

#### 6. IANA Considerations

#### 7. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4271].

#### 8. References

##### 8.1. Normative References

- [I-D.ietf-l2vpn-vpls-multihoming]  
Kothari, B., Kompella, K., Henderickx, W., Balus, F., Uttaro, J., Palislaamovic, S., and W. Lin, "BGP based Multi-homing in Virtual Private LAN Service", draft-ietf-l2vpn-vpls-multihoming-06 (work in progress), July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

## 8.2. Informative References

- [RFC2842] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.
- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.
- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, May 2012.

## Authors' Addresses

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: keyupate@cisco.com

Sami Boutros  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: sboutros@cisco.com

Jose Liste  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: jliste@cisco.com

Bin Wen  
Comcast  
1701 John F Kennedy Blvd  
Philadelphia, PA 19103  
USA

Email: bin\_wen@cable.comcast.com

Jorge Rabadan  
Alcatel-Lucent  
777 E. Middlefield Road  
Mountain View, CA 94043  
USA

Email: jorge.rabadan@alcatel-lucent.com



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 5, 2015

Z. Li  
J. Zhang  
Z. Zhuang  
Huawei Technologies  
July 04, 2014

Using BGP between PE and CE in EVPN  
draft-li-l2vpn-evpn-pe-ce-01

Abstract

This document specifies protocols and procedures of using BGP as PE-CE control protocol for carrying customer MAC routing information.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Terminology . . . . .	2
3. Application Scenarios . . . . .	2
4. BGP EVPN NLRI Extensions . . . . .	4
5. Exchanging C-MAC Routes . . . . .	4
5.1. Originating MAC Route at the CE router . . . . .	5
5.2. Receiving a MAC Route by the PE router . . . . .	6
6. IANA Considerations . . . . .	6
7. Security Considerations . . . . .	6
8. Normative References . . . . .	6
Authors' Addresses . . . . .	7

## 1. Introduction

[I-D.ietf-l2vpn-evpn] describes protocols and procedures for BGP MPLS based Ethernet VPNs. BGP is used for MAC learning by exchanging customer MAC routing information between PEs in the control plane instead of MAC learning between PEs in the data plane. It also states that MAC learning between PEs and CEs MAY be done in the control plane, but it does not define the detailed protocols and procedures. This document specifies protocols and procedures of using BGP as PE-CE control protocol for carrying customer MAC routing information. This can provide some benefits such as fast convergence in some situation.

## 2. Terminology

This document uses terminology described in [I-D.ietf-l2vpn-evpn].

## 3. Application Scenarios

There are some benefits when control plane is introduced between PE and CE in EVPN network. The following illustrates the benefits with an example of fast convergence in the event of PE to CE network failure.

[I-D.ietf-l2vpn-evpn] defines a mechanism to efficiently and quickly signal, to remote PE nodes, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet Segment. This mechanism optimizes the withdrawal of MAC Advertisement routes, and then optimizes the network convergence time

in the event of PE to CE failures. But it still cannot fully provide convergence time that is independent of the number of MAC addresses learned by the PE. There exist a situation where the network convergence time is dependent on the local MAC learning of PE and the advertisement of them to remote PE.

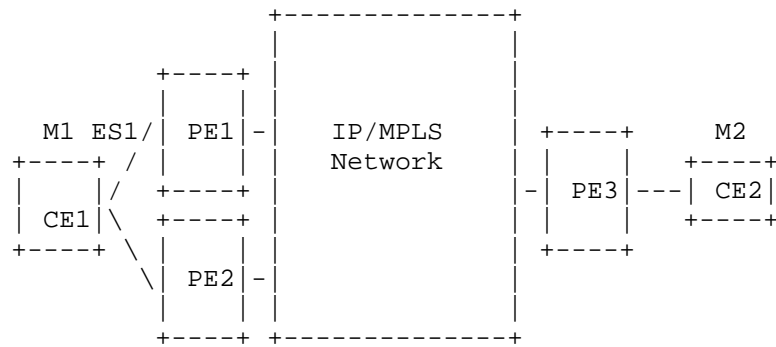


Figure 1 Multi-homed EVPN Network

To illustrate this with an example in the Figure 1, consider two PEs (PE1 and PE2) connected to a multi-homed Ethernet Segment ES1. All-Active redundancy mode is assumed. A given MAC address M1 is learned by PE1 but not PE2. On PE3, the following states may arise:

T0- PE3 receives the Ethernet A-D routes per ESI from PE1 and PE2.

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

T2- After T1, when the ES1 connected to PE1 fails, PE1 MUST withdraw its Ethernet A-D route per ESI, then PE3 forwards traffic destined to M1 to PE2 only.

T3- After T2, PE1 MUST also withdraw the MAC Advertisement routes (M1) that are impacted by the failure. Before PE2 learns M1 and advertises a MAC Advertisement route for M1, PE3 will treat traffic to M1 as unknown unicast. If the behavior is to drop the unknown unicast based on the administrative policy, the traffic to M1 on PE3 will be interrupted. Until PE2 has also advertised a MAC Advertisement route for M1 before PE1 withdraws its MAC route, then PE3 would have continued forwarding traffic destined to M1.

In the above example, once the local MAC learning of PE was done via control plane, both PE1 and PE2 will advertise a MAC Advertisement route for M1, then PE3 could continue forwarding traffic destined to M1 in the event of ES1 connected to PE1 or PE2 fails. In this case,

the network convergence time is not dependent of the local MAC learning and advertisement of MAC addresses learned by the PE any more.

The benefit can also be achieved in case of single-active redundancy mode.

#### 4. BGP EVPN NLRI Extensions

A new route type is defined for EVPN NLRI to advertise customer MAC route between PE and CE in EVPN:

+ 6 - Customer MAC Advertisement route

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

```

+-----+
| 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)           |
+-----+

```

It should be noted that the Route Distinguisher (RD) is not used since the customer MAC routes are always exchanged in the context of unawareness of Ethernet VPN.

Another solution option is to reuse EVPN MAC Advertisement Route defined in [I-D.ietf-l2vpn-evpn] to exchange MAC route information between CE and PE. In this case RD, MPLS Label1 and MPLS Label2 fields SHOULD be set as 0. In addition, the RT for the route SHOULD also be set as 0.

#### 5. Exchanging C-MAC Routes

This section describes the procedures of exchanging customer MAC routes between PE and CE. This document assumes that a CE and a PE exchange MAC routes over a direct BGP session.

### 5.1. Originating MAC Route at the CE router

When a CE receives packets in a given VLAN from interfaces, other than interfaces connected to the PE, it learns MAC addresses in the data plane. If the given VLAN is in the setting of VLANs across the Ethernet links attached to a given PE, the CE MAY advertise the MAC addresses it learns in the data plane to the given PE, using MP-BGP and the specific MAC Route, in the control plane. The MAC Route is constructed as follows:

- + The field of the Ethernet Segment Identifier is reserved for future use.
- + The Ethernet Tag ID is set to the VLAN ID from which the MAC addresses are learned.
- + The MAC address length field is in bits and it is typically set to 48. However this specification enables specifying the MAC address as a prefix; in which case, the MAC address length field is set to the length of the prefix. This provides the ability to aggregate MAC addresses if the deployment environment supports that.
- + The MAC address is set to the value of MAC address the CE learned. The encoding of a MAC address MUST be the 6-octet MAC address specified by [802.1D-ORIG] [802.1D-REV]. If the MAC address is advertised as a prefix then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as 6 octets.
- + The IP Address field is optional. By default, the IP Address Length field is set to 0 and the IP Address field is omitted from the route. When a valid IP address or address prefix needs to be advertised (e.g., for ARP suppression purposes or for inter-subnet switching), it is then encoded in this route. In this case, the IP Address Length field is in bits and it is the length of the IP prefix. This provides the ability to advertise IP address prefixes when the deployment environment supports that.
- + The encoding of an IP Address MUST be either 4 octets for IPv4 or 16 octets for IPv6. When the IP Address is advertised as a prefix, then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as either 4 or 16 octets. The length field of Ethernet NLRI is sufficient to determine whether an IP address/prefix is encoded in this route and if so, whether the encoded IP address/prefix is IPv4 or IPv6.
- + The Next Hop field of the MP\_REACH\_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising CE.

It should be noted that the BGP advertisement for the MAC route does not need to carry the Route Target (RT) attributes because of its unawareness of Ethernet VPN.

#### 5.2. Receiving a MAC Route by the PE router

When a PE receives a MAC route from a CE, it learns the MAC addresses advertised in the MAC route in the control plane and associates the MAC addresses with the Ethernet Segment from which it can reach to the advertising CE and the VLAN carried in the MAC route.

The PE SHOULD install forwarding state for the associated MAC addresses based on the Ethernet Segment and VLAN inferred from the MAC route.

In addition, the PE SHOULD advertise the MAC addresses it learns from CE in the control plane, to all the other PEs in the associated EVPN instance, using MP-BGP and the MAC Advertisement route defined in [I-D.ietf-l2vpn-evpn]. For example, the PE learns a MAC address M1 on a multi-homed Ethernet Segment (ES1) and on a VLAN 10, and the VLAN 10 is bundled to EVPN A. The PE SHOULD advertise the MAC address M1 to all the other PEs in EVPN A.

The construction of the MAC Advertisement route and procedures of handling the MAC Advertisement route on receiving it are specified in [I-D.ietf-l2vpn-evpn].

#### 6. IANA Considerations

This document requires IANA to assign a new route type value for EVPN NLRI.

#### 7. Security Considerations

There are no additional security aspects beyond those of EVPN ([I-D.ietf-l2vpn-evpn]).

#### 8. Normative References

- [I-D.ietf-l2vpn-evpn]  
Sajassi, A., Aggarwal, R., Bitar, N., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-07 (work in progress), May 2014.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

Authors' Addresses

Zhenbin Li  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com

Junlin Zhang  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: jackey.zhang@huawei.com

Shunwan Zhuang  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: zhuangshunwan@huawei.com

L2VPN Workgroup  
Internet Draft  
Intended status: Standards Track

R. Shekhar  
N. Sheth  
W. Lin  
Juniper

J. Rabadan  
S. Sathappan  
W. Henderickx  
Alcatel-Lucent

M. Katiyar  
Nuage Networks

Expires: January 5, 2015

July 4, 2014

Optimized Ingress Replication solution for EVPN  
draft-rabadan-l2vpn-evpn-optimized-ir-00

Abstract

Network Virtualization Overlay (NVO) networks using EVPN as control plane may use ingress replication (IR) or PIM-based trees to convey the overlay multicast traffic. PIM provides an efficient solution to avoid sending multiple copies of the same packet over the same physical link, however it may not always be deployed in the NVO core network. IR avoids the dependency on PIM in the NVO network core. While IR provides a simple multicast transport, some NVO networks with demanding multicast applications require a more efficient solution without PIM in the core. This document describes a solution to optimize the efficiency of IR in NVO networks.

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at



<http://www.ietf.org/ietf/lid-abstracts.txt>

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

This Internet-Draft will expire on January 5, 2015.

#### Copyright Notice

Copyright (c) 2014 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. Problem Statement . . . . .	3
2. Solution requirements . . . . .	4
3. EVPN BGP Attributes for optimized-IR . . . . .	4
4. Assisted-Replication (AR) Solution Description . . . . .	6
4.1. AR roles and control plane . . . . .	7
4.1.1. AR-REPLICATOR procedures . . . . .	7
4.1.2. AR-LEAF procedures . . . . .	8
4.1.3. RNVE procedures . . . . .	10
4.2. Multi-destination traffic forwarding behavior in AR EVIs . . . . .	10
4.2.1. Broadcast and Multicast forwarding behavior . . . . .	10
4.2.1.1. REPLICATOR BM forwarding . . . . .	10
4.2.1.2. LEAF BM forwarding . . . . .	11
4.2.1.3. RNVE BM forwarding . . . . .	11
4.2.2. Unknown unicast forwarding behavior . . . . .	12
4.2.2.1. REPLICATOR/LEAF Unknown unicast forwarding . . . . .	12
4.2.2.2. RNVE Unknown unicast forwarding . . . . .	12
5. Pruned-Flood-Lists (PFL) . . . . .	12
6. An example use-case . . . . .	13
5. Benefits of the optimized-IR solution . . . . .	14
6. Conventions used in this document . . . . .	14
7. Security Considerations . . . . .	14
8. IANA Considerations . . . . .	14

8. References	14
9. Acknowledgments	15
10. Authors' Addresses	15

## 1. Problem Statement

EVPN may be used as the control plane for a Network Virtualization Overlay (NVO) network. Network Virtualization Edge (NVE) devices and PEs that are part of the same EVI use Ingress Replication (IR) or PIM-based trees to transport the tenant's multicast traffic. In NVO networks where PIM-based trees cannot be used, IR is the only alternative. Examples of these situations are NVO networks where the core nodes don't support PIM or the network operator does not want to run PIM in the core.

In some use-cases, the amount of replication for BUM (Broadcast, Unknown unicast and Multicast traffic) is kept under control on the NVEs due to the following fairly common assumptions:

- a) Broadcast is greatly reduced due to the proxy-ARP and proxy-ND capabilities supported by EVPN on the NVEs. Some NVEs can even provide DHCP-server functions for the attached Tenant Systems (TS) reducing the broadcast even further.
- b) Unknown unicast traffic is greatly reduced in virtualized NVO networks where all the MAC and IP addresses are learnt in the control plane.
- c) Multicast applications are not used.

If the above assumptions are true for a given NVO network, then IR provides a simple solution for multi-destination traffic. However, the statement c) above is not always true and multicast applications are required in many use-cases.

When the multicast sources are attached to NVEs residing in hypervisors or low-performance-replication TORs, the ingress replication of large amounts of multicast traffic to a significant number of remote NVEs/PEs can seriously degrade the performance of the NVE and impact the application.

This document describes a solution that makes use of two IR optimizations:

- i) Assisted-Replication (AR)
- ii) Pruned-Flood-Lists (PFL)

Both optimizations may be used together or independently so that the performance and efficiency of the network to transport multicast can be improved. Both solutions require some extensions to [EVPN] that are described in section 3.

Section 2 lists the requirements of the combined optimized-IR solution, whereas section 4 describes the Assisted-Replication (AR) solution and section 5 the Pruned-Flood-Lists (PFL) solution.

## 2. Solution requirements

The IR optimization solution (optimized-IR hereafter) MUST meet the following requirements:

- a) The solution MUST provide an IR optimization for BM (Broadcast and Multicast) traffic, while preserving the packet order for unicast applications, i.e. known and unknown unicast traffic SHALL follow the same path.
- b) The solution MUST be compatible with [EVPN] and [EVPN-OVERLAY] and not have any impact on the EVPN procedures for BM traffic. In particular, the solution MUST support the following EVPN functions:
  - o All-active multi-homing, including the split-horizon and Designated Forwarder (DF) functions.
  - o Single-active multi-homing, including the DF function.
  - o Handling of multi-destination traffic and processing of broadcast and multicast as per [EVPN].
- c) The solution MUST be backwards compatible with existing NVEs using a non-optimized version of IR. A given EVI can have NVEs/PEs supporting regular-IR and optimized-IR.
- d) The solution MUST be independent of the NVO specific data plane encapsulation and the virtual identifiers being used, e.g.: VXLAN VNIs, NVGRE VSIDs or MPLS labels.

## 3. EVPN BGP Attributes for optimized-IR

This solution proposes some changes to the [EVPN] inclusive multicast routes and attributes so that an NVE/PE can signal its optimized-IR capabilities.

The Inclusive Multicast Ethernet Tag route and its PMSI Tunnel attribute's format used in EVPN are shown below:

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

Flags (1 octet)
Tunnel Type (1 octets)
MPLS Label (3 octets)
Tunnel Identifier (variable)

Where:

- o Originating Router's IP Address, Tunnel Type (0x06), MPLS Label and Tunnel Identifier MUST be used as described in [EVPN] for non-optimized-IR behavior.
- o A different Originating Router's IP Address, a new Tunnel Type (TBD), MPLS Label and Tunnel Identifier may be used for Assisted-Replication (AR).
- o The Flags field is defined as follows:

0 1 2 3 4 5 6 7
rsved  T  BM U L

Where a new type field (for AR) and two new flags (for PFL signaling) are defined:

- T is the AR Type field (2 bits):
  - + 00 (decimal 0) = RNVE (non-AR support)
  - + 01 (decimal 1) = AR REPLICATOR

- + 10 (decimal 2) = AR LEAF
- New PFL (Pruned-Flood-Lists) flags:
  - + BM= Broadcast and Multicast (BM) flag. BM=1 means "prune-me" from the BM flooding list. BM=0 means regular behavior.
  - + U= Unknown flag. U=1 means "prune-me" from the Unknown flooding list. U=0 means regular behavior.
- Flag L is an existing flag defined in RFC6514 (L=Leaf Information Required) and it has no use in this solution.

Each AR-enabled EVI node MUST understand and process the AR type field in the PMSI attribute (Flags field) and MUST signal the corresponding type (1 or 2) according to its administrative choice.

Each EVI node MAY understand and process the BM/U flags. Note that these BM/U flags may be used to optimize the delivery of multi-destination traffic and its use SHOULD be an administrative choice, regardless of the AR settings.

The T field and BM/U flags MAY be used individually or together, i.e. a given PMSI attribute may only convey the AR type information, or only the BM/U flags, or both pieces of information at the same time.

Non-optimized-IR nodes will be unaware of the new PMSI attribute flag definition, i.e. they will ignore the information contained in the flags field.

#### 4. Assisted-Replication (AR) Solution Description

The following figure illustrates an example NVO network where the AR function is enabled. This scenario will be used to describe the solution throughout the rest of the document.

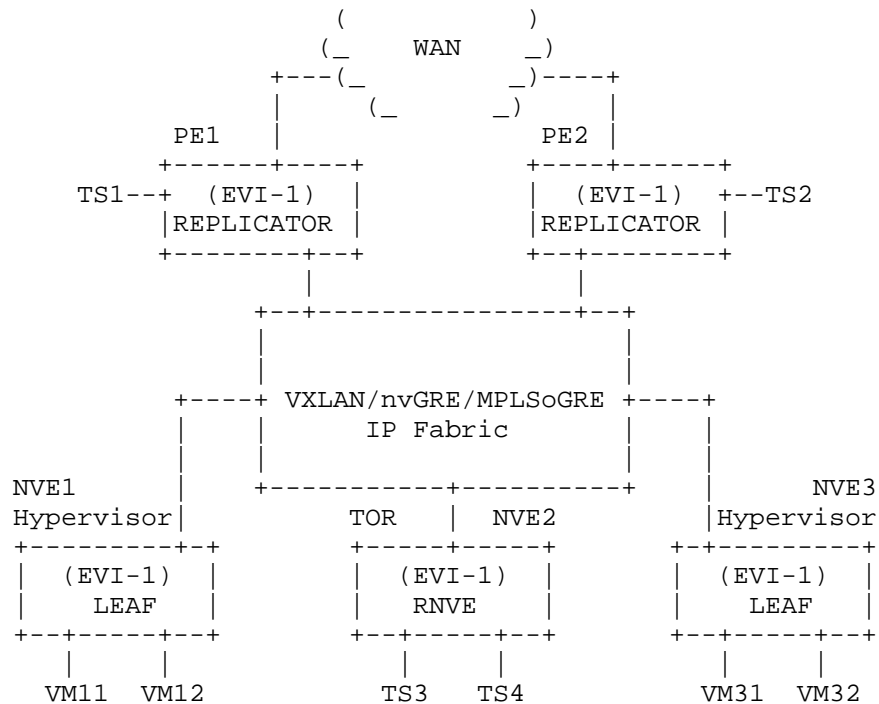


Figure 1 Optimized-IR scenario

#### 4.1. AR roles and control plane

The solution defines three different roles in an AR EVI service:

- a) AR-REPLICATOR (REPLICATOR)
- b) AR-LEAF (LEAF)
- c) Regular NVE (RNVE)

##### 4.1.1. AR-REPLICATOR procedures

REPLICATOR is defined as an NVE/PE capable of replicating ingress BM (Broadcast and Multicast) traffic received on an overlay tunnel to other overlay tunnels and local Attachment Circuits (ACs). The REPLICATOR signals its REPLICATOR role in the control plane and understands where the other roles (LEAF nodes, RNVEs and other REPLICATORS) are located. A given AR EVI service may have zero, one or more REPLICATORS. In our example in figure 1, PE1 and PE2 are defined as REPLICATORS. The following considerations apply to the REPLICATOR role:

- a) The AR-REPLICATOR role SHOULD be an administrative choice in any

NVE/PE that is part of an AR EVI. This administrative option to enable REPLICATOR capabilities MAY be implemented as a system level option as opposed to as per-EVI option.

- b) An AR-REPLICATOR MUST advertise an AR inclusive multicast route and MAY advertise an IR inclusive multicast route.
- c) An IR Inclusive Multicast Route is an Inclusive Multicast Route as defined in [EVPN] and MUST NOT be generated by the AR REPLICATOR if it does not have local attachment circuits (AC).
- d) An AR Inclusive Multicast Route MUST be generated by the AR REPLICATOR and it is comprised of:
  - o AR Originating Router's IP Address, which is different from the IR IP address used in the IR Inclusive Multicast Route.
  - o T = 1 (AR REPLICATOR)
  - o Tunnel type = TBD (AR tunnel)
  - o Tunnel Identifier MUST contain the same value as the AR Originating Router's IP Address.
  - o The rest of the route fields are used as per [EVPN].
- e) When a node defined as REPLICATOR receives a packet from an overlay tunnel, it will do a tunnel destination IP lookup and follow the following procedures:
  - o If the destination IP is the IR Originating Router's IP Address the node will process the packet normally as in [EVPN].
  - o If the destination IP is the AR Originating Router's IP Address, the node MUST replicate the packet to local ACs and overlay tunnels (excluding the overlay tunnel to the source of the packet). Selective replication to only interested AR-LEAF nodes will be added in a future revision of this document.

#### 4.1.2. AR-LEAF procedures

LEAF is defined as an NVE/PE that - given its poor replication performance - sends all the BM traffic to a REPLICATOR that can replicate the traffic further on its behalf. It signals its LEAF capability in the control plane and understands where the other roles are located (REPLICATOR and RNVEs). A given service can have zero, one or more LEAF nodes. Figure 1 shows NVE1 and NVE2 (both residing

in hypervisors) acting as LEAF. The following considerations apply to the LEAF role:

- a) The AR-LEAF role SHOULD be an administrative choice in any NVE/PE that is part of an AR EVI. This administrative option to enable LEAF capabilities MAY be implemented as a system level option as opposed to as per-EVI option.
- b) An AR-LEAF MUST advertise a single inclusive multicast route where the AR type is set to T = 2 (AR LEAF) and the rest of fields follow [EVPN].
- c) In a service where there are no REPLICATORS, the LEAF MUST use regular ingress replication. This will happen when a new update from the last former REPLICATOR is received and contains a non-REPLICATOR AR type, or when the LEAF detects that the last REPLICATOR is down (next-hop tracking in the IGP or any other detection mechanism). Ingress replication MUST use the forwarding information given by the IR Inclusive Multicast Routes as described in [EVPN].
- d) In a service where there is more than one or more REPLICATORS, the LEAF can locally select which REPLICATOR it sends the BM traffic to:
  - o A single REPLICATOR may be selected for all the BM packets received on LEAF attachment circuits (ACs). This selection is a local decision and it does not have to match other LEAF's selection within the same service.
  - o A LEAF may select more than one REPLICATOR and do either per-flow or per-service load balancing.
  - o In case of a failure on the selected REPLICATOR, another REPLICATOR will be selected.
  - o When a REPLICATOR is selected, the LEAF MUST send all the BM packets to that REPLICATOR using the forwarding information given by the AR Inclusive Multicast Route previously sent by the REPLICATOR, with tunnel type = TBD (AR tunnel). The underlay destination IP address MUST be the AR Originating Router's IP Address signaled by the REPLICATOR for the AR tunnel type.
  - o LEAF nodes SHALL send service-level BM control plane packets following regular IR procedures. An example would be IGMP, MLD or PIM multicast packets. The REPLICATORS MUST not replicate these control plane packets to other overlay tunnels since



they will use the regular Originating Router's IP Address.

#### 4.1.3. RNVE procedures

RNVE (Regular Network Virtualization Edge node) is defined as an NVE/PE without REPLICATOR or LEAF capabilities that does IR as described in [EVPN]. The RNVE does not signal any special role and is unaware of the REPLICATOR/LEAF roles in the EVI. The RNVE will ignore AR Inclusive Multicast Routes (due to an unknown tunnel type in the PMSI attribute).

This role provides EVPN with the backwards compatibility required in optimized-IR EVIs. Figure 1 shows NVE2 as RNVE.

#### 4.2. Multi-destination traffic forwarding behavior in AR EVIs

In AR EVIs, BM (Broadcast and Multicast) traffic between two NVEs may follow a different path than unicast traffic. This solution proposes the replication of BM through the REPLICATOR node, whereas unknown/known unicast will be delivered directly from the source node to the destination node without being replicated by any intermediate node. Unknown unicast SHALL follow the same path as known unicast traffic in order to avoid packet reordering for unicast applications and simplify the control and data plane procedures. Section 4.2.1 describes the expected forwarding behavior for BM traffic in nodes acting as REPLICATOR, LEAF and RNVE. Section 4.2.2 describes the forwarding behavior for unknown unicast traffic.

Note that known unicast forwarding is not impacted by this solution.

##### 4.2.1. Broadcast and Multicast forwarding behavior

The expected behavior per role is described in this section.

###### 4.2.1.1. REPLICATOR BM forwarding

The REPLICATORS will build a flooding list composed of ACs and overlay tunnels to remote nodes in the EVI. Some of those overlay tunnels MAY be flagged as non-BM receivers based on the BM flag received from the remote nodes in the EVI. The REPLICATOR will also build a list of remote REPLICATORS, LEAF nodes and RNVEs for the EVI.

- o When a REPLICATOR receives a BM packet on an AC, it will forward the BM packet to its flooding list (including local ACs and remote NVE/PEs), skipping the non-BM overlay tunnels.
- o When a REPLICATOR receives a BM packet on an overlay tunnel, it will check the destination IP of the underlay IP header and:

- If the destination IP matches its AR Originating Router IP, the REPLICATOR will forward the BM packet to its flooding list (ACs and overlay tunnels) excluding the non-BM overlay tunnels. The REPLICATOR will do source squelching to ensure the traffic is not sent back to the originating LEAF. If the overlay encapsulation is MPLS and the EVI label is not the bottom of the stack, the REPLICATOR MUST copy the rest of the labels and forward them to the egress overlay tunnels.
- If the destination IP matches its IR Originating Router IP, the REPLICATOR will skip all the overlay tunnels from the flooding list, i.e. it will only replicate to local ACs. This is the regular IR behavior described in [EVPN].

#### 4.2.1.2. LEAF BM forwarding

The LEAF nodes will build two flood-lists:

- 1) Flood-list #1 - composed of ACs and a REPLICATOR-set of overlay tunnels. The REPLICATOR-set is defined as one or more overlay tunnels to the AR Originating Router's IP Addresses of the remote REPLICATOR(s) in the EVI. The selection of more than one REPLICATOR is described in section 4.1.2 and it is a local LEAF decision.
- 2) Flood-list #2 - composed of ACs and overlay tunnels to the remote IR Originating Router's IP Addresses.

When a LEAF receives a BM packet on an AC, it will check the REPLICATOR-set:

- o If the REPLICATOR-set is empty, the LEAF will send the packet to flood-list #2.
- o If the REPLICATOR-set is NOT empty, the LEAF will send the packet to flood-list #1.

When a LEAF receives a BM packet on an overlay tunnel, will forward the BM packet to its local ACs and never to an overlay tunnel. This is the regular IR behavior described in [EVPN].

#### 4.2.1.3. RNVE BM forwarding

The RNVE is completely unaware of the REPLICATORS, LEAF nodes and BM/U flags (that information is ignored). Its forwarding behavior is the regular IR behavior described in [EVPN]. Any regular non-AR node is fully compatible with the RNVE role described in this document.

#### 4.2.2. Unknown unicast forwarding behavior

The expected behavior is described in this section.

##### 4.2.2.1. REPLICATOR/LEAF Unknown unicast forwarding

While the forwarding behavior in REPLICATORS and LEAF nodes is different for BM traffic, as far as Unknown unicast traffic forwarding is concerned, LEAF nodes behave exactly in the same way as REPLICATORS do.

The REPLICATOR/LEAF nodes will build a flood-list composed of ACs and overlay tunnels to the IR Originating Router's IP Addresses of the remote nodes in the EVI. Some of those overlay tunnels MAY be flagged as non-U (Unknown unicast) receivers based on the U flag received from the remote nodes in the EVI.

- o When a REPLICATOR/LEAF receives an unknown packet on an AC, it will forward the unknown packet to its flood-list, skipping the non-U overlay tunnels.
- o When a REPLICATOR/LEAF receives an unknown packet on an overlay tunnel will forward the unknown packet to its local ACs and never to an overlay tunnel. This is the regular IR behavior described in [EVPN].

##### 4.4.2.2. RNVE Unknown unicast forwarding

As described for BM traffic, the RNVE is completely unaware of the REPLICATORS, LEAF nodes and BM/U flags (that information is ignored). Its forwarding behavior is the regular IR behavior described in [EVPN], also for Unknown unicast traffic. Any regular non-AR node is fully compatible with the RNVE role described in this document.

#### 5. Pruned-Flood-Lists (PFL)

The second optimization supported by this solution is the ability for the all the EVI nodes to signal Pruned-Flood-Lists (PFL). As described in section 3, an EVPN node can signal a given value for the BM and U PFL flags in the IR Inclusive Multicast Routes, where:

- + BM= Broadcast and Multicast (BM) flag. BM=1 means "prune-me" from the BM flood-list. BM=0 means regular behavior.
- + U= Unknown flag. U=1 means "prune-me" from the Unknown flood-list. U=0 means regular behavior.

The ability to signal these PFL flags is an administrative choice.

Upon receiving a non-zero PFL flag, a node MAY decide to honor the PFL flag and remove the sender from the corresponding flood-list. A given EVI node receiving BUM traffic on an overlay tunnel MUST replicate the traffic normally, regardless of the signaled PFL flags.

This optimization MAY be used along with the AR solution.

## 6. An example use-case

In order to illustrate the use of the solution described in this document, we will assume that EVI-1 in figure 1 is optimized-IR enabled and:

- o PE1 and PE2 are administratively configured as REPLICATORS, due to their high-performance replication capabilities. PE1 and PE2 will signal AR type = 1 and BM/U flags = 00.
- o NVE1 and NVE3 are administratively configured as LEAF nodes, due to their low-performance software-based replication capabilities. They will signal AR type = 2. Assuming both NVEs advertise all the attached VMs in EVPN as soon as they come up and don't have any VMs interested in multicast applications, they will be configured to signal BM/U flags = 11 for EVI-1.
- o NVE2 is optimized-IR unaware; therefore it takes on the RNVE role in EVI-1.

Based on the above assumptions the following forwarding behavior will take place:

- (1) Any BM packets sent from VM11 will be sent to VM12 and PE1. PE1 will forward further the BM packets to TS1, WAN link, PE2 and NVE2, but not to NVE3. PE2 and NVE2 will replicate the BM packets to their local ACs but we will avoid NVE3 having to replicate unnecessarily those BM packets to VM31 and VM32.
- (2) Any BM packets received on PE2 from the WAN will be sent to PE1 and NVE2, but not to NVE1 and NVE3, sparing the two hypervisors from replicating unnecessarily to their local VMs. PE1 and NVE2 will replicate to their local ACs only.
- (3) Any Unknown unicast packet sent from VM31 will be forwarded by NVE3 to NVE2, PE1 and PE2 but not NVE1. The solution avoids the unnecessary replication to NVE1, since the destination of the unknown traffic cannot be at NVE1.
- (4) Any Unknown unicast packet sent from TS1 will be forwarded by PE1

to the WAN link, PE2 and NVE2 but not to NVE1 and NVE3, since the target of the unknown traffic cannot be at those NVEs.

## 5. Benefits of the optimized-IR solution

A solution for the optimization of Ingress Replication in EVPN is described in this document (optimized-IR). The solution brings the following benefits:

- o Optimizes the multicast forwarding in low-performance NVEs, by relaying the replication to high-performance NVEs (REPLICATORS) and while preserving the packet ordering for unicast applications.
- o Reduces the flooded traffic in NVO networks where some NVEs do not need broadcast/multicast and/or unknown unicast traffic.
- o It is fully compatible with existing EVPN implementations and EVPN functions for NVO overlay tunnels. Optimized-IR NVEs and regular NVEs can be even part of the same EVI.
- o It does not require any PIM-based tree in the NVO core of the network.

## 6. Conventions used in this document

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

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

## 7. Security Considerations

This section will be added in future versions.

## 8. IANA Considerations

## 8. References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-07.txt, work in progress, May, 2014

[EVPN-OVERLAY] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", draft-sd-l2vpn-evpn-overlay-02.txt, work in progress, October, 2013

## 9. Acknowledgments

The authors would like to thank Neil Hart and David Motz for their valuable feedback and contributions.

## 10. Authors' Addresses

Jorge Rabadan  
Alcatel-Lucent  
777 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: jorge.rabadan@alcatel-lucent.com

Senthil Sathappan  
Alcatel-Lucent  
Email: senthil.sathappan@alcatel-lucent.com

Mukul Katiyar  
Nuage Networks  
Email:

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

Ravi Shekhar  
Juniper Networks  
Email: rshekhar@juniper.net

Nischal Sheth  
Juniper Networks  
Email: nsheth@juniper.net

Wen Lin  
Juniper Networks  
Email: wlin@juniper.net

L2VPN Workgroup  
Internet Draft

Intended status: Standards Track

J. Drake  
Juniper

A. Sajassi  
Cisco

J. Rabadan  
W. Henderickx  
S. Palislaamovic  
Alcatel-Lucent

F. Balus  
Nuage Networks

A. Isaac  
Bloomberg

Expires: January 5, 2015

July 4, 2014

IP Prefix Advertisement in EVPN  
draft-rabadan-l2vpn-evpn-prefix-advertisement-02

Abstract

EVPN provides a flexible control plane that allows intra-subnet connectivity in an IP/MPLS and/or an NVO-based network. In NVO networks, there is also a need for a dynamic and efficient inter-subnet connectivity across Tenant Systems and End Devices that can be physical or virtual and may not support their own routing protocols. This document defines a new EVPN route type for the advertisement of IP Prefixes and explains some use-case examples where this new route-type is used.

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 5, 2015.

#### Copyright Notice

Copyright (c) 2014 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. Terminology . . . . .	3
2. Introduction and problem statement . . . . .	3
2.1 Inter-subnet connectivity requirements in Data Centers . . . .	4
2.2 The requirement for advertising IP prefixes in EVPN . . . . .	6
2.3 The requirement for a new EVPN route type . . . . .	7
3. The BGP EVPN IP Prefix route . . . . .	9
3.1 IP Prefix Route encoding . . . . .	9
4. Benefits of using the EVPN IP Prefix route . . . . .	11
5. IP Prefix next-hop use-cases . . . . .	12
5.1 TS IP address next-hop use-case . . . . .	12
5.2 Floating IP next-hop use-case . . . . .	15
5.3 IRB IP next-hop use-case . . . . .	16
5.4 ESI next-hop ("Bump in the wire") use-case . . . . .	18
5.5 IRB forwarding without core-facing IRB use-case (VRF-to-VRF) . . . . .	20
6. Conclusions . . . . .	23
7. Conventions used in this document . . . . .	24
8. Security Considerations . . . . .	24
9. IANA Considerations . . . . .	24
10. References . . . . .	24
10.1 Normative References . . . . .	24
10.2 Informative References . . . . .	24
11. Acknowledgments . . . . .	24
12. Authors' Addresses . . . . .	24



## 1. Terminology

GW IP: Gateway IP Address

IPL: IP address length

IRB: Integrated Routing and Bridging interface

ML: MAC address length

NVE: Network Virtualization Edge

TS: Tenant System

VA: Virtual Appliance

RT-2: EVPN route type 2, i.e. MAC/IP advertisement route

RT-5: EVPN route type 5, i.e. IP Prefix route

Overlay next-hop: object used in the IP Prefix route, as described in this document. It can be an IP address in the tenant space or an ESI, and identifies the next-hop to be used in IP lookups for a given IP Prefix at the routing context importing the route.

Underlay next-hop: IP address sent by BGP along with any EVPN route, i.e. BGP next-hop. It identifies the NVE sending the route and it is used at the receiving NVE as the VXLAN destination VTEP or NVGRE destination end-point.

## 2. Introduction and problem statement

Inter-subnet connectivity is required for certain tenants within the Data Center. [EVPN-INTERSUBNET] defines some fairly common inter-subnet forwarding scenarios where TSes can exchange packets with TSes located in remote subnets. In order to meet this requirement, [EVPN-INTERSUBNET] describes how MAC/IPs encoded in TS RT-2 routes are not only used to populate MAC-VRF and overlay ARP tables, but also IP-VRF tables with the encoded TS host routes (/32 or /128). In some cases, EVPN may advertise IP Prefixes and therefore provide aggregation in the IP-VRF tables, as opposed to program individual host routes. This document complements the scenarios described in [EVPN-INTERSUBNET] and defines how EVPN may be used to advertise IP Prefixes.

Section 2.1 describes the inter-subnet connectivity requirements in Data Centers. Section 2.2 and 2.3 explain why neither IP-VPN nor the existing EVPN route types meet the requirements for IP Prefix advertisements. Once the need for a new EVPN route type is justified,

sections 2 and 3 will describe this route type and how it is used in some specific use cases.

## 2.1 Inter-subnet connectivity requirements in Data Centers

[EVPN] is used as the control plane for a Network Virtualization Overlay (NVO3) solution in Data Centers (DC), where Network Virtualization Edge (NVE) devices can be located in Hypervisors or TORs, as described in [EVPN-OVERLAYS].

If we use the term Tenant System (TS) to designate a physical or virtual system identified by MAC and IP addresses, and connected to an EVPN instance, the following considerations apply:

- o The Tenant Systems may be Virtual Machines (VMs) that generate traffic from their own MAC and IP.
- o The Tenant Systems may be Virtual Appliance entities (VAs) that forward traffic to/from IP addresses of different End Devices seating behind them.
  - o These VAs can be firewalls, load balancers, NAT devices, other appliances or virtual gateways with virtual routing instances.
  - o These VAs do not have their own routing protocols and hence rely on the EVPN NVEs to advertise the routes on their behalf.
  - o In all these cases, the VA will forward traffic to the Data Center using its own source MAC but the source IP will be the one associated to the End Device seating behind or a translated IP address (part of a public NAT pool) if the VA is performing NAT.
  - o Note that the same IP address could exist behind two of these TS. One example of this would be certain appliance resiliency mechanisms, where a virtual IP or floating IP can be owned by one of the two VAs running the resiliency protocol (the master VA). VRRP is one particular example of this. Another example is multi-homed subnets, i.e. the same subnet is connected to two VAs.
  - o Although these VAs provide IP connectivity to VMs and subnets behind them, they do not always have their own IP interface connected to the EVPN NVE, e.g. layer-2 firewalls are examples of VAs not supporting IP interfaces.

The following figure illustrates some of the examples described above.

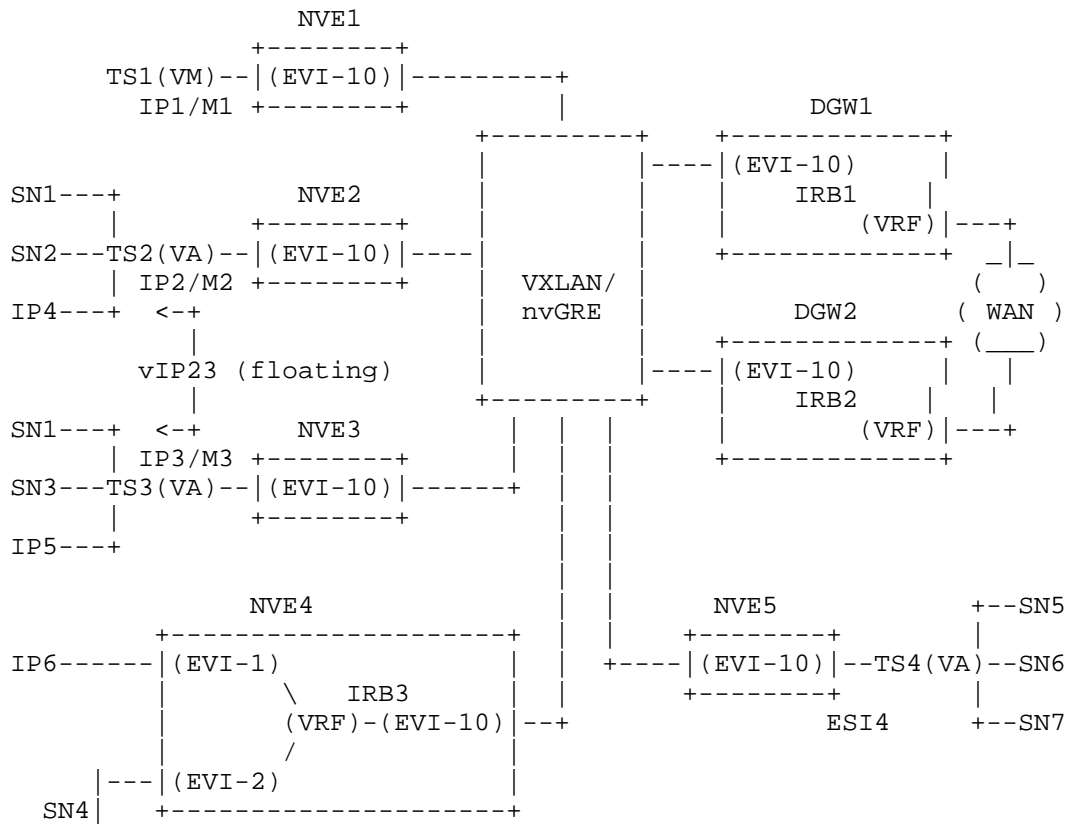


Figure 1 DC inter-subnet use-cases

Where:

NVE1, NVE2, NVE3, NVE4, NVE5, DGW1 and DGW2 share the same EVI for a particular tenant. EVI-10 is the corresponding EVPN MAC-VRF for the shared EVI on each element, i.e. core-facing EVI, and all the hosts connected to that instance belong to the same IP subnet. The hosts connected to EVI-10 are listed below:

- o TS1 is a VM that generates/receives traffic from/to IP1, where IP1 belongs to the EVI-10 subnet.
- o TS2 and TS3 are Virtual Appliances (VA) that generate/receive traffic from/to the subnets and hosts seating behind them (SN1, SN2, SN3, IP4 and IP5). Their IP addresses (IP2 and IP3) belong to the EVI-10 subnet and they can also generate/receive traffic. When these VAs receive packets destined to their own

MAC addresses (M2 and M3) they will route the packets to the proper subnet or host. These VAs do not support routing protocols to advertise the subnets connected to them and can move to a different server and NVE when the Cloud Management System decides to do so. These VAs may also support redundancy mechanisms for some subnets, similar to VRRP, where a floating IP is owned by the master VA and only the master VA forwards traffic to a given subnet. E.g.: vIP23 in figure 1 is a floating IP that can be owned by TS2 or TS3 depending on who the master is. Only the master will forward traffic to SN1.

- o Integrated Routing and Bridging interfaces IRB1, IRB2 and IRB3 have their own IP addresses that belong to the EVI-10 subnet too. These IRB interfaces connect the EVI-10 subnet to Virtual Routing and Forwarding (VRF) instances that can route the traffic to other connected subnets for the same tenant (within the DC or at the other end of the WAN).
- o TS4 is a layer-2 VA that provides connectivity to subnets SN5, SN6 and SN7, but does not have an IP address itself in the EVI-10. TS4 is connected to a physical port on NVE5 assigned to Ethernet Segment Identifier 4.

All the above DC use cases require inter-subnet forwarding and therefore the individual host routes and subnets MUST be advertised:

- a) From the NVEs (since VAs and VMs do not run routing protocols) and
- b) Associated to an overlay next-hop that can be a VA IP address, a floating IP address, and IRB IP address or an ESI.

## 2.2 The requirement for advertising IP prefixes in EVPN

In all the inter-subnet connectivity cases discussed in section 2.1 there is a need to advertise IP prefixes in the control plane. The advertisement of such prefixes must meet certain requirements, specific to NVO-based Data Centers:

- o The data plane in NVO-based Data Centers is not based on IP over a GRE or MPLS tunnel as required by [RFC4364], but Ethernet over an IP tunnel, such as VXLAN or NVGRE.
- o The IP prefixes in the DC must be advertised with a flexibility that does not exist in IP-VPNs today. For instance:
  - a) The advertised overlay next-hop for a given IP prefix can be an IRB IP address (see section 5.3), a floating IP

address (see section 5.2) or even an ESI (see section 5.4).

b) VXLAN or NVGRE virtual identifiers can have a global or a local scope. The implementation MUST support the flexibility to advertise IP Prefixes associated to a global identifier (32-bit value encoded in the EVPN Ethernet Tag ID) or a locally significant identifier (20-bit value encoded in the MPLS label field). At the moment, [RFC4364] can only advertise Prefixes associated to a locally significant identifier (MPLS label).

c) Since an NVE can potentially advertise many Prefixes with different overlay next-hops and different VXLAN/NVGRE identifiers, it is highly desirable to be able to advertise those prefixes with their corresponding overlay next-hops and VXLAN/NVGRE identifiers as attributes within the same NLRI, for a better BGP update packing. [RFC4364] does not have the capability of advertising a flexible overlay next-hop together with a prefix in the same NLRI.

- o IP prefixes must be advertised by NVE devices that have no VRF instances defined and no capability to process IP-VPN prefixes. These NVE devices just support EVPN and advertise IP Prefixes on behalf of some connected Tenant Systems. In other words: any attempt to solve this problem by simply using [RFC4364] routes requires that any EVPN deployment must be accompanied with a concurrent IP-VPN topology, which is not possible in most of the cases.
- o Finally, Data Center providers want to use a single BGP Subsequent Address Family (AFI/SAFI) for the advertisement of addresses within the Data Center, i.e. BGP EVPN only, as opposed to using EVPN and IP-VPN in a concurrent topology. This minimizes the control plane overhead in TORs and Hypervisors and simplifies the operations.

EVPN is extended - as described in this document - to advertise IP prefixes with the flexibility required by the current and future Data Center applications.

### 2.3 The requirement for a new EVPN route type

[EVPN] defines a MAC/IP route (or RT-2) where a MAC address can be advertised together with an IP address length (IPL) and IP address (IP). While a variable IPL might be used to indicate the presence of an IP prefix in a route type 2, there are several specific use cases in which using this route type to deliver IP Prefixes is not suitable.

One example of such use cases is the "floating IP" example described in section 2.1. In this example we need to decouple the advertisement of the prefixes from the advertisement of the floating IP (vIP23 in figure 1) and MAC associated to it, otherwise the solution gets highly inefficient and does not scale.

E.g.: if we are advertising 1k prefixes from M2 (using route type 2) and the floating IP owner changes from M2 to M3, we would need to withdraw 1k routes from M2 and re-advertise 1k routes from M3. However if we use a separate route type, we can advertise the 1k routes associated to the floating IP address (vIP23) and only one route type 2 for advertising the ownership of the floating IP, i.e. vIP23 and M2 in the route type 2. When the floating IP owner changes from M2 to M3, a single route type 2 withdraw/update is required to indicate the change. The remote DGW will not change any of the 1k prefixes associated to vIP23, but will only update the ARP resolution entry for vIP23 (now pointing at M3).

Other reasons to decouple the IP Prefix advertisement from the MAC route are listed below:

- o Clean identification, operation of troubleshooting of IP Prefixes, not subject to interpretation and independent of the IPL and the IP value. E.g.: a default IP route 0.0.0.0/0 must always be easily and clearly distinguished from the absence of IP information.
- o MAC address information must not be compared by BGP when selecting two IP Prefix routes. If IP Prefixes are to be advertised using MAC routes, the MAC information is always present and part of the route key.
- o IP Prefix routes must not be subject to MAC route procedures such as MAC Mobility or aliasing. Prefixes advertised from two different ESIs do not mean mobility; MACs advertised from two different ESIs do mean mobility. Similarly load balancing for IP prefixes is achieved through IP mechanisms such as ECMP, and not through MAC route mechanisms such as aliasing.
- o NVEs that do not require processing IP Prefixes must have an easy way to identify an update with an IP Prefix and ignore it, rather than processing the MAC route only to find out later that it carries a Prefix that must be ignored.

The following sections describe how EVPN is extended with a new route type for the advertisement of prefixes and how this route is used to address the current and future inter-subnet connectivity requirements existing in the Data Center.

### 3. The BGP EVPN IP Prefix route

The current BGP EVPN NLRI as defined in [EVPN] is shown below:

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

Where the route type field can contain one of the following specific values:

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

This document defines an additional route type that will be used for the advertisement of IP Prefixes:

- + 5 - IP Prefix Route

The support for this new route type is OPTIONAL.

Since this new route type is OPTIONAL, an implementation not supporting it MUST ignore the route, based on the unknown route type value.

The detailed encoding of this route and associated procedures are described in the following sections.

#### 3.1 IP Prefix Route encoding

An IP Prefix advertisement route type specific EVPN NLRI consists of the following fields:

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

Where:

- o RD, Ethernet Tag ID and MPLS Label fields will be used as defined in [EVPN] and [EVPN-OVERLAYS].
- o The Ethernet Segment Identifier will be a non-zero 10-byte identifier if the ESI is used as an overlay next-hop. It will be zero otherwise.
- o The IP address length can be set to a value between 0 and 32 (bits) for ipv4 and between 0 and 128 for ipv6.
- o The IP address will be a 32 or 128-bit field (ipv4 or ipv6).
- o The GW IP (Gateway IP Address) will be a 32 or 128-bit field (ipv4 or ipv6), and will encode the overlay IP next-hop for the IP Prefixes. The GW IP field can be zero if it is not used as an overlay next-hop.
- o The total route length will indicate the type of prefix (ipv4 or ipv6) and the type of GW IP address (ipv4 or ipv6). Note that the IP Address + the GW IP should have a length of either 64 or 256 bits, but never 160 bits (ipv4 and ipv6 mixed values are not allowed).

The Eth-Tag ID, IP address length and IP address will be part of the route key used by BGP to compare routes. The rest of the fields will be out of the route key.

The route will contain a single overlay next-hop, i.e. if the ESI field is zero, the GW IP field will not, and vice versa. The following table shows the different inter-subnet use-cases described



in this document and the corresponding coding of the overlay next-hop in the route-type 5 (RT-5).

Overlay next-hop use-case	Field in the RT-5
TS IP address	GW IP Address
Floating IP address	GW IP Address
IRB IP address	GW IP Address
"Bump in the wire"	ESI
VRF-to-VRF	GW MAC Address (Tunnel Attribute)

#### 4. Benefits of using the EVPN IP Prefix route

This section clarifies the different functions accomplished by the EVPN RT-2 and RT-5 routes, and provides a list of benefits derived from using a separate route type for the advertisement of IP Prefixes in EVPN.

[EVPN] describes the content of the BGP EVPN route type 2 specific NLRI, i.e. MAC/IP Advertisement Route, where the IP address length (IPL) and IP address (IP) of a specific advertised MAC are encoded. The subject of the MAC advertisement route is the MAC address (M) and MAC address length (ML) encoded in the route. The MAC mobility and other complex procedures are defined around that MAC address. The IP address information carries the host IP address required for the ARP resolution of the MAC according to [EVPN] and the host route to be programmed in the IP-VRF [EVPN-INTERSUBNET].

The BGP EVPN route type 5 defined in this document, i.e. IP Prefix Advertisement route, decouples the advertisement of IP prefixes from the advertisement of any MAC address related to it. This brings some major benefits to NVO-based networks where certain inter-subnet forwarding scenarios are required. Some of those benefits are:

- a) Upon receiving a route type 2 or type 5, an egress NVE can easily distinguish MACs and IPs from IP Prefixes. E.g. an IP prefix with IPL=32 being advertised from two different ingress NVEs (as RT-5) can be identified as such and be imported in the designated routing context as two ECMP routes, as opposed to two MACs competing for the same IP.
- b) Similarly, upon receiving a route, an egress NVE not supporting processing IP Prefixes can easily ignore the update, based on the route type.
- c) A MAC route includes the ML, M, IPL and IP in the route key that

is used by BGP to compare routes, whereas for IP Prefix routes, only IPL and IP (as well as Ethernet Tag ID) are part of the route key. Advertised IP Prefixes are imported into the designated routing context, where there is no MAC information associated to IP routes. In the example illustrated in figure 1, subnet SN1 should be advertised by NVE2 and NVE3 and interpreted by DGW1 as the same route coming from two different next-hops, regardless of the MAC address associated to TS2 or TS3. This is easily accomplished in the route type 5 by including only the IP information in the route key.

- d) By decoupling the MAC from the IP Prefix advertisement procedures, we can leave the IP prefix advertisements out of the MAC mobility procedures defined in [EVPN] for MACs. In addition, this allows us to have an indirection mechanism for IP prefixes advertised from a MAC/IP that can move between hypervisors. E.g. if there are 1,000 prefixes seating behind TS2 (figure 1), NVE2 will advertise all those prefixes in RT-5 routes associated to the next-hop IP2. Should TS2 move to a different NVE, a single MAC advertisement route withdraw for the M2/IP2 route from NVE2 will invalidate the 1,000 prefixes, as opposed to have to wait for each individual prefix to be withdrawn. This may be easily accomplished by using IP Prefix routes that are not tied to a MAC address, and use a different MAC route to advertise the location and resolution of the overlay next-hop to a MAC address.

## 5. IP Prefix next-hop use-cases

The IP Prefix route can use a GW IP, an ESI or a GW MAC as an overlay next-hop. This section describes some use-cases for these next-hop types.

### 5.1 TS IP address next-hop use-case

The following figure illustrates an example of inter-subnet forwarding for subnets seating behind Virtual Appliances (on TS2 and TS3).

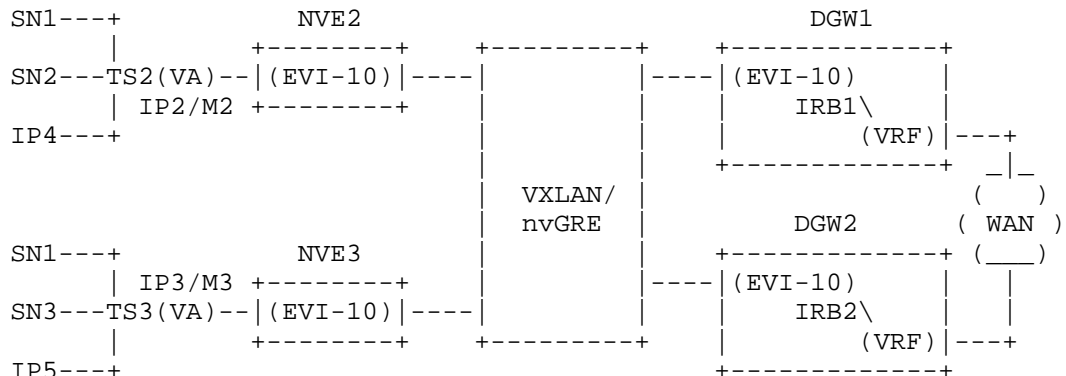


Figure 2 TS IP address use-case

An example of inter-subnet forwarding between subnet SN1/24 and a subnet seating in the WAN is described below. NVE2, NVE3, DGW1 and DGW2 are running BGP EVPN. TS2 and TS3 do not support routing protocols, only a static route to forward the traffic to the WAN.

(1) NVE2 advertises the following BGP routes on behalf of TS2:

- o Route type 2 (MAC route) containing: ML=48, M=M2, IPL=32, IP=IP2
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP2

(2) NVE3 advertises the following BGP routes on behalf of TS3:

- o Route type 2 (MAC route) containing: ML=48, M=M3, IPL=32, IP=IP3
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP3

(3) DGW1 and DGW2 import both received routes based on the RT:

- o Based on the EVI-10 route-target in DGW1 and DGW2, the MAC route is imported and M2 is added to the EVI-10 MAC-VRF along with its corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop (underlay next-hop) and VNI from the Ethernet Tag or MPLS fields. IP2 - M2 is added to the ARP table.
- o Based on the EVI-10 route-target in DGW1 and DGW2, the IP Prefix route is also imported and SN1/24 is added to the

designated routing context with next-hop IP2 pointing at the local EVI-10. Should ECMP be enabled in the routing context, SN1/24 would also be added to the routing table with next-hop IP3.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop=IP2 is found. The tunnel information to encapsulate the packet will be derived from the route-type 2 (MAC route) received for M2/IP2.
- o IP2 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC FIB (remote VTEP and VNI for the VXLAN case).
- o The IP packet destined to IPx is encapsulated with:
  - . Source inner MAC = IRB1 MAC
  - . Destination inner MAC = M2
  - . Tunnel information provided by the MAC-VRF (VNI, VTEP IPs and MACs for the VXLAN case)

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup.
- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(6) Should TS2 move from NVE2 to NVE3, MAC Mobility procedures will be applied to the MAC route IP2/M2, as defined in [EVPN]. Route type 5 prefixes are not subject to MAC mobility procedures, hence no changes in the DGW VRF routing table will occur for TS2 mobility, i.e. all the prefixes will still be pointing at IP2 as next-hop. There is an indirection for e.g. SN1/24, which still points at next-hop IP2 in the routing table, but IP2 will be simply resolved to a different tunnel, based on the outcome of the MAC mobility procedures for the MAC route IP2/M2.

Note that in the opposite direction, TS2 will send traffic based on its static-route next-hop information (IRB1 and/or IRB2), and regular EVPN procedures will be applied.

## 5.2 Floating IP next-hop use-case

Sometimes Tenant Systems (TS) work in active/standby mode where an upstream floating IP - owned by the active TS - is used as the next-hop to get to some subnets behind. This redundancy mode, already introduced in section 2.1 and 2.3, is illustrated in Figure 3.

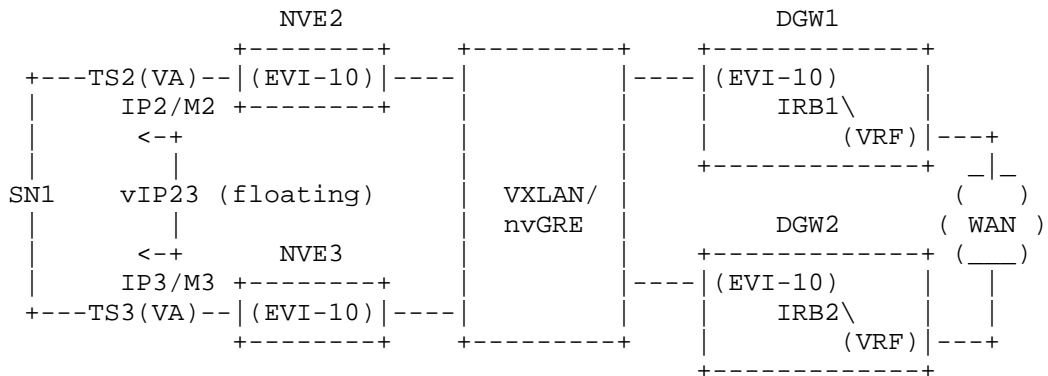


Figure 3 Floating IP next-hop for redundant TS

In this example, assuming TS2 is the active TS and owns IP23:

- (1) NVE2 advertises the following BGP routes for TS2:
  - o Route type 2 (MAC route) containing: ML=48, M=M2, IPL=32, IP=IP23
  - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP23
- (2) NVE3 advertises the following BGP routes for TS3:
  - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP23
- (3) DGW1 and DGW2 import both received routes based on the RT:
  - o M2 is added to the EVI-10 MAC FIB along with its corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop and VNI from the Ethernet Tag or MPLS fields. IP23 - M2 is added to the ARP table.
  - o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop IP23 pointing at the local EVI-10.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 IP-VRF routing table and next-hop=IP23 is found. The tunnel information to encapsulate the packet will be derived from the route-type 2 (MAC route) received for M2/IP23.
- o IP23 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC-VRF (remote VTEP and VNI for the VXLAN case).
- o The IP packet destined to IPx is encapsulated with:
  - . Source inner MAC = IRB1 MAC
  - . Destination inner MAC = M2
  - . Tunnel information provided by the MAC FIB (VNI, VTEP IPs and MACs for the VXLAN case)

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup.
- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(6) When the redundancy protocol running between TS2 and TS3 appoints TS3 as the new active TS for SN1, TS3 will now own the floating IP23 and will signal this new ownership (GARP message or similar). Upon receiving the new owner's notification, NVE3 will issue a route type 2 for M3-IP23. DGW1 and DGW2 will update their ARP tables with the new MAC resolving the floating IP. No changes are carried out in the VRF routing table.

In the DGW1/2 BGP RIB, there will be two route type 5 routes for SN1 (from NVE2 and NVE3) but only the one with the same BGP next-hop as the IP23 route type 2 BGP next-hop will be valid.

### 5.3 IRB IP next-hop use-case

In some other cases, the NVEs and DGWs will have just IRB interfaces as hosts in the EVPN instance. This use-case is referred as "IRB

forwarding on NVEs with core-facing IRB Interface" in [EVPN-INTERSUBNET], however the new requirement here is the advertisement of IP Prefixes as opposed to only host routes. Figure 4 illustrates an example.

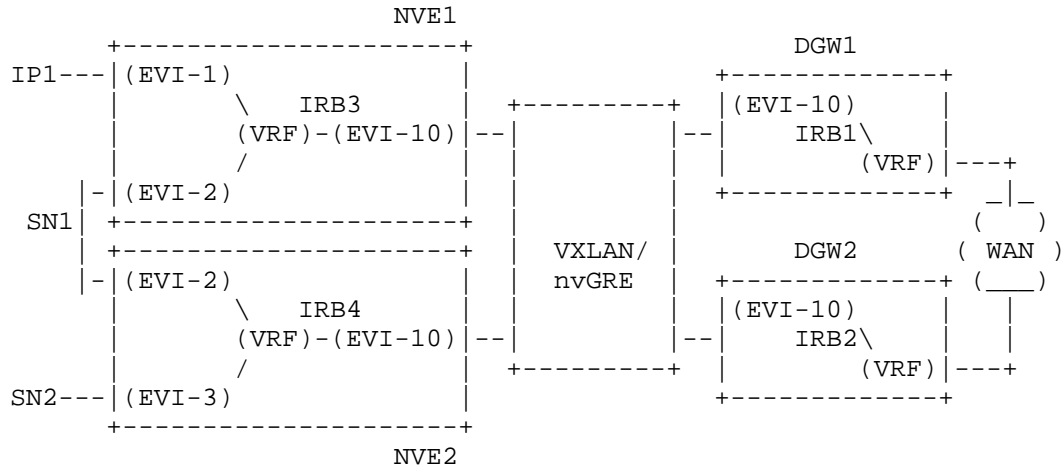


Figure 4 IRB IP next-hop use-case

In this case:

(1) NVE1 advertises the following BGP routes for SN1 resolution:

- o Route type 2 (MAC route) containing: ML=48, M=IRB3-MAC, IPL=32, IP=IRB3-IP
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IRB3-IP

(2) NVE2 advertises the following BGP routes for SN1 resolution:

- o Route type 2 (MAC route) containing: ML=48, M=IRB4-MAC, IPL=32, IP=IRB4-IP
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IRB4-IP

(3) DGW1 and DGW2 import both received routes based on the RT:

- o IRB3-MAC and IRB4-MAC are added to the EVI-10 MAC-VRF along with their corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop and VNI from the Ethernet Tag or MPLS fields. IRB3-MAC - IRB3-

IP and IRB4-MAC - IRB4-IP are added to the ARP table.

- o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop IRB3-IP (and/or IRB4-IP) pointing at the local EVI-10.

Similar forwarding procedures as the ones described in the previous use-cases are followed.

#### 5.4 ESI next-hop ("Bump in the wire") use-case

The following figure illustrates an example of inter-subnet forwarding for a subnet route that uses an ESI as an overlay next-hop. In this use-case, TS2 and TS3 are layer-2 VA devices without any IP address that can be included as an overlay next-hop in the GW IP field of the IP Prefix route.

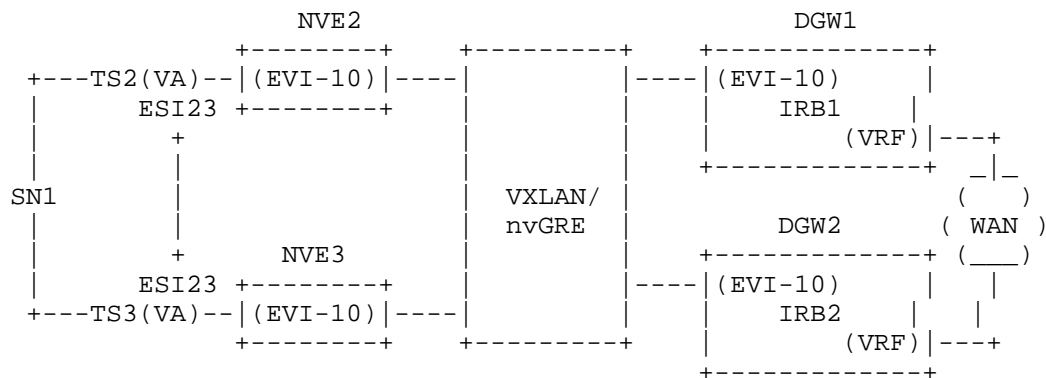


Figure 5 ESI next-hop use-case

Since neither TS2 nor TS3 can run any routing protocol and have no IP address assigned, an ESI, i.e. ESI23, will be provisioned on the attachment ports of NVE2 and NVE3. This model supports VA redundancy in a similar way as the one described in section 4.2 for the floating IP next-hop use-case, only using the EVPN A-D route instead of the MAC advertisement route to advertise the location of the overlay next-hop. The procedure is explained below:

(1) NVE2 advertises the following BGP routes for TS2:

- o Route type 1 (A-D route for EVI-10) containing: ESI=ESI23 and the corresponding tunnel information (Ethernet Tag and/or MPLS label). Assuming the ESI is active on NVE2, NVE2 will advertise this route.



- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=ESI23, GW IP address=0.

(2) NVE3 advertises the following BGP routes for TS3:

- o Route type 1 (A-D route for EVI-10) containing: ESI=ESI23 and the corresponding tunnel information (Ethernet Tag and/or MPLS label). NVE3 will advertise this route assuming the ESI is active on NVE2. Note that if the resiliency mechanism for TS2 and TS3 is in active-active mode, both NVE2 and NVE3 will send the A-D route. Otherwise, that is, the resiliency is active-standby, only the NVE owning the active ESI will advertise the A-D route for ESI23.
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=23, GW IP address=0.

(3) DGW1 and DGW2 import the received routes based on the RT:

- o The tunnel information to get to ESI23 is installed in DGW1 and DGW2. For the VXLAN use case, the VTEP will be derived from the A-D route BGP next-hop and VNI from the Ethernet Tag or MPLS fields (see [EVPN-OVERLAYS]).
- o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop ESI23 pointing at the local EVI-10.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop=ESI23 is found. The tunnel information to encapsulate the packet will be derived from the route-type 1 (A-D route) received for ESI23.
- o The IP packet destined to IPx is encapsulated with:
  - . Source inner MAC = IRB1 MAC
  - . Destination inner MAC = M2 (this MAC will be obtained after a looked up in the VRF ARP table or in the EVI-10 FDB table associated to ESI23).
  - . Tunnel information provided by the A-D route for ESI23 (VNI, VTEP IP and MACs for the VXLAN case).

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup (assuming MAC disposition model).
- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly forwarded.

(6) If the redundancy protocol running between TS2 and TS3 follows an active/standby model and there is a failure, appointing TS3 as the new active TS for SN1, TS3 will now own the connectivity to SN1 and will signal this new ownership (GARP message or similar). Upon receiving the new owner's notification, NVE3 will issue a route type 1 for ESI23, whereas NVE2 will withdraw its A-D route for ESI23. DGW1 and DGW2 will update their tunnel information to resolve ESI23. No changes are carried out in the VRF routing table.

In the DGW1/2 BGP RIB, there will be two route type 5 routes for SN1 (from NVE2 and NVE3) but only the one with the same BGP next-hop as the ESI23 route type 1 BGP next-hop will be valid.

#### 5.5 IRB forwarding without core-facing IRB use-case (VRF-to-VRF)

This use-case is referred as "IRB forwarding on NVEs without core-facing IRB Interface" in [EVPN-INTERSUBNET], however the new requirement here is the advertisement of IP Prefixes as opposed to only host routes. In the previous examples, the EVI instance can connect IRB interfaces and any other Tenant Systems connected to it. EVPN provides connectivity for:

- a) Traffic destined to the IRB IP interfaces as well as
- b) Traffic destined to IP subnets seating behind the IRB interfaces, e.g. SN1 or SN2.

In order to provide connectivity for (a) we need MAC/IP routes (RT-2) distributing IRB MACs and IPs. Connectivity type (b) is accomplished by the exchange of IP Prefix routes (route type 5) for IPs and subnets seating behind IRBs.

In some cases, connectivity type (a) (see above) is not required and the EVI instance is connecting only IRB interfaces, which are never the final destination of any packet. This use case is depicted in the diagram below and we refer to it as the "IRB forwarding on NVEs without core-facing IRB Interface" use-case:

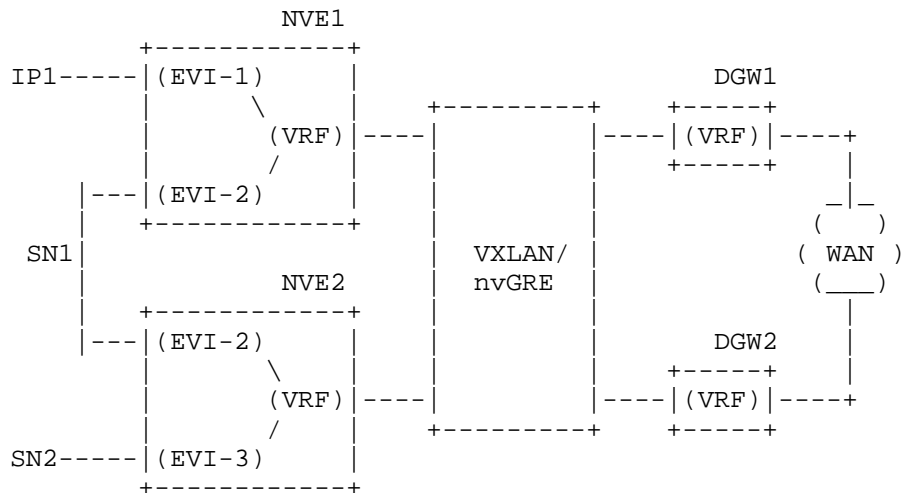


Figure 6 Inter-subnet forwarding without core-facing IRB interfaces

In this case, we need to provide connectivity from/to IP hosts in SN1, SN2, IP1 and hosts seating at the other end of the WAN. The EVI in the core just connects all the IRBs in NVE1, NVE2, DGW1 and DGW2 but there will not be any IP host in this core EVI that is the final destination of any IP packet.

Therefore there is no need to define IRB interfaces (IRBs are not represented in the diagram). This is the reason why we refer to this solution as "Inter-subnet forwarding without core-facing IRB interfaces" or "VRF-to-VRF" solution.

In this case, the proposal is to use EVPN type 5 routes and a BGP tunnel encapsulation attribute as in [EVPN-INTERSUBNET], where the following information is carried:

- o Route type 5 Eth-Tag ID can contain the core instance VNI (if the VNI is global, otherwise, for local significant VNIs, an MPLS label field may be added with a 20-bit VNI encoded in the label space).
- o Route type 5 IP address length and IP address, as explained in the previous sections.
- o Route type 5 GW IP address=0 and ESI=0.
- o Tunnel Encapsulation Attribute as per [EVPN-INTERSUBNET] containing the following fields and including the GW MAC to be

used in the overlay encapsulation:

- o Tunnel Type (2 octets) is:
  - + TBD - VXLAN Encapsulation
  - + TBD - NVGRE Encapsulation
- o Length (2 octets): the total number of octets of the value field.
- o Address Length= 6 bytes (for MAC address)
- o Address= GW MAC Address, a MAC address associated to the system advertising the route. This MAC address identifies the NVE/DGW and can be re-used for all the IP-VRFs in the node.

Example of prefix advertisement for the ipv4 prefix SN1/24 advertised from NVE1:

(1) NVE1 advertises the following BGP route for SN1:

- o Route type 5 (IP Prefix route) containing: Eth-Tag=VNI=10 (assuming global VNI), IPL=24, IP=SN1. In addition to that, a Tunnel Encapsulation Attribute will be sent, where: Tunnel-type= VXLAN or NVGRE, and the address value will contain a GW MAC address= NVE1 MAC.

(2) DGW1 imports the received route from NVE1 and SN1/24 is added to the designated routing context. The next-hop for SN1/24 will be given by the route type 5 BGP next-hop (NVE1), which is resolved to a tunnel. For instance: if the tunnel is VXLAN based, the BGP next-hop will be resolved to a VXLAN tunnel where: destination-VTEP= NVE1 IP, VNI=10, inner destination MAC = NVE1 MAC (derived from the GW MAC value in the Tunnel Encapsulation attribute).

(3) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop= "NVE1 IP" is found. The tunnel information to encapsulate the packet will be derived from the route-type 5 received for SN1.
- o The IP packet destined to IPx is encapsulated with: Source inner MAC = DGW1 MAC, Destination inner MAC = NVE1 MAC, Source outer IP (source VTEP) = DGW1 IP, Destination outer IP (destination VTEP) =

NVE1 IP

(4) When the packet arrives at NVE1:

- o Based on the tunnel information (VNI for the VXLAN case), the routing context is identified for an IP lookup.
- o An IP lookup is performed in the routing context, where SN1 turns out to be a local subnet associated to EVI-2. A subsequent lookup in the ARP table and the EVI-2 MAC-VRF will return the forwarding information for the packet in EVI-2.

## 6. Conclusions

A new EVPN route type 5 for the advertisement of IP Prefixes is proposed in this document. This new route type will have a differentiated role from the RT-2 route and will address all the Data Center (or NVO-based networks in general) inter-subnet connectivity scenarios in which IP Prefix advertisement is required. Using this new RT-5 route, an IP Prefix will be advertised along with an overlay next-hop that can be a GW IP address, an ESI or a GW MAC address. As discussed throughout the document, IP-VPN cannot address all the inter-subnet use-cases in an NVO-based DC and the existing EVPN RT-2 does not meet the requirements for all the DC use cases, therefore a new EVPN route type is required.

This new EVPN route type 5 decouples the IP Prefix advertisements from the MAC route advertisements in EVPN, hence:

- a) Allows the clean and clear announcements of ipv4 or ipv6 prefixes in an NLRI with no MAC addresses in the route key, so that only IP information is used in BGP route comparisons.
- b) Since the route type is different from the MAC/IP advertisement route, the advertisement of prefixes will be excluded from all the procedures defined for the advertisement of VM MACs, e.g. MAC Mobility or aliasing. As a result of that, the current EVPN procedures do not need to be modified.
- c) Allows a flexible implementation where the prefix can be linked to different types of next-hops: MAC address, IP address, IRB IP address, ESI, etc. and these MAC or IP addresses do not need to reside in the advertising NVE.
- d) An EVPN implementation not requiring IP Prefixes can simply discard them by looking at the route type value. An unknown route type MUST be ignored by the receiving NVE/PE.

## 7. Conventions used in this document

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

## 8. Security Considerations

## 9. IANA Considerations

## 10. References

### 10.1 Normative References

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

### 10.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03.txt, work in progress, February, 2013

[EVPN-OVERLAYS] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", draft-sd-l2vpn-evpn-overlay-03.txt, work in progress, June, 2014

[EVPN-INTERSUBNET] Sajassi et al., "IP Inter-Subnet Forwarding in EVPN", draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-04.txt, work in progress, July, 2014

## 11. Acknowledgments

The authors would like to thank Mukul Katiyar and Senthil Sathappan for their valuable feedback and contributions.

## 12. Authors' Addresses

Jorge Rabadan  
Alcatel-Lucent  
777 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: jorge.rabadan@alcatel-lucent.com

Wim Henderickx

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

Florin Balus  
Nuage Networks  
Email: florin@nuagenetworks.net

Aldrin Isaac  
Bloomberg  
Email: aisaac71@bloomberg.net

Senad Palislamovic  
Alcatel-Lucent  
Email: senad.palislamovic@alcatel-lucent.com

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

Ali Sajassi  
Cisco  
Email: sajassi@cisco.com

L2VPN Workgroup  
Internet Draft

Intended status: Informational

J. Uttaro  
AT&T

A. Isaac  
T. Boyes  
Bloomberg

J. Rabadan  
S. Palislaamovic  
W. Henderickx  
F. Balus  
Alcatel-Lucent

K. Patel  
A. Sajassi  
Cisco

Expires: December 20, 2014

June 18, 2014

Usage and applicability of BGP MPLS based Ethernet VPN  
draft-rp-l2vpn-evpn-usage-02.txt

Abstract

This document discusses the usage and applicability of BGP MPLS based Ethernet VPN (EVPN) in a simple and fairly common deployment scenario. The different EVPN procedures will be explained on the example scenario, analyzing the benefits and trade-offs of each option. Along with [EVPN], this document is intended to provide a simplified guide for the deployment of EVPN in Service Provider networks.

Status of this Memo

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

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

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



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

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

This Internet-Draft will expire on December 28, 2013.

#### Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

1. Introduction . . . . .	4
2. Use-case scenario description . . . . .	4
3. Provisioning Model . . . . .	6
3.1. Common provisioning tasks . . . . .	7
3.1.1. Non-service specific parameters . . . . .	7
3.1.2. Service specific parameters . . . . .	8
3.2. Service interface dependent provisioning tasks . . . . .	8
3.2.1. VLAN-based service interface EVI . . . . .	8
3.2.2. VLAN-bundle service interface EVI . . . . .	9
3.2.3. VLAN-aware bundling service interface EVI . . . . .	9
4. BGP EVPN NLRI usage . . . . .	9
5. MAC-based forwarding model use-case . . . . .	10
5.1. EVPN Network Startup procedures . . . . .	10
5.2. VLAN-based service procedures . . . . .	11
5.2.1. Service startup procedures . . . . .	11
5.2.2. Packet walkthrough . . . . .	12
5.3. VLAN-bundle service procedures . . . . .	15
5.3.1. Service startup procedures . . . . .	15
5.3.2. Packet Walkthrough . . . . .	16
5.4. VLAN-aware bundling service procedures . . . . .	16
5.4.1. Service startup procedures . . . . .	17
5.4.2. Packet Walkthrough . . . . .	17
6. MPLS-based forwarding model use-case . . . . .	18

6.1. Impact of MPLS-based forwarding on the EVPN network startup . . . . .	19
6.2. Impact of MPLS-based forwarding on the VLAN-based service procedures . . . . .	19
6.3. Impact of MPLS-based forwarding on the VLAN-bundle service procedures . . . . .	20
6.4. Impact of MPLS-based forwarding on the VLAN-aware service procedures . . . . .	20
7. Comparison between MAC-based and MPLS-based forwarding models . . . . .	21
8. Traffic flow optimization . . . . .	22
8.1. Control Plane Procedures . . . . .	22
8.1.1. MAC learning options . . . . .	22
8.1.2. Proxy-ARP/ND . . . . .	23
8.1.3. Unknown Unicast flooding suppression . . . . .	23
8.1.4. Optimization of Inter-subnet forwarding . . . . .	24
8.2. Packet Walkthrough Examples . . . . .	25
8.2.1. Proxy-ARP example for CE2 to CE3 traffic . . . . .	25
8.2.2. Flood suppression example for CE1 to CE3 traffic . . . . .	25
8.2.3. Optimization of inter-subnet forwarding example for CE3 to CE2 traffic . . . . .	26
9. Conventions used in this document . . . . .	27
10. Security Considerations . . . . .	28
11. IANA Considerations . . . . .	28
12. References . . . . .	28
12.1. Normative References . . . . .	28
12.2. Informative References . . . . .	28
13. Acknowledgments . . . . .	29
14. Authors' Addresses . . . . .	29

## 1. Introduction

This document complements [EVPN] by discussing the applicability of the technology in a simple and fairly common deployment scenario, which is described in section 2.

After describing the topology of the use-case scenario and the characteristics of the service to be deployed, section 3 will describe the provisioning model, comparing the EVPN procedures with the provisioning tasks required for other VPN technologies, such as VPLS or IP-VPN.

Once the provisioning model is analyzed, sections 4, 5 and 6 will describe the control plane and data plane procedures in the example scenario, for the two potential disposition/forwarding models: MAC-based and MPLS-based models. While both models can interoperate in the same network, each one has different trade-offs that are analyzed in section 7.

Finally, EVPN provides some potential traffic flow optimization tools that are also described in section 8, in the context of the example scenario.

## 2. Use-case scenario description

The following figure depicts the scenario that will be referenced throughout the rest of the document.

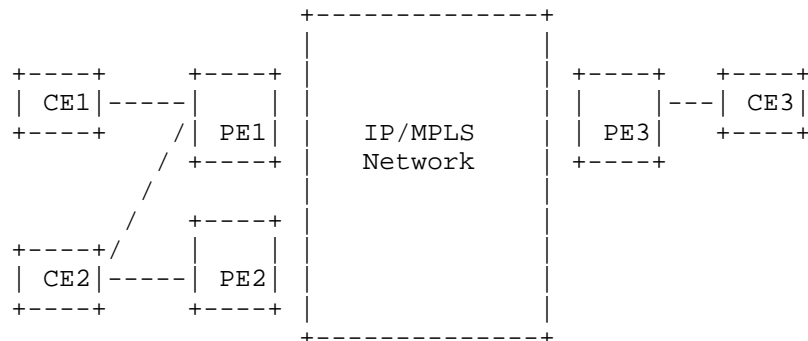


Figure 1 EVPN use-case scenario

There are three PEs and three CEs considered in this example: PE1, PE2, PE3, as well as CE1, CE2 and CE3. Layer-2 traffic must be extended among the three CEs. The following service requirements are assumed in this scenario:

- o Redundancy requirements: CE1 and CE3 are single-homed to PE1 and PE3 respectively. CE2 requires multi-homing connectivity to PE1 and PE2, not only for redundancy purposes, but also for adding more upstream/downstream connectivity bandwidth to/from the network. If CE2 has a single CE-VID (or a few CE-VIDs) the current VPLS multi-homing solutions (based on load-balancing per CE-VID or service) do not provide the optimized link utilization required in this example. Another redundancy requirement that must be met is fast convergence. E.g.: if the link between CE2 and PE1 goes down, a fast convergence mechanism must be supported so that PE3 can immediately send the traffic to PE2, irrespectively of the number of affected services and MAC addresses. EVPN provides the flow-based load-balancing multi-homing solution required in this scenario to optimize the upstream/downstream link utilization between CE2 and PE1-PE2. EVPN also provides a fast convergence solution so that PE3 can immediately send the traffic to PE2 upon failure on the link between CE2 and PE1.

- o Service interface requirements: service definition must be flexible in terms of CE-VID-to-broadcast-domain assignment and service contexts in the core. The following three services are required in this example:

EVI100 - It will use VLAN-based service interfaces in the three CEs with a 1:1 mapping (VLAN-to-EVI). The CE-VIDs at the three CEs can be the same, e.g.: VID 100, or different at each CE, e.g.: VID 101 in CE1, VID 102 in CE2 and VID 103 in CE3. A single broadcast domain needs to be created for EVI100 in any case; therefore CE-VIDs will require translation at the egress PEs if they are not consistent across the three CEs. The case when the same CE-VID is used across the three CEs for EVI100 is referred in [EVPN] as the "Unique VLAN" EVPN case. This term will be used throughout this document too.

EVI200 - It will use VLAN-bundle service interfaces in CE1, CE2 and CE3, based on an N:1 VLAN-to-EVI mapping. In this case, the service provider just needs to assign a pre-configured number of CE-VIDs on the ingress PE to EVI200, and send the customer frames with the original CE-VIDs. The Service Provider will build a single broadcast domain for the customer. The customer will be responsible for the CE-VID handling.

EVI300 - It will use VLAN-aware bundling service interfaces in CE1, CE2 and CE3. At the ingress PE, an N:1 VLAN-to-EVI mapping will be done, however and as opposed to EVI200, a separate core broadcast domain is required per CE-VID. In addition to that, the CE-VIDs can be different (hence CE-VID translation is required). Note that, while the requirements stated for EVI100 and EVI200 might be met

with the current VPLS solutions, the VLAN-aware bundling service interfaces required by EVI300 are not supported by the current VPLS tools.

NOTE: in section 3.2.1, only EVI100 is used as an example of VLAN-based service provisioning. In sections 5.2 and 6.2, 4k VLAN-based EVIs (EVI1 to EVI4k) are used so that the impact of MAC vs. MPLS disposition models in the control plane can be evaluated. In the same way, EVI200 and EVI300 will be described with a 4k:1 mapping (CE-VIDs-to-EVI mapping) in sections 5.3-4 and 6.3-4.

- o BUM (Broadcast, Unknown unicast, Multicast) optimization requirements: The solution must be able to support ingress replication, P2MP MPLS LSPs and MP2MP MPLS LSPs and the user must be able to decide what kind of provider tree will be used by each EVI service. For example, if we assume that EVI100 and EVI200 will not carry much BUM traffic, we can use ingress replication for those service instances. The benefit is that the core will not need to maintain any states for the multicast trees associated to EVI100 and EVI200. On the contrary, if EVI300 is presumably carrying a significant amount of multicast traffic, P2MP MPLS LSPs or MP2MP LSPs can be used for this service. Note that ingress replication and P2MP LSPs are supported by VPLS solutions (see [VPLS-MCAST]), however VPLS solutions do not support MP2MP LSPs, since the source of the tree must be identified for the data plane MAC learning, and that identification is challenging when using MP2MP LSPs. Since EVPN uses the control plane for MAC learning, any type of provider multicast tree is supported in the core.

As already outlined above, the current VPLS solutions, based on [RFC4761][RFC4762][RFC6074], cannot meet all the above set of requirements and therefore a new solution is needed. The rest of the document will describe how EVPN can be used to meet those service requirements and even optimize the network further by:

- o Providing the user with an option to reduce (and even suppress) the ARP-flooding.
- o Supporting ARP termination for inter-subnet forwarding

### 3. Provisioning Model

One of the requirements stated in [RFC7209] is the ease of provisioning. BGP parameters and service context parameters should be auto-provisioned so that the addition of a new MAC-VRF to the EVI requires a minimum number of single-sided provisioning touches. However this is only possible in a limited number of cases. This section describes the provisioning tasks required for the services

described in section 2, i.e. EVI100 (VLAN-based service interfaces), EVI200 (VLAN-bundle service interfaces) and EVI300 (VLAN-aware bundling service interfaces).

### 3.1. Common provisioning tasks

Regardless of the service interface type (VLAN-based, VLAN-bundle or VLAN-aware), the following sub-sections describe the parameters to be provisioned in the three PEs.

#### 3.1.1. Non-service specific parameters

The multi-homing function in EVPN requires the provisioning of certain parameters which are not service-specific and that are shared by all the MAC-VRFs in the node using the multi-homing capabilities. In our use-case, these parameters are only provisioned in PE1 and PE2, and are listed below:

- o Ethernet Segment Identifier (ESI): only the ESI associated to CE2 needs to be considered in our example. Single-homed CEs such as CE1 and CE3 do not require the provisioning of an ESI (the ESI will be coded as zero in the BGP NLRI). In our example, a LAG is used between CE2 and PE1-PE2 (since all-active multi-homing is a requirement) therefore the ESI can be auto-derived from the LACP information as described in [EVPN]. Note that the ESI MUST be unique across all the PEs in the network, therefore the auto-provisioning of the ESI is only recommended in case the CEs are managed by the Service Provider. Otherwise the ESI should be manually provisioned (type 0 as in [EVPN]) in order to avoid potential conflicts.
- o ES-Import Route Target (ES-Import RT): this is the RT that will be sent by PE1 and PE2, along with the ES route. Regardless of how the ESI is provisioned in PE1 and PE2, the ES-Import RT must always be auto-derived from the 6-byte MAC address portion of the ESI value.
- o Ethernet Segment Route Distinguisher (ES RD): this is the RD to be encoded in the ES route and Ethernet Auto-Discovery (A-D) route to be sent by PE1 and PE2 for the CE2 ESI. This RD should always be auto-derived from the PE IP address, as described in [EVPN].
- o Multi-homing type: the user must be able to provision the multi-homing type to be used in the network. In our use-case, the multi-homing type will be set to all-active for the CE2 ESI. This piece of information is encoded in the ESI Label extended community flags and sent by PE1 and PE2 along with the Ethernet A-D route for the CE2 ESI.

In our use-case, besides the above parameters, the same LACP parameters will be configured in PE1 and PE2 for the ESI, so that CE2 can send different flows to PE1 and PE2 for the same CE-VID as though they were forming a single system from the CE2 perspective.

### 3.1.2. Service specific parameters

The following parameters must be provisioned in PE1, PE2 and PE3 per EVI service:

- o EVI identifier: global identifier per EVI that is shared by all the PEs part of the EVI, i.e. PE1, PE2 and PE3 will be provisioned with EVI100, 200 and 300. The EVI identifier can be associated to (or be the same value as) the EVI default Ethernet Tag (4-byte default broadcast domain identifier for the EVI). The Ethernet Tag is different from zero in the EVPN BGP routes only if the service interface type (of the source PE) is VLAN-aware.
- o EVI Route Distinguisher (EVI RD): This RD is a unique value across all the MAC-VRFs in a PE. Auto-derivation of this RD might be possible depending on the service interface type being used in the EVI. Next section discusses the specifics of each service interface type.
- o EVI Route Target(s) (EVI RT): one or more RTs can be provisioned per MAC-VRF. The RT(s) imported and exported can be equal or different, just as the RT(s) in IP-VPNs. Auto-derivation of this RT(s) might be possible depending on the service interface type being used in the EVI. Next section discusses the specifics of each service interface type.
- o CE-VID and port/LAG binding to EVI identifier or Ethernet Tag: see section 3.2.

### 3.2. Service interface dependent provisioning tasks

Depending on the service interface type being used in the EVI, a specific CE-VID binding provisioning must be specified.

#### 3.2.1. VLAN-based service interface EVI

In our use-case, EVI100 is a VLAN-based service interface EVI.

EVI100 can be a "unique-VLAN" EVPN if the CE-VID being used for this service in CE1, CE2 and CE3 is equal, e.g. VID 100. In that case, the VID 100 binding must be provisioned in PE1, PE2 and PE3 for EVI100 and the associated port or LAG. The MAC-VRF RD and RT can be auto-derived from the CE-VID:

- o The auto-derived MAC-VRF RD will be a Type 1 RD, as recommended in [EVPN], and it will be comprised of [PE-IP]:[zero-padded-VID]; where PE-IP is the IP address of the PE (a loopback address) and [zero-padded-VID] is a 2-byte value where the low order 12 bits are the VID (VID 100 in our example) and the high order 4 bits are zero.
- o The auto-derived MAC-VRF RT will be composed of [AS]:[zero-padded-VID]; where AS is the Autonomous System that the PE belongs to and [zero-padded-VID] is a 4-byte value where the low order 12 bits are the VID (VID 100 in our example) and the high order 20 bits are zero. Note that auto-deriving the RT implies supporting a basic any-to-any topology in the EVI and using the same import and export RT in the EVI.

If EVI100 is not a "unique-VLAN" EVPN, each individual CE-VID must be configured in each PE, and MAC-VRF RDs and RTs cannot be auto-derived, hence they must be provisioned by the user.

### 3.2.2. VLAN-bundle service interface EVI

Assuming EVI200 is a VLAN-bundle service interface EVI, and VIDs 200-250 are assigned to EVI200, the CE-VID bundle 200-250 must be provisioned on PE1, PE2 and PE3. Note that this model does not allow CE-VID translation and the CEs must use the same CE-VIDs for EVI200. No auto-derived EVI RDs or EVI RTs are possible.

### 3.2.3. VLAN-aware bundling service interface EVI

If EVI300 is a VLAN-aware bundling service interface EVI, CE-VID binding to EVI300 does not have to match on the three PEs (only on PE1 and PE2, since they are part of the same ES). E.g.: PE1 and PE2 CE-VID binding to EVI300 can be set to the range 300-310 and PE3 to 321-330. Note that each individual CE-VID will be assigned to a core broadcast domain, i.e. Ethernet Tag, which will be encoded in the BGP EVPN routes.

Therefore, besides the CE-VID bundle range bound to EVI300 in each PE, associations between each individual CE-VID and the EVPN Ethernet Tag must be provisioned by the user. No auto-derived EVI RDs/RTs are possible.

## 4. BGP EVPN NLRI usage

[EVPN] defines four different types of routes and four different extended communities advertised along with the different routes. However not all the PEs in a network must generate and process all the different routes and extended communities. The following table



shows the routes that must be exported and imported in the use-case described in this document. "Export", in this context, means that the PE must be capable of generating and exporting a given route, assuming there are no BGP policies to prevent it. In the same way, "Import" means the PE must be capable of importing and processing a given route, assuming the right RTs and policies. "N/A" means neither import nor export actions are required.

BGP EVPN routes	PE1-PE2	PE3
ES	Export/import	N/A
A-D per ESI	Export/import	Import
A-D per EVI	Export/import	Import
MAC	Export/import	Export/import
Inclusive mcast	Export/import	Export/import

PE3 is only required to export MAC and Inclusive multicast routes and be able to import and process A-D routes, as well as MAC and Inclusive multicast routes. If PE3 did not support importing and processing A-D routes per ESI and per EVI, fast convergence and aliasing functions (respectively) would not be possible in this use-case.

## 5. MAC-based forwarding model use-case

This section describes how the BGP EVPN routes are exported and imported by the PEs in our use-case, as well as how traffic is forwarded assuming that PE1, PE2 and PE3 support a MAC-based forwarding model. In order to compare the control and data plane impact in the two forwarding models (MAC-based and MPLS-based) and different service types, we will assume that CE1, CE2 and CE3 need to exchange traffic for up to 4k CE-VIDs.

### 5.1. EVPN Network Startup procedures

Before any EVI is provisioned in the network, the following procedures are required:

- o Infrastructure setup: the proper MPLS infrastructure must be setup among PE1, PE2 and PE3 so that the EVPN services can make use of P2P, P2MP and/or MP2MP LSPs. In addition to the MPLS transport, PE1 and PE2 must be properly configured with the same LACP configuration to CE2. Details are provided in [EVPN]. Once the LAG is properly setup, the ESI for the CE2 Ethernet Segment, e.g. ESI12, can be auto-generated by PE1 and PE2 from the LACP information exchanged with CE2 (ESI type 1), as discussed in

section 3.1. Alternatively, the ESI can also be manually provisioned on PE1 and PE2 (ESI type 0). PE1 and PE2 will auto-configure a BGP policy that will import any ES route matching the auto-derived ES-import RT for ESI12.

- o Ethernet Segment route exchange and DF election: PE1 and PE2 will advertise a BGP Ethernet Segment route for ESI12, where the ESI RD and ES-Import RT will be auto-generated as discussed in section 3.1.1. PE1 and PE2 will import the ES routes of each other and will run the DF election algorithm for any existing EVI (if any, at this point). PE3 will simply discard the route. Note that the DF election algorithm can support service carving, so that the downstream BUM traffic from the network to CE2 can be load-balanced across PE1 and PE2 on a per-service basis.

At the end of this process, the network infrastructure is ready to start deploying EVPN services. PE1 and PE2 are aware of the existence of a shared Ethernet Segment, i.e. ESI12.

## 5.2. VLAN-based service procedures

Assuming that the EVPN network must carry traffic among CE1, CE2 and CE3 for up to 4k CE-VIDs, the Service Provider can decide to implement VLAN-based service interface EVIs to accomplish it. In this case, each CE-VID will be individually mapped to a different EVI. While this means a total number of 4k MAC-VRFs is required per PE, the advantages of this approach are the auto-provisioning of most of the service parameters if no VLAN translation is needed (see section 3.2.1) and great control over each individual customer broadcast domain. We assume in this section that the range of EVIs from 1 to 4k is provisioned in the network.

### 5.2.1. Service startup procedures

As soon as the EVIs are created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (up to 4k routes per PE) so that the flooding tree per EVI can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created.
- o Ethernet A-D routes per ESI (a set of routes for ESI12): A set of A-D routes with a list of 4k RTs (one per EVI) for ESI12 will be issued from PE1 and PE2 (it has to be a set of routes so that the total number of RTs can be conveyed). This set will also include

ESI Label extended communities with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used for split-horizon functions). These routes will be imported by the three PEs, since the RTs match the EVI RTs locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as discussed in [EVPN].

- o Ethernet A-D routes per EVI (4k routes): An A-D route per EVI will be sent by PE1 and PE2 for ESI12. Each individual route includes the corresponding EVI RT and an MPLS label to be used by PE3 for the aliasing function. These routes will be imported by the three PEs.

#### 5.2.2. Packet walkthrough

Once the services are setup, the traffic can start flowing. Assuming there are no MAC addresses learnt yet and that MAC learning at the access is performed in the data plane in our use-case, this is the process followed upon receiving packets from each CE (example for EVI1).

(1) BUM packet example from CE1:

- a) An ARP-request with CE-VID=1 is issued from source MAC CE1-MAC (MAC address coming from CE1 or from a device connected to CE1) to find the MAC address of CE3-IP.
- b) Based on the CE-VID, the packet is identified to be forwarded in the MAC-VRF-1 (EVI1) context. A source MAC lookup is done in the MAC FIB and the sender's CE1-IP in the proxy-ARP table within the MAC-VRF-1 (EVI1) context. If CE1-MAC/CE1-IP are unknown in both tables, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port and CE-VID, (2) the ARP-request is snooped and the tuple CE1-MAC/CE1-IP is added to the proxy-ARP table and (3) a BGP MAC advertisement route is triggered from PE1 containing the EVI1 RD and RT, ESI=0, Ethernet-Tag=0 and CE1-MAC/CE1-IP along with an MPLS label assigned to MAC-VRF-1 from the PE1 label space. Note that depending on the implementation, the MAC FIB and proxy-ARP learning processes can independently send two BGP MAC advertisements instead of one (one containing only the CE1-MAC and another one containing CE1-MAC/CE1-IP).

Since we assume a MAC forwarding model, a label per MAC-VRF is normally allocated and signaled by the three PEs for MAC advertisement routes. Based on the RT, the route is imported by PE2 and PE3 and the forwarding state plus ARP entry are added to their MAC-VRF-1 context. From this moment on, any ARP request from

CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding for CE1-IP is not needed in the core.

- c) Since the ARP packet is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI1 (CE-VID=1 should be kept if translation is required). Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication, the packet is replicated to PE2 and PE3 with the downstream allocated labels and the P2P LSP transport labels. No other labels are added to the stack.
- d) Assuming PE1 is the DF for EVI1 on ESI12, the packet is locally replicated to CE2.
- e) The MPLS-encapsulated packet gets to PE2 and PE3. Since PE2 is non-DF for EVI1 on ESI12, and there is no other CE connected to PE2, the packet is discarded. At PE3, the packet is de-encapsulated, CE-VID translated if needed and replicated to CE3.

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet example from CE2:

- a) An ARP-request with CE-VID=1 is issued from source MAC CE2-MAC to find the MAC address of CE3-IP.
- b) CE2 will hash the packet and will forward it to e.g. PE2. Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB and the sender's CE2-IP in the proxy-ARP table within the MAC-VRF-1 context. If both are unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI and CE-VID, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the proxy-ARP table and (3) a BGP MAC advertisement route is triggered from PE2 containing the EVI1 RD and RT, ESI=12, Ethernet-Tag=0 and CE2-MAC/CE2-IP along with an MPLS label assigned from the PE2 label space (one label per MAC-VRF). Again, depending on the implementation, the MAC FIB and proxy-ARP learning processes can independently send two BGP MAC advertisements instead of one.

Note that, since PE3 is not part of ESI12, it will install a forwarding state for CE2-MAC as long as the A-D routes for ESI12 are also active on PE3. On the contrary, PE1 is part of ESI12,

therefore PE1 will not modify the forwarding state for CE2-MAC if it has previously learnt CE2-MAC locally attached to ESI12. Otherwise it will add forwarding state for CE2-MAC associated to the local ESI12 port.

- c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI1 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI1, adding the ESI label for ESI12 at the bottom of the stack. The ESI label has been previously allocated and signaled by the A-D routes for ESI12. Note that, as per [EVPN], if the result of the CE2 hashing is different and the packet sent to PE1, PE1 SHOULD add the ESI label too (PE1 is the DF for EVI1 on ESI12).
- d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI identified by the ESI label. At PE3, the Inclusive multicast tree label is popped and the packet forwarded to CE3. If a P2MP LSP is used as Inclusive multicast tree for EVI1, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will simply be ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

- a) A unicast packet with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE1-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE1-IP, and has added CE3-MAC/CE3-IP to its proxy-ARP table).
- b) Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB within the MAC-VRF-1 context and this time, since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and the label stack associated to the MAC CE1-MAC is found (including the label associated to MAC-VRF-1 in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1.
- c) At PE1, the packet is identified to be part of EVI1 and a destination MAC lookup is performed in the MAC-VRF-1 context. The labels are popped and the packet forwarded to CE1 with CE-VID=1.

Unicast packets from CE1 to CE3 or from CE2 to CE3 follow the same procedures described above.

(4) Unicast packet example from CE3 to CE2:

- a) A unicast packet with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE2-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE2-IP).
- b) Based on the CE-VID, the packet is identified to be forwarded in the MAC-VRF-1 context. We assume CE3-MAC is known. A destination MAC lookup is performed next and PE3 finds CE2-MAC associated to PE2 on ESI12, an Ethernet Segment for which PE3 has two active A-D routes per ESI (from PE1 and PE2) and two active A-D routes for EVI1 (from PE1 and PE2). Based on a hashing function for the packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI1). Either way, the packet is encapsulated and sent to the remote PE.
- c) At PE2 (or PE1), the packet is identified to be part of EVI1 based on the bottom label, and a destination MAC lookup is performed. At either PE (PE2 or PE1), the FIB lookup yields a local ESI12 port to which the packet is sent.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

### 5.3. VLAN-bundle service procedures

Instead of using VLAN-based interfaces, the Service Provider can choose to implement VLAN-bundle interfaces to carry the traffic for the 4k CE-VIDs among CE1, CE2 and CE3. If that is the case, the 4k CE-VIDs can be mapped to the same EVI, e.g. EVI200, at each PE. The main advantage of this approach is the low control plane overhead (reduced number of routes and labels) and easiness of provisioning, at the expense of no control over the customer broadcast domains, i.e. a single inclusive multicast tree for all the CE-VIDs and no CE-VID translation in the Provider network.

#### 5.3.1. Service startup procedures

As soon as the EVI200 is created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI (one route): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (hence only one route per PE) so that the flooding tree per EVI can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally

be signaled in the PMSI Tunnel attribute and the corresponding tree be created.

- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI200), an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the EVI200 RT locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [EVPN].
- o Ethernet A-D routes per EVI (one route): An A-D route (EVI200) will be sent by PE1 and PE2 for ESI12. This route includes the EVI200 RT and an MPLS label to be used by PE3 for the aliasing function. This route will be imported by the three PEs.

#### 5.3.2. Packet Walkthrough

The packet walkthrough for the VLAN-bundle case is similar to the one described for EVI1 in the VLAN-based case except for the way the CE-VID is handled by the ingress PE and the egress PE:

- o No VLAN translation is allowed and the CE-VIDs are kept untouched from CE to CE, i.e. the ingress CE-VID MUST be kept at the imposition PE and at the disposition PE.
- o The packet is identified to be forwarded in the MAC-VRF-200 context as long as its CE-VID belongs to the VLAN-bundle defined in the PE1/PE2/PE3 port to CE1/CE2/CE3. Our example is a special VLAN-bundle case, since the entire CE-VID range is defined in the ports, therefore any CE-VID would be part of EVI200.

Please refer to section 5.2.2 for more information about the control plane and forwarding plane interaction for BUM and unicast traffic from the different CEs.

#### 5.4. VLAN-aware bundling service procedures

The last potential service type analyzed in this document is VLAN-aware bundling. When this type of service interface is used to carry the 4k CE-VIDs among CE1, CE2 and CE3, all the CE-VIDs will be mapped to the same EVI, e.g. EVI300. The difference, compared to the VLAN-bundle service type in the previous section, is that each incoming CE-VID will also be mapped to a different "normalized" Ethernet-Tag in addition to EVI300. If no translation is required, the Ethernet-tag will match the CE-VID. Otherwise a translation

between CE-VID and Ethernet-tag will be needed at the imposition PE and at the disposition PE. The main advantage of this approach is the ability to control customer broadcast domains while providing a single EVI to the customer.

#### 5.4.1. Service startup procedures

As soon as the EVI300 is created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI per Ethernet-Tag (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI and per Ethernet-Tag (hence 4k routes per PE) so that the flooding tree per customer broadcast domain can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use-case, since all the CE-VIDs and Ethernet-Tags are defined on the three PEs, multicast tree aggregation might make sense in order to save forwarding states.
- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI300), an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different than zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the EVI300 RT locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [EVPN].
- o Ethernet A-D routes per EVI (one route): An A-D route (EVI300) will be sent by PE1 and PE2 for ESI12. This route includes the EVI300 RT and an MPLS label to be used by PE3 for the aliasing function. This route will be imported by the three PEs.

#### 5.4.2. Packet Walkthrough

The packet walkthrough for the VLAN-aware case is similar to the one described before. Compared to the other two cases, VLAN-aware services allow for CE-VID translation and for an N:1 CE-VID to EVI mapping. Both things are not supported at once in either of the two other service interfaces. Note that this model requires qualified learning on the MAC FIBs. Some differences compared to the packet walkthrough described in section 5.2.2 are:

- o At the ingress PE, the packets are identified to be forwarded in the EVI300 context as long as their CE-VID belong to the range defined in the PE port to the CE. In addition to it, CE-VID=x is



mapped to a "normalized" Ethernet-Tag=y at the MAC-VRF-300 (where x and y might be equal if no translation is needed). Qualified learning is now required (a different FIB space is allocated within MAC-VRF-300 for each Ethernet-Tag). Potentially the same MAC could be learnt in two different Ethernet-Tag bridge domains of the same MAC-VRF.

- o Any new locally learnt MAC on the MAC-VRF-300/Ethernet-Tag=y interface is advertised by the ingress PE in a MAC advertisement route, using now the Ethernet-Tag field (Ethernet-Tag=y) so that the remote PE learns the MAC associated to the MAC-VRF-300/Ethernet-Tag=y FIB. Note that the Ethernet-Tag field is not used in advertisements of MACs learnt on VLAN-based or VLAN-bundle service interfaces.
- o At the ingress PE, BUM packets are sent to the corresponding flooding tree for the particular Ethernet-Tag they are mapped to. Each individual Ethernet-Tag can have a different flooding tree within the same EVI300. For instance, Ethernet-Tag=y can use ingress replication to get to the remote PEs whereas Ethernet-Tag=z can use a p2mp LSP.
- o At the egress PE, Ethernet-Tag=y, for a given broadcast domain within MAC-VRF-300, can be translated to egress CE-VID=x. That is not possible for VLAN-bundle interfaces. It is possible for VLAN-based interfaces, but it requires a separate EVI per CE-VID.

## 6. MPLS-based forwarding model use-case

EVPN supports an alternative forwarding model, usually referred to as MPLS-based forwarding or disposition model as opposed to the MAC-based forwarding or disposition model described in section 5. Using MPLS-based forwarding model instead of MAC-based model might have an impact on:

- o The number of forwarding states required
- o The FIB where the forwarding states are handled: MAC FIB or MPLS LFIB.

The MPLS-based forwarding model avoids the destination MAC lookup at the egress PE MAC FIB, at the expense of increasing the number of next-hop forwarding states at the egress MPLS LFIB. This also has an impact on the control plane and the label allocation model, since an MPLS-based disposition PE MUST send as many routes and labels as required next-hops in the egress MAC-VRF. This concept is equivalent to the forwarding models supported in IP-VPNs at the egress PE, where an IP lookup in the IP-VPN FIB might be necessary or not depending on

the available next-hop forwarding states in the LFIB.

The following sub-sections highlight the impact on the control and data plane procedures described in section 5 when and MPLS-based forwarding model is used.

Note that both forwarding models are compatible and interoperable in the same network. The implementation of either model in each PE is a local decision to the PE node.

#### 6.1. Impact of MPLS-based forwarding on the EVPN network startup

The MPLS-based forwarding model has no impact on the procedures explained in section 5.1.

#### 6.2. Impact of MPLS-based forwarding on the VLAN-based service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of number of routes, when all the service interfaces are VLAN-based. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (4k routes per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one set of routes for ESI12 per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per EVI (4k routes per PE/ESI): no impact compared to the MAC-based model.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same MAC-VRF, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. E.g. if CE2 sends traffic from two different MACs to PE1, CE2-MAC1 and CE2-MAC2, the same MPLS label=x can be re-used for both MAC advertisements since they both share the same source ESI12. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment (even if only one label per ESI is enough).
- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will

rather add forwarding states to the MPLS LFIB.

#### 6.3. Impact of MPLS-based forwarding on the VLAN-bundle service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of number of routes when all the service interfaces are VLAN-bundle type. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (one route): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per EVI (one route per PE/ESI): no impact compared to the MAC-based model since no VLAN translation is required.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same MAC-VRF, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment (even if only one label per ESI is enough).
- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

#### 6.4. Impact of MPLS-based forwarding on the VLAN-aware service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has definitively an impact in terms of number of A-D routes when all the service interfaces are VLAN-aware bundle type. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (4k routes per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact

compared to the MAC-based model.

- o Ethernet A-D routes per EVI (4k routes per PE/ESI): PE1 and PE2 will send 4k routes for EVI300, one per <ESI, Ethernet-Tag ID> tuple. This will allow the egress PE to find out all the forwarding information in the MPLS LFIB and even support Ethernet-Tag to CE-VID translation at the egress. The MAC-based forwarding model would allow the PEs to send a single route per PE/ESI for EVI300, since the packet with the embedded Ethernet-Tag would be used to perform a MAC lookup and find out the egress CE-VID.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same MAC-VRF, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment. Note that, in this model, the Ethernet-Tag will be set to a non-zero value for the MAC-advertisement routes. The same MAC address can be announced with different Ethernet-Tag value. This will make the advertising PE install two different forwarding states in the MPLS LFIB.
- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

## 7. Comparison between MAC-based and MPLS-based forwarding models

Both forwarding models are possible in a network deployment and each one has its own trade-offs.

The MAC-based forwarding model can save A-D routes per EVI when VLAN-aware bundling services are deployed and therefore reduce the control plane overhead. This model also saves a significant amount of MPLS labels compared to the MPLS-based forwarding model. All the MACs and A-D routes for the same EVI can signal the same MPLS label, saving labels from the local PE space. A MAC FIB lookup at the egress PE is required in order to do so.

The MPLS-based forwarding model can save forwarding states at the egress PEs if labels per next hop CE (as opposed to per MAC) are implemented. No egress MAC lookup is required. An A-D route per <EVI, Ethernet-Tag> is required for VLAN-aware services, as opposed to an

A-D route per EVI. Also, a different label per next-hop CE per MAC-VRF is consumed, as opposed to a single label per MAC-VRF.

The following table summarizes the implementation details of both models for the VLAN-aware bundling service type.

4k CE-VID VLANs	MAC-based Model	MPLS-based Model
A-D routes/EVI	1 per ESI/EVI	4k per ESI/EVI
MPLS labels consumed	1 per MAC-VRF	1 per CE/EVI
Egress PE Forwarding states	1 per MAC	1 per next-hop
Egress PE Lookups	2 (MPLS+MAC)	1 (MPLS)

The egress forwarding model is an implementation local to the egress PE and is independent of the model supported on the rest of the PEs, i.e. in our use-case, PE1, PE2 and PE3 could have either egress forwarding model without any dependencies.

## 8. Traffic flow optimization

In addition to the procedures described across sections 1 through 7, EVPN [EVPN] procedures allow for optimized traffic handling in order to minimize unnecessary flooding across the entire infrastructure. Optimization is provided through specific ARP termination and the ability to block unknown unicast flooding. Additionally, EVPN procedures allow for intelligent, close to the source, inter-subnet forwarding and solves the commonly known sub-optimal routing problem. Besides the traffic efficiency, ingress based inter-subnet forwarding also optimizes packet forwarding rules and implementation at the egress nodes as well. Details of these procedures are outlined in sections 8.1 and 8.2.

### 8.1. Control Plane Procedures

#### 8.1.1. MAC learning options

The fundamental premise of [EVPN] is the notion of a different approach to MAC address learning compared to traditional IEEE 802.1 bridge learning methods; specifically EVPN differentiates between data and control plane driven learning mechanisms.

Data driven learning implies that there is no separate communication channel used to advertise and propagate MAC addresses. Rather, MAC addresses are learned through IEEE defined bridge-learning procedures as well as by snooping on DHCP and ARP requests. As different MAC

addresses show up on different ports, the L2 FIB is populated with the appropriate MAC addresses.

Control plane driven learning implies a communication channel that could be either a control-plane protocol or a management-plane mechanism. In the context of EVPN, two different learning procedures are defined, i.e. local and remote procedures:

- o Local learning defines the procedures used for learning the MAC addresses of network elements locally connected to a MAC-VRF. Local learning could be implemented through all three learning procedures: control plane, management plane as well as data plane. However, the expectation is that for most of the use cases, local learning through data plane should be sufficient.
- o Remote learning defines the procedures used for learning MAC addresses of network elements remotely connected to a MAC-VRF, i.e. far-end PEs. Remote learning procedures defined in [EVPN] advocate using only control plane learning; specifically BGP. Through the use of BGP EVPN NLRIs, the remote PE has the capability of advertising all the MAC addresses present in its local FIB.

#### 8.1.1.2. Proxy-ARP/ND

In EVPN, MAC addresses are advertised via the MAC/IP Advertisement Route, as discussed in [EVPN]. Optionally an IP address can be advertised along with the MAC address announcement. However, there are certain rules put in place in terms of IP address usage: if the MAC Advertisement Route contains an IP address, and the IP Address Length is 32 bits (or 128 in the IPv6 case), this particular IP address correlates directly with the advertised MAC address. Such advertisement allows us to build a proxy-ARP/ND table populated with the IP<>MAC bindings received from all the remote nodes.

Furthermore, based on these bindings, a local MAC-VRF can now provide Proxy-ARP/ND functionality for all ARP requests and ND solicitations directed to the IP address pool learned through BGP. Therefore, the amount of unnecessary L2 flooding, ARP/ND requests/solicitations in this case, can be further reduced by the introduction of Proxy-ARP/ND functionality across all EVI MAC-VRFs.

#### 8.1.1.3. Unknown Unicast flooding suppression

Given that all locally learned MAC addresses are advertised through BGP to all remote PEs, suppressing flooding of any Unknown Unicast traffic towards the remote PEs is a feasible network optimization.

The assumption in the use case is made that any network device that appears on a remote MAC-VRF will somehow signal its presence to the network. This signaling can be done through e.g. gratuitous ARPs. Once the remote PE acknowledges the presence of the node in the MAC-VRF, it will do two things: install its MAC address in its local FIB and advertise this MAC address to all other BGP speakers via EVPN NLRI. Therefore, we can assume that any active MAC address is propagated and learnt through the entire EVI. Given that MAC addresses become pre-populated - once nodes are alive on the network - there is no need to flood any unknown unicast towards the remote PEs. If the owner of a given destination MAC is active, the BGP route will be present in the local RIB and FIB, assuming that the BGP import policies are successfully applied; otherwise, the owner of such destination MAC is not present on the network.

It is worth noting that unless: a) control or management plane learning is performed through the entire EVI or b) all the EVI-attached devices signal their presence when they come up (GARPs or similar), unknown unicast flooding MUST be enabled.

#### 8.1.4. Optimization of Inter-subnet forwarding

In a scenario in which both L2 and L3 services are needed over the same physical topology, some interaction between EVPN and IP-VPN is required. A common way of stitching the two service planes is through the use of an IRB interface, which allows for traffic to be either routed or bridged depending on its destination MAC address. If the destination MAC address is the one of the IRB interface, traffic needs to be passed through a routing module and potentially be either routed to a remote PE or forwarded to a local subnet. If the destination MAC address is not the one of the IRB, the MAC-VRF follows standard bridging procedures.

A typical example of EVPN inter-subnet forwarding would be a scenario in which multiple IP subnets are part of a single or multiple EVIs, and they all belong to a single IP-VPN. In such topologies, it is desired that inter-subnet traffic can be efficiently routed without any tromboning effects in the network. Due to the overlapping physical and service topology in such scenarios, all inter-subnet connectivity will be locally routed through the IRB interface.

In addition to optimizing the traffic patterns in the network, local inter-subnet forwarding also optimizes greatly the amount of processing needed to cross the subnets. Through EVPN MAC advertisements, the local PE learns the real destination MAC address associated with the remote IP address and the inter-subnet forwarding can happen locally. When the packet is received at the egress PE, it is directly mapped to an egress MAC-VRF, bypassing any egress IP-VPN

processing.

Please refer to [EVPN-INTERSUBNET] for more information about the IP inter-subnet forwarding procedures in EVPN.

## 8.2. Packet Walkthrough Examples

Assuming that the services are setup according to figure 1 in section 2, the following flow optimization processes will take place in terms of creating, receiving and forwarding packets across the network.

### 8.2.1. Proxy-ARP example for CE2 to CE3 traffic

Using figure 1 in section 2, consider EVI 400 residing on PE1, PE2 and PE3 connecting CE2 and CE3 networks. Also, consider that PE1 and PE2 are part of the all-active multi-homing ES for CE2, and that PE2 is elected designated-forwarder for EVI400. We assume that all the PEs implement the proxy-ARP functionality in the MAC-VRF-400 context.

In this scenario, PE3 will not only advertise the MAC addresses through the EVPN MAC Advertisement Route but also IP addresses of individual hosts, i.e. /32 prefixes, behind CE3. Upon receiving the EVPN routes, PE1 and PE2 will install the MAC addresses in the MAC-VRF-400 FIB and based on the associated received IP addresses, PE1 and PE2 can now build a proxy-ARP table within the context of MAC-VRF-400.

From the forwarding perspective, when a node behind CE2 sends a packet destined to a node behind CE3, it will first send an ARP request to e.g. PE2 (based on the result of the CE2 hashing). Assuming that PE2 has populated its proxy-ARP table for all active nodes behind the CE3, and that the IP address in the ARP message matches the entry in the table, PE2 will respond to the ARP request with the actual MAC address on behalf of the node behind CE3.

Once the nodes behind CE2 learn the actual MAC address of the nodes behind CE3, all the MAC-to-MAC communications between the two networks will be unicast.

### 8.2.2. Flood suppression example for CE1 to CE3 traffic

Using figure 1 in section 2, consider EVI 500 residing on PE1 and PE3 connecting CE1 and CE3 networks. Consider that both PE1 and PE3 have disabled unknown unicast flooding for this specific EVI context. Once the network devices behind CE3 come online they will learn their MAC addresses and create local FIB entries for these devices. Note that local FIB entries could also be created through either a control or management plane between PE and CE as well. Consequently, PE3 will



automatically create EVPN Type 2 MAC Advertisement Routes and advertise all locally learned MAC addresses. The routes will also include the corresponding MPLS label.

Given that PE1 automatically learns and installs all MAC addresses behind CE3, its MAC-VRF FIB will already be pre-populated with the respective next-hops and label assignments associated with the MAC addresses behind CE3. As such, as soon as the traffic sent by CE1 to nodes behind CE3 is received into the context of EVI 500, PE1 will push the MPLS Label(s) onto the original Ethernet frame and send the packet to the MPLS network. As usual, once PE3 receives this packet, and depending on the forwarding model, PE3 will either do a next-hop lookup in the EVI 500 context, or will just forward the traffic directly to the CE3. In the case that PE1 MAC-VRF-500 does not have a MAC entry for a specific destination that CE1 is trying to reach, PE1 will drop the packet since unknown unicast flooding is disabled.

Based on the assumption that all the MAC entries behind the CEs are pre-populated through gratuitous-ARP and/or DHCP requests, if one specific MAC entry is not present in the MAC-VRF-500 FIB on PE1, the owner of that MAC is not alive on the network behind the CE3, hence the traffic can be dropped at PE1 instead of be flooded and consume network bandwidth.

#### 8.2.3. Optimization of inter-subnet forwarding example for CE3 to CE2 traffic

Using figure 1 in section 2 consider that there is an IP-VPN 666 context residing on PE1, PE2 and PE3 which connects CE1, CE2 and CE3 into a single IP-VPN domain. Also consider that there are two EVIs present on the PEs, EVI 600 and EVI 60. Each IP subnet is associated to a different MAC-VRF context. Thus there is a single subnet, subnet 600, between CE1 and CE3 that is established through EVI 600. Similarly, there is another subnet, subnet 60, between CE2 and CE3 that is established through EVI 60. Since both subnets are part of the same IP VPN, there is a mapping of each EVI (or individual subnet) to a local IRB interface on the three PEs.

If a node behind CE2 wants to communicate with a node on the same subnet seating behind CE3, the communication flow will follow the standard EVPN procedures, i.e. FIB lookup within the PE1 (or PE2) after adding the corresponding EVPN label to the MPLS label stack (downstream label allocation from PE3 for EVI 60).

When it comes to crossing the subnet boundaries, the ingress PE implements local inter-subnet forwarding. For example, when a node behind CE2 (EVI 60) sends a packet to a node behind CE1 (EVI 600) the destination IP address will be in the subnet 600, but the destination

MAC address will be the address of source node's default gateway, which in this case will be an IRB interface on PE1 (connecting EVI 60 to IP-VPN 666). Once PE1 sees the traffic destined to its own MAC address, it will route the packet to EVI 600, i.e. it will change the source MAC address to the one of the IRB interface in EVI 600 and change the destination MAC address to the address belonging to the node behind CE1, which is already populated in the MAC-VRF-600 FIB, either through data or control plane learning.

An important optimization to be noted is the local inter-subnet forwarding in lieu of IP VPN routing. If the node from subnet 60 (behind CE2) is sending a packet to the remote end node on subnet 600 (behind CE3), the mechanism in place still honors the local inter-subnet (inter-EVI) forwarding.

In our use-case, therefore, when node from subnet 60 behind CE2 sends traffic to the node on subnet 600 behind CE3, the destination MAC address is the PE1 MAC-VRF-60 IRB MAC address. However, once the traffic locally crosses EVIs, to EVI 600, via the IRB interface on PE1, the source MAC address is changed to that of the IRB interface and the destination MAC address is changed to the one advertised by PE3 via EVPN and already installed in MAC-VRF-600. The rest of the forwarding through PE1 is using the MAC-VRF-600 forwarding context and label space.

Another very relevant optimization is due to the fact that traffic between PEs is forwarded through EVPN, rather than through IP-VPN. In the example described above for traffic from EVI 60 on CE2 to EVI 600 on CE3, there is no need for IP-VPN processing on the egress PE3. Traffic is forwarded either to the EVI 600 context in PE3 for further MAC lookup and next-hop processing, or directly to the node behind CE3, depending on the egress forwarding model being used.

## 9. Conventions used in this document

In the examples, the following conventions are used:

- o CE-VIDs refer to the VLAN tag identifiers being used at CE1, CE2 and CE3 to tag customer traffic sent to the Service Provider E-VPN network
- o CE1-MAC, CE2-MAC and CE3-MAC refer to source MAC addresses "behind" each CE respectively. Those MAC addresses can belong to the CEs themselves or to devices connected to the CEs.
- o CE1-IP, CE2-IP and CE3-IP refer to IP addresses associated to the above MAC addresses.

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

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

## 10. Security Considerations

## 11. IANA Considerations

## 12. References

### 12.1. Normative References

[RFC4761]Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

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

[RFC6074]Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

[RFC4364]Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC709] A. Sajassi, R. Aggarwal et al., "Requirements for Ethernet VPN", RFC7209, May 2014

### 12.2. Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-07.txt, work in progress, May, 2014

[VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et al., draft-ietf-l2vpn-vpls-mcast-13.txt

[EVPN-INTERSUBNET] Sajassi et al., "IP Inter-subnet forwarding in EVPN", draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-02.txt

### 13. Acknowledgments

The authors want to thank Giles Heron for his detailed review of the document.

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

### 14. Authors' Addresses

Jorge Rabadan  
Alcatel-Lucent  
777 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: jorge.rabadan@alcatel-lucent.com

Senad Palislamovic  
Alcatel-Lucent  
Email: senad.palislamovic@alcatel-lucent.com

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

Florin Balus  
Alcatel-Lucent  
Email: Florin.Balus@alcatel-lucent.com

Keyur Patel  
Cisco  
Email: keyupate@cisco.com

Ali Sajassi  
Cisco  
Email: sajassi@cisco.com

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

Aldrin Isaac  
Bloomberg  
Email: aisaac71@bloomberg.net

Truman Boyes  
Bloomberg  
Email: tboyes@bloomberg.net

L2VPN Workgroup  
INTERNET-DRAFT  
Intended Status: Standards Track

Ali Sajassi  
Samer Salam  
Samir Thoria  
Cisco

Wim Henderickx  
Jorge Rabadan  
Alcatel-Lucent

Yakov Rekhter  
John Drake  
Juniper

Florin Balus  
Nuage Networks

Lucy Yong  
Linda Dunbar  
Huawei

Dennis Cai  
Cisco

Expires: January 4, 2015

July 4, 2014

Integrated Routing and Bridging in EVPN  
draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-04

Abstract

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

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at

<http://www.ietf.org/lid-abstracts.html>

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

#### Copyright and License Notice

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

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

#### Table of Contents

1	Introduction . . . . .	4
2	Inter-Subnet Forwarding Scenarios . . . . .	5
2.1	Switching among Subnets within a DC . . . . .	6
2.2	Switching among EVIs in different DCs without route aggregation . . . . .	7
2.3	Switching among EVIs in different DCs with route aggregation . . . . .	7
2.4	Switching among IP-VPN sites and EVIs with route aggregation . . . . .	7
3	Default L3 Gateway Addressing . . . . .	8
3.1	Homogeneous Environment . . . . .	8
3.1	Heterogeneous Environment . . . . .	9
4	Operational Models for Asymmetric Inter-Subnet Forwarding . . . . .	9
4.1	Among EVPN NVEs within a DC . . . . .	9
4.2	Among EVPN NVEs in Different DCs Without Route Aggregation . . . . .	10
4.3	Among EVPN NVEs in Different DCs with Route Aggregation . . . . .	12
4.4	Among IP-VPN Sites and EVPN NVEs with Route Aggregation . . . . .	13
4.5	Use of Centralized Gateway . . . . .	14
5	Operational Models for Symmetric Inter-Subnet Forwarding . . . . .	14
5.1	IRB forwarding on NVEs without core-facing IRB Interface . . . . .	14
5.1.1	Control Plane Operation for IRB forwarding without core-facing I/F . . . . .	15
5.1.2	Data Plane Operation for IRB forwarding without	

core-facing I/F . . . . .	16
5.2 IRB forwarding on NVEs with core-facing IRB Interface . . .	17
5.2.1 Control Plane Operation for IRB forwarding with core-facing I/F . . . . .	18
5.2.2 Data Plane Operation for IRB forwarding with core-facing I/F . . . . .	19
6 BGP Encoding . . . . .	20
7 VM Mobility . . . . .	21
7.1 VM Mobility & Optimum Forwarding for VM's Outbound Traffic .	21
7.2 VM Mobility & Optimum Forwarding for VM's Inbound Traffic .	21
7.2.1 Mobility without Route Aggregation . . . . .	22
7.2.2 Mobility with Route Aggregation . . . . .	22
8 Acknowledgements . . . . .	22
9 Security Considerations . . . . .	22
10 IANA Considerations . . . . .	22
11 References . . . . .	22
11.1 Normative References . . . . .	22
11.2 Informative References . . . . .	22
Authors' Addresses . . . . .	23

## Terminology

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

IRB: Integrated Routing and Bridging

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

NVE: Network Virtualization Endpoint

TS: Tenant System



## 1 Introduction

EVPN provides an extensible and flexible multi-homing VPN solution for intra-subnet connectivity among hosts/VMs over an MPLS/IP network. However, there are scenarios where, in addition to intra-subnet forwarding, inter-subnet forwarding is required among hosts/VMs across different IP subnets at the EVPN PE nodes, also known as EVPN NVE nodes throughout this document, while maintaining the multi-homing capabilities of EVPN. This document describes an Integrated Routing and Bridging (IRB) solution based on EVPN to address such requirements.

The inter-subnet communication is traditionally achieved at centralized L3 Gateway nodes where all the inter-subnet communication policies are enforced. When two Tenant Systems belonging to two different subnets connected to the same PE node wanted to talk to each other, their traffic needed to be back hauled from the PE node all the way to the centralized gateway nodes where inter-subnet switching is performed and then back to the PE node. For today's large multi-tenant data center, this scheme is very inefficient and sometimes impractical.

In order to overcome the drawback of centralized approach, IRB functionality is needed on the PE nodes (i.e., NVE devices) as close to TS as possible to avoid hair pinning of user traffic unnecessarily. Under this design, all traffic between hosts attached to one NVE can be routed and bridged locally, thus avoiding traffic hair-pinning issue at the centralized L3GW.

There can be scenarios where both centralized and decentralized approaches may be preferred simultaneously. For example, to allow NVEs to switch inter-subnet traffic belonging to one tenant or one security zone locally; whereas, to back haul inter-subnet traffic belonging to two different tenants or security zones to the centralized gateway nodes and perform switching there after the traffic is subjected to Firewall or Deep Packet Inspection (DPI).

Some TSes run non-IP protocols in conjunction with their IP traffic. Therefore, it is important to handle both kinds of traffic optimally - e.g., to bridge non-IP traffic and to route IP traffic.

Therefore, the solution needs to meet the following requirements:

R1: The solution MUST allow for inter-subnet traffic to be locally switched at NVEs.

R2: The solution MUST allow for both inter-subnet and intra-subnet traffic belonging to the same tenant to be locally routed and bridged

respectively. The solution MUST provide IP routing for inter-subnet traffic and Ethernet Bridging for intra-subnet traffic.

R3: The solution MUST support bridging non-IP traffic.

R4: The solution MUST allow inter-subnet switching to be disabled on a per VLAN basis on NVEs where the traffic needs to be back hauled to another node (e.g., for performing FW or DPI functionality).

## 2 Inter-Subnet Forwarding Scenarios

The inter-subnet forwarding scenarios performed by an EVPN NVE can be divided into the following five categories. The last scenario, along with their corresponding solutions, are described in [EVPN-IPVPN-INTEROP]. The solutions for the first four scenarios are the focus of this document.

1. Switching among EVIs (subnets) within a DC
2. Switching among EVIs (subnets) in different DCs without route aggregation
3. Switching among EVIs (subnets) in different DCs with route aggregation
4. Switching among IP-VPN sites and EVPN instances with route aggregation
5. Switching among IP-VPN sites and EVPN instances without route aggregation

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

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

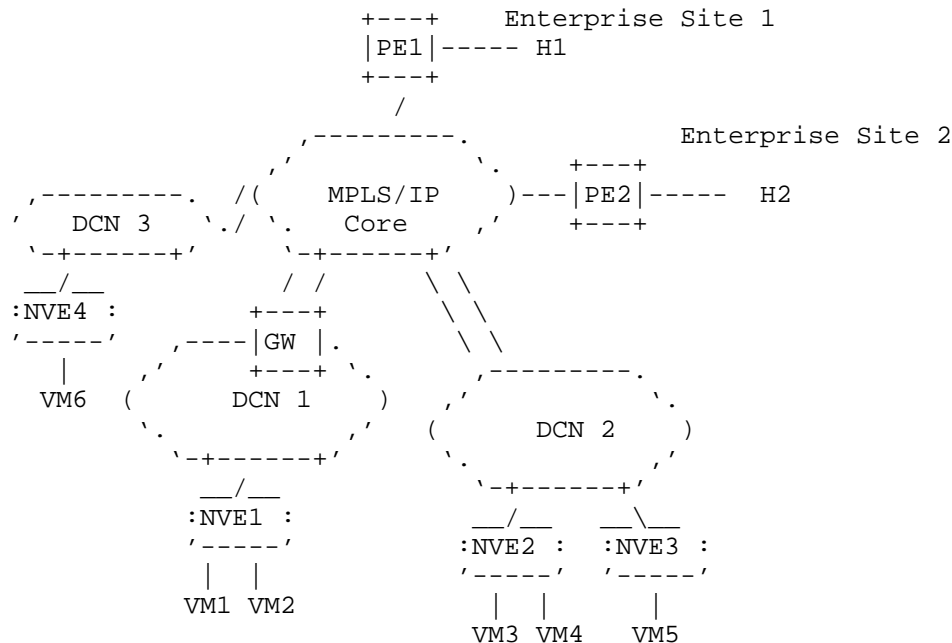


Figure 2: Interoperability Use-Cases

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

### 2.1 Switching among Subnets within a DC

In this scenario, connectivity is required between hosts (e.g. VMs) in the same data center, where those hosts belong to different IP subnets. All these subnets belong to the same tenant or are part of the same IP VPN. Each subnet is associated with a single EVPN instance, where each such EVI is realized by a collection of MAC-VRFs (one per NVE) residing on the NVEs configured for that EVI.

As an example, consider VM3 and VM5 of Figure 2 above. Assume that connectivity is required between these two VMs where VM3 belongs to the IP-subnet 3 (SN3) whereas VM5 belongs to the IP-subnet 5 (SN5). Both SN3 and SN5 subnets belong to the same tenant (e.g., are part of the same IP VPN). NVE2 has an EVI3 associated with the SN3 and this EVI is represented by a MAC-VRF which is connected to an IP-VRF (for that IP VPN) via an IRB interface. NVE3 respectively has an EVI5 associated with the SN5 and this EVI is represented by a MAC-VRF which is connected to an IP-VRF (for the same IP VPN) via an IRB interface.

## 2.2 Switching among EVIs in different DCs without route aggregation

This case is similar to that of section 2.1 above albeit for the fact that the hosts belong to different data centers that are interconnected over a WAN (e.g. MPLS/IP PSN). The data centers in question here are seamlessly interconnected to the WAN, i.e., the WAN edge devices does not maintain any host/VM-specific addresses in the forwarding path - e.g., there is no WAN edge GW(s) between these DCs.

As an example, consider VM3 and VM6 of Figure 2 above. Assume that connectivity is required between these two VMs where VM3 belongs to the SN3 whereas VM6 belongs to the SN6. NVE2 has an EVI3 associated with SN3 and NVE4 has an EVI6 associated with the SN6. Both SN3 and SN6 are part of the same IP VPN.

## 2.3 Switching among EVIs in different DCs with route aggregation

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

As an example, consider VM1 and VM5 of Figure 2 above. Assume that connectivity is required between these two VMs where VM1 belongs to the SN1 whereas VM5 belongs to the SN5 thus SN1 and SN5 belong to the same IP VPN. NVE3 has an EVI5 associated with the SN5 and this EVI is represented by the MAC-VRF which is connected to the IP-VRF via an IRB interface. NVE1 has an EVI1 associated with the SN1 and this EVI is represented by the MAC-VRF which is connected to the IP-VRF representing the same IP VPN. Due to the gateway at the edge of DCN 1, NVE1's IP-VRF does not need to have the address of VM5 but instead it has a default route in its IP-VRF with the next-hop being the GW.

## 2.4 Switching among IP-VPN sites and EVIs with route aggregation

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

As an example, consider end-station H1 and VM2 of Figure 2. Assume

that connectivity is required between the end-station and the VM, where VM2 belongs to the SN2 that is realized using EVPN, whereas H1 belongs to an IP VPN site connected to PE1 (PE1 maintains an IP-VRF associated with that IP VPN). NVE1 has an EVI2 associated with the SN2. Moreover, EVI2 on NVE1 is connected to an IP-VRF associated with that IP VPN. PE1 originates a VPN-IP route that covers H1. The gateway at the edge of DCN1 performs interworking function between IP-VPN and EVPN. As a result of this, a default route in the IP-VRF on the NVE1, pointing to the gateway as the next hop, and a route to the VM2 (or maybe SN2) on the PE1's IP-VRF are sufficient for the connectivity between H1 and VM2. In this scenario, the NVE1's IP-VRF does not need to maintain a route to H1 because it has the default route to the gateway.

### 3 Default L3 Gateway Addressing

#### 3.1 Homogeneous Environment

This is an environment where all NVEs to which an EVPN instance could potentially be attached (or moved), perform inter-subnet switching. Therefore, inter-subnet traffic can be locally switched by the EVPN NVE connecting the VMs belonging to different subnets.

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

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

Both of these models enable a packet forwarding paradigm for asymmetric IRB forwarding where a packet can bypass the VRF processing on the egress (i.e. disposition) NVE. The egress NVE merely needs to perform a lookup in the associated EVI and forward the Ethernet frames unmodified, i.e. without rewriting the source MAC address. This is different from symmetric IRB forwarding where a packet is forwarded through the bridge module followed by the routing module on the ingress NVE, and then forwarded through the routing

module followed by the bridging module on the egress NVE.

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

### 3.1 Heterogeneous Environment

For large data centers with thousands of servers and ToR (or Access) switches, some of them may not have the capability of maintaining or enforcing policies for inter-subnet switching. Even though policies among multiple subnets belonging to same tenant can be simpler, hosts belonging to one tenant can also send traffic to peers belonging to different tenants or security zones. A L3GW not only needs to enforce policies for communication among subnets belonging to a single tenant, but also it needs to know how to handle traffic destined towards peers in different tenants. Therefore, there can be a mixed environment where an NVE performs inter-subnet switching for some EVPN instances but not others.

## 4 Operational Models for Asymmetric Inter-Subnet Forwarding

### 4.1 Among EVPN NVEs within a DC

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

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated MAC-VRF for that EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup identifies both the next-hop (i.e. egress) NVE to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the EVPN MAC route), instead of the MAC address of the next-hop NVE. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) NVE after encapsulating it with the MPLS

label stack. Note that this label stack includes the LSP label as well as the EVI label that was advertised by the egress NVE. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVI label to identify the MAC-VRF table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 2 below depicts the packet flow, where NVE1 and NVE2 are the ingress and egress NVEs, respectively.

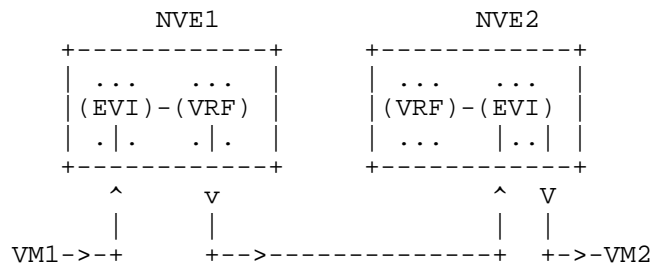


Figure 2: Inter-Subnet Forwarding Among EVPN NVEs within a DC

Note that the forwarding behavior on the egress NVE is similar to EVPN intra-subnet forwarding. In other words, all the packet processing associated with the inter-subnet forwarding semantics is confined to the ingress NVE and that is why it is called Asymmetric IRB.

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

#### 4.2 Among EVPN NVEs in Different DCs Without Route Aggregation

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

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress

NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup identifies both the next-hop (i.e. egress) Gateway to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the EVPN MAC route), instead of the MAC address of the next-hop Gateway. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) Gateway after encapsulating it with the MPLS label stack.

Note that this label stack includes the LSP label as well as an EVI label. The EVI label could be either advertised by the ingress Gateway, if inter-AS option B is used, or advertised by the egress NVE, if inter-AS option C is used. When the MPLS encapsulated packet is received by the ingress Gateway, the processing again differs depending on whether inter-AS option B or option C is employed: in the former case, the ingress Gateway swaps the EVI label in the packets with the EVI label value received from the egress Gateway. In the latter case, the ingress Gateway does not modify the EVI label and performs normal label switching on the LSP label. Similarly on the egress Gateway, for option B, the egress Gateway swaps the EVI label with the value advertised by the egress NVE. Whereas, for option C, the egress Gateway does not modify the EVI label, and performs normal label switching on the LSP label. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVI label to identify the bridge-domain table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 3 below depicts the packet flow.

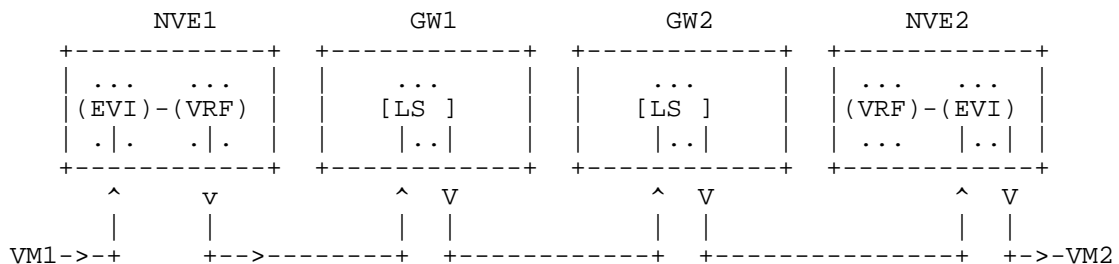


Figure 3: Inter-Subnet Forwarding Among EVPN NVEs in Different DCs without Route Aggregation



#### 4.3 Among EVPN NVEs in Different DCs with Route Aggregation

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

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated MAC-VRF table. If the MAC address corresponds to the IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup, in this case, matches the default route which points to the local WAN gateway. The ingress NVE then rewrites the destination MAC address in the packet with the IRB Interface MAC address of the local WAN gateway. It also rewrites the source MAC address with its own IRB Interface MAC address. The ingress NVE, then, forwards the frame to the WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the IP-VPN label that was advertised by the local WAN gateway. When the MPLS encapsulated packet is received by the local WAN gateway, it uses the IP-VPN label to identify the IP-VRF table. It then performs an IP lookup in that table. The lookup identifies both the remote WAN gateway (of the remote data center) to which the packet must be forwarded, in addition to an adjacency that contains a MAC rewrite and an MPLS label stack. The MAC rewrite holds the MAC address associated with the ultimate destination host (as populated by the EVPN MAC route). The local WAN gateway then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The local WAN gateway, then, forwards the frame to the remote WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as a EVI label that was advertised by the remote WAN gateway. When the MPLS encapsulated packet is received by the remote WAN gateway, it simply swaps the EVI label and forwards the packet to the egress NVE. This implies that the GW1 needs to keep the remote host MAC addresses along with the corresponding EVI labels in the adjacency entries of the IP-VRF table. The remote WAN gateway then forward the packet to the egress NVE. The egress NVE then performs a MAC lookup in the MAC-VRF (identified by the received EVI label) to determine the outbound port to send the traffic on.

Figure 4 below depicts the forwarding model.



where to forward the traffic.

Figure 5 below depicts the forwarding model.

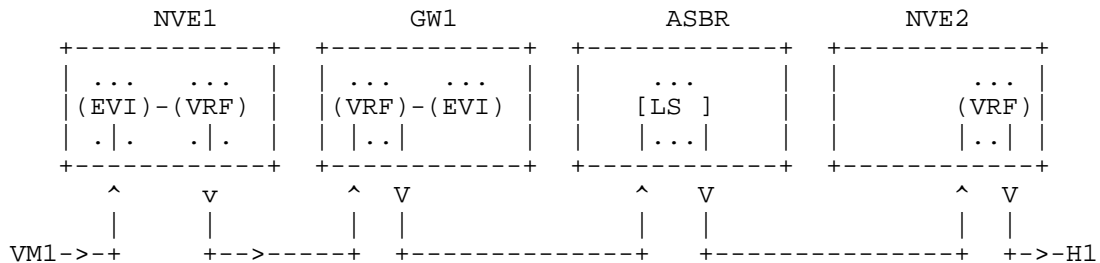


Figure 5: Inter-Subnet Forwarding Among IP-VPN Sites and EVPN NVEs with Route Aggregation

#### 4.5 Use of Centralized Gateway

In this scenario, the NVEs within a given data center need to forward traffic in L2 to a centralized L3GW for a number of reasons: a) they don't have IRB capabilities or b) they don't have required policy for switching traffic between different tenants or security zones. The centralized L3GW performs both the IRB function for switching traffic among different EVPN instances as well as it performs interworking function when the traffic needs to be switched between IP-VPN sites and EVPN instances.

### 5 Operational Models for Symmetric Inter-Subnet Forwarding

The following sections describe several main symmetric IRB forwarding scenarios.

#### 5.1 IRB forwarding on NVEs without core-facing IRB Interface

In this scenario, for a given tenant or IP-VPN, an NVE has an access-facing EVI for each tenant's subnet (VLAN) that is configured for. Assuming VLAN-based service which is typically the case for VxLAN and NVGRE encapsulation, each of these EVIs represent a MAC-VRF with one bridge domain. In case of MPLS encapsulation with VLAN-aware bundling, then each EVI may represent a MAC-VRF with multiple bridge domains (one bridge domain per VLAN). The EVIs (or MAC-VRFs) on an NVE for a given tenant are connected to an IP-VRF corresponding to that tenant (or IP-VPN) via their associated IRB interfaces.

Since in this scenario, there is no core-facing IRB interface, there

is no need for a core-facing EVI or MAC-VRF. The advantage of not having a core-facing IRB interface may be operational simplicity as there is no need to configure an IRB interface and have a MAC-VRF associated with it and no additional BGP MAC address advertisements are needed. However, the disadvantage for not having a core-facing IRB interface is that no QoS or security policies can be enforced for the core-facing traffic on a per tenant basis.

Since VxLAN and NVGRE encapsulations require inner Ethernet header (inner MAC SA/DA), and since for inter-subnet traffic, TS MAC address cannot be used, the ingress NVE's MAC address is used as inner MAC SA. It should be noted that if there was a core-facing IRB interface, then the MAC address of IRB interface would have been used as inner MAC SA. The NVE's MAC address is the device MAC address and the same MAC address is used across all EVIs and IP-VPNs.

Figure below illustrates this scenario where a given tenant (e.g., IP-VPN) has three subnets represented by EVI-1, EVI-2, and EVI3 across two NVEs. There are five TSEs connected to these three EVIs - i.e., TS1, TS5 are connected to EVI-1 on NVE1, TS4 is connected to EVI-1 on NVE2, TS2 is connected to EVI-2 on NVE1, and TS3 is connected to EVI3 on NVE2. When TS1, TS5, and TS4 exchange traffic with each other, only L2 forwarding (bridging) part of the IRB solution is used because all these TSEs sit on the same subnet. However, when TS1 wants to exchange traffic with TS2 or TS3 which belong to different subnets, then both bridging and routing parts of the IRB solution are used. The following subsections describe the control and data planes operations for this IRB scenario in details.

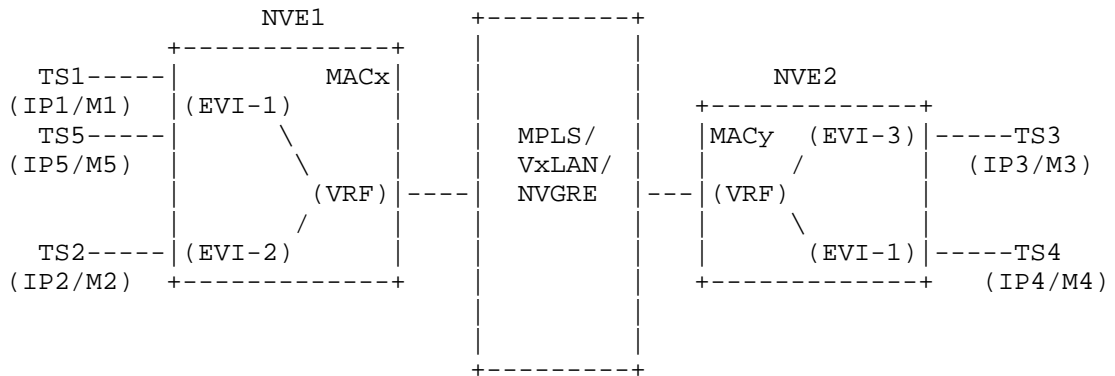


Figure 6: IRB forwarding on NVEs without core-facing IRB Interface

#### 5.1.1 Control Plane Operation for IRB forwarding without core-facing I/F

Each NVE advertises an RT-2 (MAC/IP Advertisement Route) for each of its TSeS with the following field set:

- RD and ESI per [EVPN]
- Ethernet Tag = 0; assuming VLAN-based service
- MAC Address Length = 48
- MAC Address =  $M_i$  ; where  $i = 1, 2, 3, 4$ , or 5 in the above example
- IP Address Length = 32 or 128
- IP Address =  $I P_i$  ; where  $i = 1, 2, 3, 4$ , or 5 in the above example
- Label-1 = MPLS Label or VNID corresponding to EVI
- Label-2 = MPLS Label or VNID corresponding to IP-VRF

Each RT-2 route is advertised with two RTs (one corresponding to the EVI and the other corresponding to the IP-VPN) and with a new BGP attribute (section 6) that includes the tunnel type and the MAC address of the NVE (e.g.,  $MAC_x$  for NVE1 or  $MAC_y$  for NVE2) .

Upon receiving this advertisement, the receiving NVE performs the following:

- It uses Route Targets corresponding to EVI and IP-VPN for importing this route into the corresponding MAC-VRF and IP-VRF tables.
- It imports the MAC address into the MAC-VRF with BGP Next Hop address as underlay tunnel destination address (e.g., VTEP DA for VxLAN encapsulation) and Label-1 as EVI VNID for VxLAN encapsulation or EVPN label for MPLS encapsulation.
- It imports the IP address into IP-VRF with NVE's MAC address (from the new BGP attribute) as inner MAC DA and BGP Next Hop address as underlay tunnel destination address (e.g., VTEP DA for VxLAN encapsulation) and Label-2 as IP-VPN VNID for VxLAN encapsulation or IP-VPN label for MPLS encapsulation.

#### 5.1.2 Data Plane Operation for IRB forwarding without core-facing I/F

The following description of the data-plane operation describes just the logical functions and the actual implementation may differ. Lets consider data-plane operation when TS1 in subnet-1 (EVI-1) on NVE1 wants to send traffic to TS3 in subnet-3 (EVI-3) on NVE2.

- TS1 send an Ethernet frame with MAC DA corresponding to the EVI-1 IRB interface of NVE1, and VLAN-tag corresponding to EVI-1.
- Upon receiving the Ethernet frame, the NVE1 uses VLAN-tag to identify the MAC-VRF corresponding to EVI-1. It then looks up the MAC DA and forwards the frame to its IRB interface.

- The Ethernet header of the frame is stripped and the packet is fed to the IP-VRF where IP lookup is performed on the destination address. This lookup yields a MAC address to be used as inner MAC DA for VxLAN/NVGRE encapsulation, an IP address to be used as VTEP DA for VxLAN encap or tunnel label for MPLS encap , and a VPN-ID to be used as VNID for VxLAN encap or IP-VPN label.
- The packet is then encapsulated with the proper header based on the above info. The inner MAC SA and VTEP SA is set to NVE's MAC and IP addresses respectively. The packet is then forwarded to the egress NVE.
- On the egress NVE, if the packet is VxLAN encapsulated, the VxLAN header is removed. Since the inner MAC DA is that of egress NVE, the NVE knows that it needs to perform an IP lookup. It uses VNID to identify the IP-VRF table and then performs an IP lookup which results in destination TS (TS3) MAC address and the access-facing IRB interface over which the packet needs to be sent.
- The IP packet is encapsulated with an Ethernet header with MAC SA set to that of NVE-2 MAC address(MACy) and MAC DA set to that of destination TS (TS3) MAC address. The packet is sent to the corresponding MAC-VRF and after a lookup of MAC DA, is forwarded to the destination TS (TS3) over the corresponding interface.

## 5.2 IRB forwarding on NVEs with core-facing IRB Interface

The only difference between this scenario and the previous scenario is that there is a core-facing IRB interface per tenant (or IP-VPN) on each NVE. Each core-facing IRB interface has a MAC and IP addresses associated with it and it allows for QoS/security policies to be configured on a per tenant basis on this interface. Furthermore, it allows for better OAM coverage (e.g., fault isolation) by running OAM on this interface. Other than that, the rest of the functionality is the same as the solution describe in section 5.1.

This core-facing IRB interface results in additional control-plane processing (e.g., BGP routes advertisements) and additional data-plane processing as detail in the next two sub-sections.

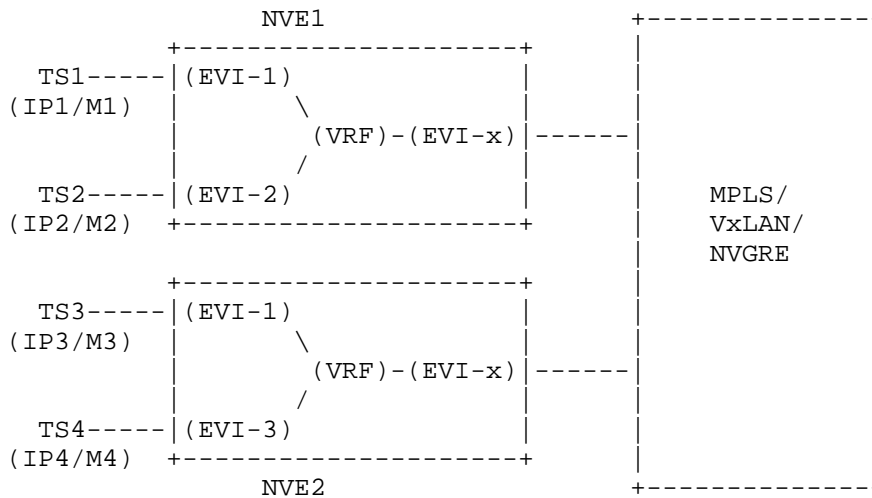


Figure 7: IRB forwarding on NVEs with core-facing IRB Interface

## 5.2.1 Control Plane Operation for IRB forwarding with core-facing I/F

Each NVE advertises an RT-2 (MAC/IP Advertisement Route) for each of its TSeS and it also advertises a single RT-2 for core-facing IRB interface (which is per tenant or per IP-VPN).

The fields of RT-2 for each TS are set as follow:

- RD and ESI per [EVPN]
- Ethernet Tag = 0; assuming VLAN-based service
- MAC Address Length = 48
- MAC Address = Mi ; MAC address of TS
- IP Address Length = 32 or 128
- IP Address = Ipi ; IP address of TS
- Label-1 = MPLS Label or VNID corresponding to access-facing EVI

Furthermore, this RT-2 is also advertised with two RTs (one corresponding to the EVI and the other corresponding to the IP-VPN) as described in section 5.1.1. The main difference in terms of BGP advertisement for this per-TS RT-2 is that it is advertised with a new BGP attribute (section 6) that includes the tunnel type and the IP address of the core-facing IRB interface (which is per tenant).

Upon receiving this per-TS RT-2 advertisement, the receiving NVE performs the following:

- It uses the Route Targets corresponding to EVI and IP-VPN for

importing this route into the corresponding MAC-VRF and IP-VRF tables similar to section 5.1.1.

- It imports the MAC address into the MAC-VRF just like section 5.1.1.
- It imports the IP address into IP-VRF with next hop pointing to the IP address of core-facing IRB interface (carried in the new BGP attribute).

The fields of RT-2 advertised for core-facing IRB interface, are set as follow. This RT-2 is advertised with an RT corresponding to the core-facing EVI (e.g., EVI-x). This RT-2 is also advertised as a sticky MAC per section 15.2 of [EVPN] in order to ensure mis-configuration is caught quickly.

- RD per [EVPN]
- ESI = 0
- Ethernet Tag = 0
- MAC Address Length = 48
- MAC Address = Ma ; MAC address of core-facing IRB interface
- IP Address Length = 32 or 128
- IP Address = IPa ; IP address of core-facing IRB interface
- Label-1 = MPLS Label or VNID corresponding to core-facing EVI

Upon receiving the RT-2 advertisement corresponding to core-facing IRB interface, the receiving NVE performs the following:

- It uses the Route Target corresponding to the EVI-x, to identify MAC-VRF associated with EVI-x.
- It imports the MAC address into the MAC-VRF associated with EVI-x with BGP Next Hop address as underlay tunnel destination address (e.g., VTEP DA for VxLAN encapsulation) and Label-1 as EVI VNID for VxLAN encapsulation or EVPN label for MPLS encapsulation.
- It imports (MAC/IP ) pair associated with core-facing IRB interface into the overlay ARP table. This overlay ARP table is used to resolve per-TS IP addresses imported into the IP-VRF table previously.

#### 5.2.2 Data Plane Operation for IRB forwarding with core-facing I/F

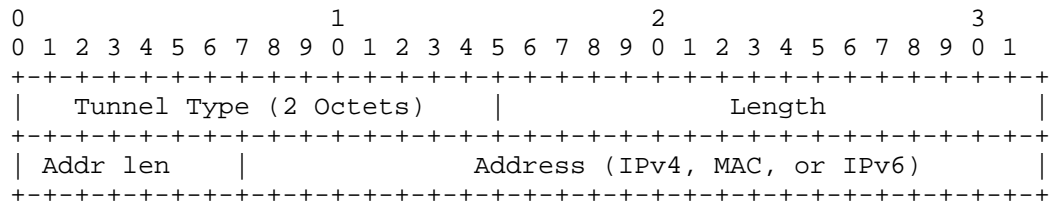
The following description of the data-plane operation describes just the logical functions and the actual implementation may differ. Lets consider data-plane operation when TS1 in subnet-1 (EVI-1) on NVE1 wants to send traffic to TS4 in subnet-3 (EVI-3) on NVE2.



- TS1 send an Ethernet frame with MAC DA corresponding to the EVI-1 IRB interface of NVE1, and VLAN-tag corresponding to EVI-1 just like section 5.1.1.
- Upon receiving the Ethernet frame, the ingress NVE1 uses VLAN-tag to identify the MAC-VRF corresponding to EVI-1. It then looks up the MAC DA and forwards the frame to its IRB interface just like section 5.1.1.
- The Ethernet header of the frame is stripped and the packet is fed to the IP-VRF where IP lookup is performed on the destination address. This lookup yield a MAC address (corresponding to the destination core-facing IRB interface) and its local core-facing IRB interface over which the packet is sent.
- The packet is encapsulated with an Ethernet header where MAC SA is set to that of the local core-facing IRB interface and MAC DA is set to that of the remote core-facing IRB interface. The packet is then sent to the core-facing EVI of the ingress NVE.
- MAC DA lookup is performed in the core-facing IRB of the ingress NVE. This lookup yields an IP address to be used as VTEP DA for VxLAN encap or tunnel label for MPLS encap , and a VPN-ID to be used as VNID for VxLAN encap or IP-VPN label.
- The packet is then encapsulated with the proper header based on the above info and is forwarded to the egress NVE.
- On the egress NVE, if the packet is VxLAN encapsulated, the VxLAN header is removed and the resultant Ethernet frame is fed into the core-facing MAC-VRF associated with that tenant based on the VNID.
- The MAC DA lookup yields the core-facing IRB interface of the egress NVE over which the frame is sent. Next, the Ethernet header is removed and a lookup is performed based on IP DA in the associated IP-VRF for that tenant. The IP lookup yields the destination TS (TS3) MAC address and the access-facing IRB interface over which the packet needs to be sent.
- The IP packet is encapsulated with an Ethernet header with the MAC SA set to that of the access-facing IRB interface of the egress NVE (NVE2) and the MAC DA is set to that of destination TS (TS4) MAC address. The packet is sent to the corresponding MAC-VRF and after a lookup of MAC DA, is forwarded to the destination TS (TS3) over the corresponding interface.

## 6 BGP Encoding

A new BGP attribute with the following encoding is introduced.



Tunnel Type (2 octets): identifies the type of tunneling technology being signaled. This document specifies the following types:

This document defines the following types:

- VXLAN: Tunnel Type = 8
- NVGRE: Tunnel Type = 9
- GTP: Tunnel Type = 10

Unknown types MUST be ignored and skipped upon receipt.

Length (2 octets): the total number of octets of the value field.

Address Length - Addr len (1 octet): Length of Address. Set to 4 bytes for an IPv4 address, 6 bytes for MAC address, and 16 bytes for an IPv6 address.

## 7 VM Mobility

### 7.1 VM Mobility & Optimum Forwarding for VM's Outbound Traffic

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

### 7.2 VM Mobility & Optimum Forwarding for VM's Inbound Traffic

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

In what follows, we use the following terminology:

- source NVE refers to the NVE behind which the VM used to reside prior to the VM mobility event.

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

#### 7.2.1 Mobility without Route Aggregation

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

#### 7.2.2 Mobility with Route Aggregation

This section will be completed in the next revision.

### 8 Acknowledgements

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

### 9 Security Considerations

### 10 IANA Considerations

### 11 References

#### 11.1 Normative References

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

#### 11.2 Informative References

- [EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-12vpn-evpn-04.txt, work in progress, July, 2014.

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

[DC-MOBILITY] Aggarwal et al., "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-05.txt, work in progress, June, 2013.

#### Authors' Addresses

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

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

Yakov Rekhter  
Juniper Networks  
Email: [yakov@juniper.net](mailto:yakov@juniper.net)

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

Lucy Yong  
Huawei Technologies  
Email: [lucy.yong@huawei.com](mailto:lucy.yong@huawei.com)

Linda Dunbar  
Huawei Technologies  
Email: [linda.dunbar@huawei.com](mailto:linda.dunbar@huawei.com)

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

Florin Balus  
Alcatel-Lucent

Email: Florin.Balus@alcatel-lucent.com

Samir Thoria

Cisco

Email: sthoria@cisco.com

L2VPN Workgroup  
INTERNET-DRAFT  
Intended Status: Standards Track

A. Sajassi (Editor)  
Cisco

J. Drake (Editor)  
Juniper

Y. Rekhter  
R. Shekhar  
B. Schliesser  
Juniper

Nabil Bitar  
Verizon

S. Salam  
K. Patel  
D. Rao  
S. Thoria  
Cisco

Aldrin Isaac  
Bloomberg

James Uttaro  
AT&T

L. Yong  
Huawei

W. Henderickx  
Alcatel-Lucent

D. Cai  
S. Sinha  
Cisco

Expires: December 18, 2014

June 18, 2014

A Network Virtualization Overlay Solution using EVPN  
draft-sd-l2vpn-evpn-overlay-03

Abstract

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

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months

and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

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

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

## Copyright and License Notice

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

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

## Table of Contents

1	Introduction . . . . .	4
2	Specification of Requirements . . . . .	5
3	Terminology . . . . .	5
4	EVPN Features . . . . .	6
5	Encapsulation Options for EVPN Overlays . . . . .	7
5.1	VXLAN/NVGRE Encapsulation . . . . .	7
5.1.1	Virtual Identifiers Scope . . . . .	8
5.1.1.1	Data Center Interconnect with Gateway . . . . .	8
5.1.1.2	Data Center Interconnect without Gateway . . . . .	9
5.1.2	Virtual Identifiers to EVI Mapping . . . . .	9
5.1.2.1	Auto Derivation of RT . . . . .	10
5.1.3	Constructing EVPN BGP Routes . . . . .	11
5.2	MPLS over GRE . . . . .	12
6	EVPN with Multiple Data Plane Encapsulations . . . . .	13
7	NVE Residing in Hypervisor . . . . .	13
7.1	Impact on EVPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation . . . . .	14
7.2	Impact on EVPN Procedures for VXLAN/NVGRE Encapsulation . . . . .	14

8	NVE Residing in ToR Switch . . . . .	15
8.1	EVPN Multi-Homing Features . . . . .	15
8.1.1	Multi-homed Ethernet Segment Auto-Discovery . . . . .	16
8.1.2	Fast Convergence and Mass Withdraw . . . . .	16
8.1.3	Split-Horizon . . . . .	16
8.1.4	Aliasing and Backup-Path . . . . .	16
8.1.5	DF Election . . . . .	17
8.2	Impact on EVPN BGP Routes & Attributes . . . . .	17
8.3	Impact on EVPN Procedures . . . . .	17
8.3.1	Split Horizon . . . . .	18
8.3.2	Aliasing and Backup-Path . . . . .	19
9	Support for Multicast . . . . .	19
10	Inter-AS . . . . .	20
11	Acknowledgement . . . . .	21
12	Security Considerations . . . . .	21
13	IANA Considerations . . . . .	22
14	References . . . . .	22
14.1	Normative References . . . . .	22
14.2	Informative References . . . . .	22
	Authors' Addresses . . . . .	23



## 1 Introduction

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

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

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

This document describes how Ethernet VPN (EVPN) can be used as an NVO solution and explores applicability of EVPN functions and procedures.

In particular, it describes the various tunnel encapsulation options for EVPN over IP, and their impact on the EVPN control-plane and procedures for two main scenarios:

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

Note that the use of EVPN as an NVO solution does not necessarily mandate that the BGP control-plane be running on the NVE. For such scenarios, it is still possible to leverage the EVPN 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 EVPN overlays that are analyzed in this document are:

- VXLAN and NVGRE
- MPLS over GRE

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

and any impact that the encapsulation has on those features.

## 2 Specification of Requirements

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

## 3 Terminology

NVO: Network Virtualization Overlay

NVE: Network Virtualization Endpoint

VNI: Virtual Network Identifier (for VxLAN)

VSID: Virtual Subnet Identifier (for NVGRE)

EVPN: Ethernet VPN

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

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

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

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

Single-Active Multihoming: When a device or a network is multihomed to a group of two or more PEs and when only a single PE in such a redundancy group can forward traffic to/from the multihomed device or network for a given VLAN, such multihoming is referred to as "Single-Active"

All-Active Multihoming: When a device is multihomed to a group of two

or more PEs and when all PEs in such redundancy group can forward traffic to/from the multihomed device or network for a given VLAN, such multihoming is referred to as "All-Active".

#### 4 EVPN Features

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

- 1) Control plane traffic is distributed with BGP and Broadcast and Multicast traffic is sent using a shared multicast tree or with ingress replication.
- 2) Control plane learning is used for MAC (and IP) addresses instead of data plane learning. The latter requires the flooding of unknown unicast and ARP frames; whereas, the former does not require any flooding.
- 3) Route Reflector is used to reduce a full mesh of BGP sessions among PE devices to a single BGP session between a PE and the RR. Furthermore, RR hierarchy can be leveraged to scale the number BGP routes on the RR.
- 4) Auto-discovery via BGP is used to discover PE devices participating in a given VPN, PE devices participating in a given redundancy group, tunnel encapsulation types, multicast tunnel type, multicast members, etc.
- 5) All-Active multihoming is used. This allows a given customer device (CE) to have multiple links to multiple PEs, and traffic to/from that CE fully utilizes all of these links. This set of links is termed an Ethernet Segment (ES).
- 6) When a link between a CE and a PE fails, the PEs for that EVI are notified of the failure via the withdrawal of a single EVPN route. This allows those PEs to remove the withdrawing PE as a next hop for every MAC address associated with the failed link. This is termed 'mass withdrawal'.
- 7) BGP route filtering and constrained route distribution are leveraged to ensure that the control plane traffic for a given EVI is only distributed to the PEs in that EVI.
- 8) When a 802.1Q interface is used between a CE and a PE, each of the VLAN ID (VID) on that interface can be mapped onto a bridge domain (for upto 4094 such bridge domains). All these bridge domains can also be mapped onto a single EVI (in case of VLAN-aware bundle

service).

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

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

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

## 5 Encapsulation Options for EVPN Overlays

### 5.1 VXLAN/NVGRE Encapsulation

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

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

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

[VXLAN] also provides an option of including an inner VLAN tag in the encapsulated frame, if explicitly configured at the VTEP. This mode of operation can either map to VLAN Based Service or VLAN Bundle Service in [EVPN] because inner VLAN tag is not used for lookup by the disposition PE when performing VXLAN decapsulation as described in section 6 of [VXLAN].

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

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

#### 5.1.1 Virtual Identifiers Scope

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

##### 5.1.1.1 Data Center Interconnect with Gateway

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

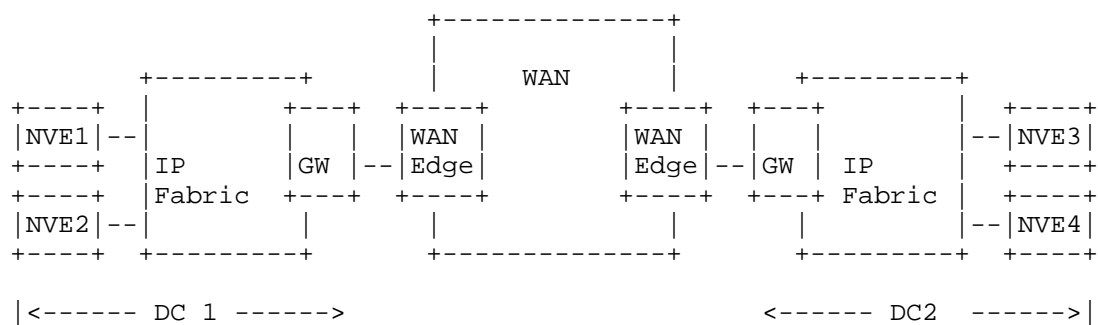


Figure 1: Data Center Interconnect with Gateway

## 5.1.1.2 Data Center Interconnect without Gateway

In the case where NVEs in different data centers need to be interconnected, and the NVEs need to use locally assigned VNIs or VSIDs (e.g., as MPLS labels), then there may be no need to employ Gateways at the edge of the data center network. More specifically, the VNI or VSID value that is used by the transmitting NVE is allocated by the NVE that is receiving the traffic (in other words, this is a "downstream assigned" MPLS label). This allows the VNI or VSID space to be decoupled between different data center networks without the need for a dedicated Gateway at the edge of the data centers.

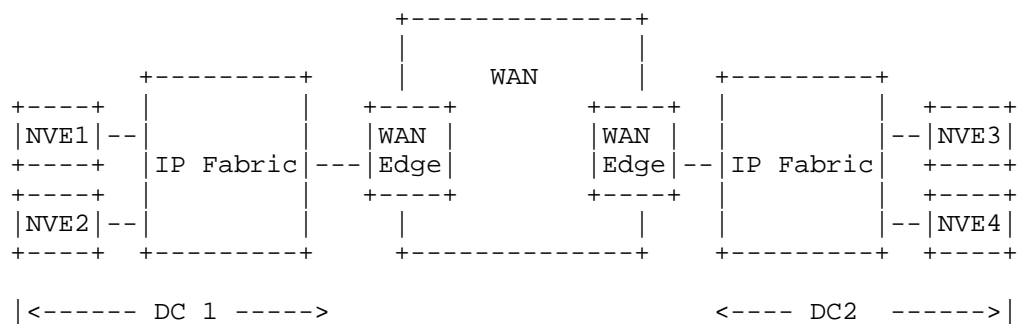


Figure 2: Data Center Interconnect without Gateway

## 5.1.2 Virtual Identifiers to EVI Mapping

When the EVPN control plane is used in conjunction with VXLAN or NVGRE, two options for mapping the VXLAN VNI or NVGRE VSID to an EVI are possible:

## 1. Option 1: Single Subnet per EVI

In this option, a single subnet represented by a VNI or VSID is mapped to a unique EVI. As such, a BGP RD and RT is needed per VNI / VSID on every VTEP. The advantage of this model is that it allows the BGP RT constraint mechanisms to be used in order to limit the propagation and import of routes to only the VTEPs that are interested in a given VNI or VSID. The disadvantage of this model may be the provisioning overhead if RD and RT are not derived automatically from VNI or VSID.

In this option, the MAC-VRF table is identified by the RT in the control plane and by the VNI or VSID for the data-plane. In this

option, the specific the MAC-VRF table corresponds to only a single bridge domain (e.g., a single subnet).

## 2. Option 2: Multiple Subnets per EVI

In this option, multiple subnets each represented by a unique VNI or VSID are mapped to a unique EVI. For example, if a tenant has multiple segments/subnets each represented by a VNI or VSID, then all the VNIs (or VSIDs) for that tenant are mapped to a single EVI - e.g., the EVI in this case represents the tenant and not a subnet . The advantage of this model is that it doesn't require the provisioning of RD/RT per VNI or VSID. However, this is a moot point if option 1 with if auto-derivation is used. The disadvantage of this model is that routes would be imported by VTEPs that may not be interested in a given VNI or VSID.

In this option the MAC-VRF table is identified by the RT in the control plane and a specific bridge domain for that MAC-VRF is identified by the <RT, Ethernet Tag ID> in the control plane. In this option, the VNI/VSID in the data-plane is sufficient to identify a specific bridge domain - e.g., no need to do a lookup based on VNI/VSID field and Ethernet Tag ID fields to identify a bridge domain.

#### 5.1.2.1 Auto Derivation of RT

When the option of a single VNI or VSID per EVI is used, it is important to auto-derive RT for EVPN BGP routes in order to simplify configuration for data center operations. RD can be derived easily as described in [EVPN] and RT can be auto-derived as described next.

Since a gateway PE as depicted in figure-1 participates in both the DCN and WAN BGP sessions, it is important that when RT values are auto-derived for VNIs (or VSIDs), there is no conflict in RT spaces between DCN and WAN networks assuming that both are operating within the same AS. Also, there can be scenarios where both VXLAN and NVGRE encapsulations may be needed within the same DCN and their corresponding VNIs and VSIDs are administered independently which means VNI and VSID spaces can overlap. In order to ensure that no such conflict in RT spaces arises, RT values for DCNs are auto-derived as follow:

[illegible]

+++++

- 2 bytes of global admin field of the RT is set to the AS number.
- Three least significant bytes of the local admin field of the RT is set to the VNI or VSID, I-SID, or VID. The most significant bit of the local admin field of the RT is set as follow:
  - 0: auto-derived
  - 1: manually-derived
- The next 3 bits of the most significant byte of the local admin field of the RT identifies the space in which the other 3 bytes are defined. The following spaces are defined:
  - 0 : VID
  - 1 : VXLAN
  - 2 : NVGRE
  - 3 : I-SID
  - 4 : EVI
  - 5 : dual-VID
- The remaining 4 bits of the most significant byte of the local admin field of the RT identifies the domain-id. The default value of domain-id is zero indicating that only a single numbering space exist for a given technology. However, if there are more than one number space exist for a given technology (e.g., overlapping VXLAN spaces), then each of the number spaces need to be identify by their corresponding domain-id starting from 1.

### 5.1.3 Constructing EVPN BGP Routes

In EVPN, an MPLS label is distributed by the egress PE via the EVPN control plane and is placed in the MPLS header of a given packet by the ingress PE. This label is used upon receipt of that packet by the egress PE for disposition of that packet. This is very similar to the use of the VNI or VSID by the egress VTEP or NVE, respectively, with the difference being that an MPLS label has local significance while a VNI or VSID typically has global significance. Accordingly, and specifically to support the option of locally assigned VNIs, the MPLS label field in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast Ethernet Tag routes is used to carry the VNI or VSID. For the balance of this memo, the MPLS label field will be referred to as the VNI/VSID field. The VNI/VSID field is used for both locally and globally assigned VNIs or VSIDs.

For the VNI based mode (a single VNI per EVI), the Ethernet Tag field



in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast route MUST be set to zero just as in the VLAN Based service in [EVPN]. For the VNI bundle mode (multiple VNIs per EVI with a single bridge domain), the Ethernet Tag field in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast Ethernet Tag routes MUST be set to zero just as in the VLAN Bundle service in [EVPN].

For the VNI-aware bundle mode (multiple VNIs per EVI each with its own bridge domain), the Ethernet Tag field in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast route MUST identify a bridge domain within an EVI and the set of Ethernet Tags for that EVI needs to be configured consistently on all PEs within that EVI. The value advertised in the Ethernet Tag field MAY be a VNI as long as it matches the existing semantics of the Ethernet Tag, i.e., it identifies a bridge domain within an EVI and the set of VNIs are configured consistently on each PE in that EVI.

In order to indicate that which type of data plane encapsulation (i.e., VXLAN, NVGRE, MPLS, or MPLS in GRE) is to be used, the BGP Encapsulation extended community defined in [RFC5512] is included with all EVPN routes (i.e. MAC Advertisement, Ethernet AD per EVI, Ethernet AD per ESI, Inclusive Multicast Ethernet Tag, and Ethernet Segment) advertised by an egress PE. Four new values will be defined to extend the list of encapsulation types defined in [RFC5512]:

- + TBD (IANA assigned) - VXLAN Encapsulation
- + TBD (IANA assigned) - NVGRE Encapsulation
- + TBD (IANA assigned) - MPLS Encapsulation
- + TBD (IANA assigned) - MPLS in GRE Encapsulation

If the BGP Encapsulation extended community is not present, then the default MPLS encapsulation or a statically configured encapsulation is assumed.

The Next Hop field of the MP\_REACH\_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the NVE. The remaining fields in each route are set as per [EVPN].

## 5.2 MPLS over GRE

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

operation.

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

## 6 EVPN with Multiple Data Plane Encapsulations

The use of the BGP Encapsulation extended community allows each PE in a given EVI to know each of the encapsulations supported by each of the other PEs in that EVI. I.e., each of the PEs in a given EVI may support multiple data plane encapsulations. An ingress PE can send a frame to an egress PE only if the set of encapsulations advertised by the egress PE in the subject MAC Advertisement or Per EVI Ethernet AD route, forms a non-empty intersection with the set of encapsulations supported by the ingress PE, and it is at the discretion of the ingress PE which encapsulation to choose from this intersection. (As noted in section 5.1.3, if the BGP Encapsulation extended community is not present, then the default MPLS encapsulation or a statically configured encapsulation is assumed.)

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

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

It is the responsibility of the operator of a given EVI to ensure that all of the PEs in that EVI support at least one common encapsulation. If this condition is violated, it could result in service disruption or failure. The use of the BGP Encapsulation extended community provides a method to detect when this condition is violated but the actions to be taken are at the discretion of the operator and are outside the scope of this document.

## 7 NVE Residing in Hypervisor

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

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

#### 7.1 Impact on EVPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation

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

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

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

However, as noted in section 8.6 of [EVPN] in order to enable a single-homed ingress PE to take advantage of fast convergence, aliasing, and backup-path when interacting with multi-homed egress PEs attached to a given Ethernet segment, a single-homed ingress PE SHOULD be able to receive and process Ethernet AD per ES and Ethernet AD per EVI routes."

#### 7.2 Impact on EVPN Procedures for VXLAN/NVGRE Encapsulation

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

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

However, as noted in section 8.6 of [EVPN] in order to enable a single-homed ingress PE to take advantage of fast convergence, aliasing, and back-up path when interacting with multi-homed egress PEs attached to a given Ethernet segment, a single-homed ingress PE SHOULD implement the ingress node processing of Ethernet AD per ES and Ethernet AD per EVI routes as defined in sections 8.2 Fast Convergence and 8.4 Aliasing and Backup-Path of [EVPN].

## 8 NVE Residing in ToR Switch

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

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

### 8.1 EVPN Multi-Homing Features

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

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

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

#### 8.1.2 Fast Convergence and Mass Withdraw

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

#### 8.1.3 Split-Horizon

If a CE that is multi-homed to two or more NVEs on an Ethernet segment ES1 operating in all-active redundancy mode sends a multicast, broadcast or unknown unicast packet to a one of these NVEs, then that NVE will forward that packet to all of the other PEs in that EVI including the other NVEs attached to ES1 and those NVEs MUST drop the packet and not forward back to the originating CE. This is termed 'split horizon filtering'.

#### 8.1.4 Aliasing and Backup-Path

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

To alleviate this issue, EVPN introduces the concept of Aliasing. This refers to the ability of an NVE to signal that it has reachability to a given locally attached Ethernet segment, even when it has learnt no MAC addresses from that segment. The Ethernet A-D route per EVI is used to that end. Remote NVEs which receive MAC advertisement routes with non-zero ESI SHOULD consider the MAC address as reachable via all NVEs that advertise reachability to the relevant Segment using Ethernet A-D routes with the same ESI and with the Single-Active flag reset.

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

#### 8.1.5 DF Election

If a CE is multi-homed to two or more NVEs on an Ethernet segment operating in all-active redundancy mode, then for a given EVI only one of these NVEs, termed the Designated Forwarder (DF) is responsible for sending it broadcast, multicast, and, if configured for that EVI, unknown unicast frames.

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.

#### 8.2 Impact on EVPN BGP Routes & Attributes

Since multi-homing is supported in this scenario, then the entire set of BGP routes and attributes defined in [EVPN] are used. As discussed in Section 3.1.3, the VSID or VNI is carried in the VNI/VSID field in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast Ethernet Tag routes.

#### 8.3 Impact on EVPN Procedures

Two cases need to be examined here, depending on whether the NVEs are

operating in Active/Standby or in All-Active redundancy.

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

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

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

### 8.3.1 Split Horizon

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

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

Every NVE track the IP address(es) associated with the other NVE(s) with which it has shared multi-homed Ethernet Segments. When the NVE

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

In order to prevent unhealthy interactions between the split horizon procedures defined in [EVPN] and the local bias procedures described in this document, a mix of MPLS over GRE encapsulations on the one hand and VXLAN/NVGRE encapsulations on the other on a given Ethernet Segment is prohibited.

### 8.3.2 Aliasing and Backup-Path

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

## 9 Support for Multicast

The E-VPN Inclusive Multicast BGP route is used to discover the multicast tunnels among the endpoints associated with a given VXLAN VNI or NVGRE VSID. The Ethernet Tag field of this route is used to encode the VNI for VLXAN or VSID for NVGRE. The Originating router's IP address field is set to the NVE's IP address. 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 tunnel encapsulation is encoded by adding the BGP Encapsulation extended community as per section 3.1.1. The following tunnel types as defined in [RFC6514] can be used in the PMSI tunnel attribute for VXLAN/NVGRE:

+ 3 - PIM-SSM Tree



- + 4 - PIM-SM Tree
- + 5 - BIDIR-PIM Tree
- + 6 - Ingress Replication

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

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

## 10 Inter-AS

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

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

In the first scenario, inter-AS operation is straight-forward and follows existing BGP inter-AS procedures. However, in the first scenario where the tunnel endpoint IP addresses are public, there may be security concern regarding the distribution of these addresses among different ASes. This security concern is one of the main reasons for having the so called inter-AS "option-B" in MPLS VPN solutions such as EVPN.

The second scenario is more challenging, because the absence of the MPLS client layer from the VXLAN encapsulation creates a situation where the ASBR has no fully qualified indication within the tunnel header as to where the tunnel endpoint resides. To elaborate on this, recall that with MPLS, the client layer labels (i.e. the VPN labels) are downstream assigned. As such, this label implicitly has a connotation of the tunnel endpoint, and it is sufficient for the ASBR to look up the client layer label in order to identify the label translation required as well as the tunnel endpoint to which a given packet is being destined. With the VXLAN encapsulation, the VNI is globally assigned and hence is shared among all endpoints. The destination IP address is the only field which identifies the tunnel endpoint in the tunnel header, and this address is privately managed by every data center network. Since the tunnel address is allocated

out of a private address pool, then we either need to do a lookup based on VTEP IP address in context of a VRF (e.g., use IP-VPN) or terminate the VXLAN tunnel and do a lookup based on the tenant's MAC address to identify the egress tunnel on the ASBR. This effectively mandates that the ASBR to either run another overlay solution such as IP-VPN over MPLS/IP core network or to be aware of the MAC addresses of all VMs in its local AS, at the very least.

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

## 11 Acknowledgement

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

## 12 Security Considerations

This document uses IP-based tunnel technologies to support data plane transport. Consequently, the security considerations of those tunnel technologies apply. This document defines support for [VXLAN] and [NVGRE]. The security considerations from those documents as well as [RFC4301] apply to the data plane aspects of this document.

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

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

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

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

the scope of this document.

### 13 IANA Considerations

### 14 References

#### 14.1 Normative References

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

#### 14.2 Informative References

- [EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (EVPN)", draft-ietf-l2vpn-evpn-req-01.txt, work in progress, October 21, 2012.
- [NVGRE] Sridhavan, M., et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-01.txt, July 8, 2012.
- [VXLAN] Dutt, D., et al, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-02.txt, August 22, 2012.
- [EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-02.txt, work in progress, February, 2012.
- [Problem-Statement] Narten et al., "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement-01, September 2012.
- [L3VPN-ENDSYSTEMS] Marques et al., "BGP-signaled End-system IP/VPNs", draft-ietf-l3vpn-end-system, work in progress, October 2012.

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

#### Authors' Addresses

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

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

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

Aldrin Isaac  
Bloomberg  
Email: [aisaac71@bloomberg.net](mailto:aisaac71@bloomberg.net)

James Uttaro  
AT&T  
Email: [uttaro@att.com](mailto:uttaro@att.com)

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

Ravi Shekhar  
Juniper Networks  
Email: [rshekhar@juniper.net](mailto:rshekhar@juniper.net)

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

Keyur Patel

Cisco  
Email: Keyupate@cisco.com

Dhananjaya Rao  
Cisco  
Email: dhrao@cisco.com

Samir Thoria  
Cisco  
Email: sthoria@cisco.com

L2VPN Working Group  
INTERNET-DRAFT  
Intended Status: Proposed Standard

R. Singh  
K. Kompella  
Juniper Networks  
S. Palislaovic  
Alcatel-Lucent  
June 17, 2014

Expires: December 19, 2014

Updated processing of control flags for BGP VPLS  
draft-singh-l2vpn-bgp-vpls-control-flags-01

## Abstract

This document updates the meaning of the "control flags" fields inside the "layer2 info extended community" used for BGP-VPLS NLRI.

## Status of this Memo

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

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

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

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

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

## Copyright and License Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents  
(<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Introduction . . . . .	3
1.1	Terminology . . . . .	3
2	Problem . . . . .	3
3	Updated meaning of control flags in the layer2 info extended community . . . . .	4
3.1	Control word (C-bit) . . . . .	4
3.2	Sequence flag (S-bit) . . . . .	4
4	Using p2mp LSP as transport for BGP VPLS . . . . .	5
5.	Treatment of C and S bits in multi-homing scenarios . . . . .	5
5.1	Control word (C-bit) . . . . .	5
5.2	Sequence flag (S-bit) . . . . .	6
6	Illustrative diagram . . . . .	6
7	Security Considerations . . . . .	7
8	IANA Considerations . . . . .	7
9	References . . . . .	7
9.1	Normative References . . . . .	7
	Authors' Addresses . . . . .	8

## 1 Introduction

[RFC4761] describes the concepts and signaling for using BGP to setup a VPLS. It specifies the BGP VPLS NLRI that a PE may require other PEs in the same VPLS to include (or not) control-word and sequencing information in VPLS frames sent to this PE.

The use of control word (CW) helps prevent mis-ordering of IPv4 or IPv6 PW traffic over ECMP-paths/LAG-bundles. [RFC4385] describes the format for control-word that may be used over point-2-point PWs and over a VPLS. It along with [RFC3985] also describes sequencing of frames.

However, [RFC4761] does not specify the behavior of PEs in a mixed environment where some PEs support control-word/sequencing and others do not.

### 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 Problem

[RFC4761] specifies the use of a VPLS BGP NLRI by which a given PE advertises the required behavior off multiple PEs participating in the same VPLS. The behavior required off the multiple PEs identified by the NLRI indicates the VPLS label they should use in the VPLS traffic being forwarded to this PE. Additionally, by using the "control flags" this PE specifies whether the other PEs (in the same VPLS) should use control-word or sequenced-delivery for frames forwarded to this PE. These are respectively indicated by the C and the S bits in the "control flags" as specified in section 3.2.4 in [RFC4761].

[RFC4761] requires that if the advertising PE sets the C and S bits, the receiving PE MUST honor the same by inserting control word (CW) and by including sequence numbers respectively.

However, in a BGP VPLS deployment there would often be cases where a PE receiving the VPLS BGP NLRI may not have the ability to insert a CW or include sequencing information inside PW frames. Thus, the behavior of BGP VPLS needs to be further specified.

This document updates the meaning of the control flags in layer2 extended community in the BGP VPLS NLRI and specifies the resulting forwarding behavior for a mixed mode environment where not every PE



in a VPLS has the ability or the configuration to honor the control flags received from the PE advertising the BGP NLRI.

### 3 Updated meaning of control flags in the layer2 info extended community

Current specification does not allow for the CW setting to be negotiated. Rather, if a PE sets the C-bit, it expects to receive VPLS frames with a control word, and will send frames the same way. If the PEs at both ends of a pseudowire do not agree on the setting of the C-bit, the PW does not come up. The expected behavior is similar for the S-bit.

This memo updates the meaning of the C-bit and the S-bit in the control flags.

#### 3.1 Control word (C-bit)

If a PE sets the C-bit in its NLRI, it means that the PE has ability to send and receive frames with a control word. If the PEs at both ends of a PW set the C-bit, control words **MUST** be used in both directions of the PW. If both PEs send a C-bit of 0, control words **MUST** not be used on the PW. These two cases behave as before.

However, if the PEs don't agree on the setting of the C-bit, control words **MUST** not be used on that PW but the PW **MUST NOT** be prevented from coming up due to this mismatch. So, the PW **MUST** still come up. This behavior is new; the old behavior was that the PW doesn't come up.

#### 3.2 Sequence flag (S-bit)

Current BGP VPLS implementation do not allow for S-bit setting to be negotiated either. If the PE sets the S-bit, it expects to receive VPLS frames with sequence numbers, and will send the frames with the sequence numbers as well. This memo further specifies the existing behavior. If the PEs on the both ends of the PW set the S-bit, then both PEs **MUST** include the PW sequence numbers. If the PEs at both ends of the PW do not agree on the setting of the S-bit, the PW **SHOULD NOT** come up at all.

#### 4 Using p2mp LSP as transport for BGP VPLS

BGP VPLS can be used over point-2-point LSPs acting as transport between the VPLS PEs. Alternately, BGP VPLS may also be used over p2mp LSPs with the source of the p2mp LSP rooted at the PE advertising the VPLS BGP NLRI.

In a network that uses p2mp LSPs as transport for BGP VPLS, in a given VPLS there may be some PEs that support control-word while others do not. Similarly, for sequencing of frames.

In such a setup, a source PE that supports control-word should setup 2 different p2mp LSPs such that:

- one p2mp LSP will carry CW-marked frames to those PEs that advertised C-bit as 1, and
- the other p2mp LSP will carry frames without CW to those PEs that advertised C-bit as 0.

However, the set of leaves on the 2 p2mp LSPs (rooted at the given PE) MUST NOT contain any PEs that advertised a value for S-bit different from what this PE itself is advertising.

Using 2 different p2mp LSPs to deliver frames with and without CW to different PEs ensures that this PE honors the C-bit advertised by the other PEs.

By not having PEs that advertised their S-bit value differently (from what this PE advertised) on either of the p2mp LSPs, it is ensured that this PE is sending VPLS frames only to those PEs that agree on the setting of S-bit with this PE.

#### 5. Treatment of C and S bits in multi-homing scenarios

##### 5.1 Control word (C-bit)

In multi-homed environment, different PEs may effectively represent the same service destination end point. It could be assumed that the end-to-end PW establishment process should follow the same rules when it comes to control word requirement, meaning setting the C-bit would be enforced equally toward both primary and backup designated forwarder together.

However, it is to be noted that in the multi-homing case, each PW SHOULD be evaluated independently. Assuming the below specified network topology, there could be the case where PW between PE2 and PE1 could have control word signaled via extended community and would be used in the VPLS frame, while PE2 to PE4 PW would not

insert the control word in the VPLS frame due to C-bit mismatch. The rest of PEs multi-homing behavior should simply follow the rules specified in draft-ietf-l2vpn-vpls-multihoming-06.

## 5.2 Sequence flag (S-bit)

In multi-homed environment, different PEs may effectively represent the same service destination end point. In this case, the rules for end-to-end PW establishment SHOULD follow the same rules when it comes to sequence bit requirements. Consider the case below with CE5 being multi-homed to PE4 and PE1. The PW behavior is similar to the C-word scenario so that the insertion of S-bit evaluation SHOULD be independent per PW. However, because S-bit mismatch between two end-point PEs yields in no PW establishment, in the case where PE4 doesn't support S-bit, only one PW would be established, between PE1 and PE2. Thus, even though CE5 is physically multi-homed, due to PE4's lack of support for S-bit, and no PW between PE1 and PE4, CE5 would not be multi-homed any more.

## 6 Illustrative diagram

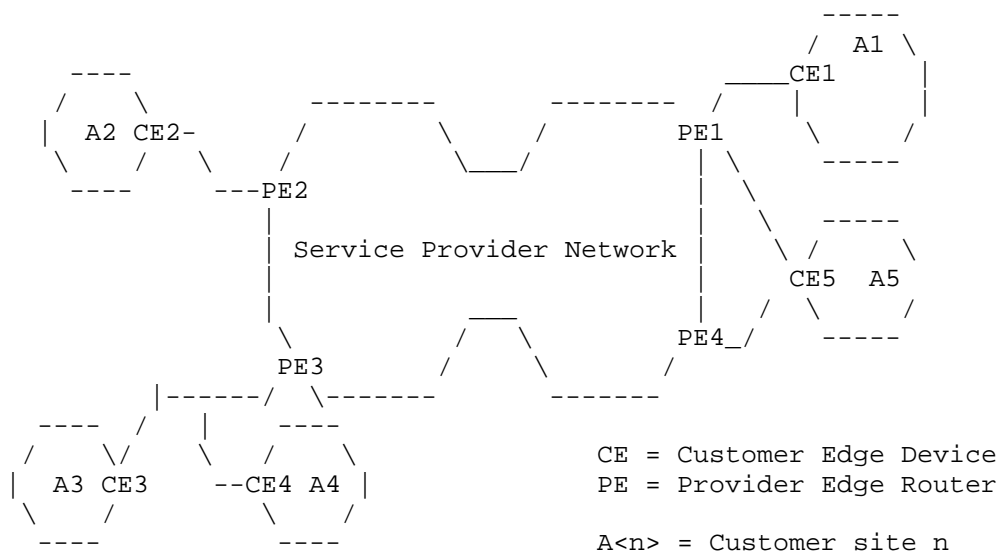


Figure 1: Example of a VPLS

In the above topology, let there be a VPLS configured with the PEs as displayed. Let PE1 be the PE under consideration that is CW enabled. Let PE2 and PE3 also be CW enabled. Let PE4 not be CW enabled. PE1

will advertise a VPLS BGP NLRI, containing the C/S bits marked as 1. PE2 and PE3 on learning of NLRI from PE1, shall include the control word in VPLS frames being forwarded to PE1. However, PE4 which does not have the ability to include control-word.

As per [RFC4761], PE1 would have an expectation that all other PEs forward traffic to it by including CW. That expectation cannot be met by PE4 in this example. Thus, as per [RFC4761] the PW between PE1 and PE4 does not come up.

However, this document addresses how to support the mixed-CW environment as above. PE1 will bring up the PW with PE4 despite the CW mismatch. Additionally, it will setup its data-plane such that it will strip the control-word only for those VPLS frames that are received from PEs that are themselves indicating their desire to receive CW marked frames. So, PE1 will setup its data plane to strip-off the CW only for VPLS frames received from PEs PE2 and PE3. PE1 will setup its data plane to not strip CW from frames received from PE4.

## 7 Security Considerations

No new security issues.

## 8 IANA Considerations

None.

## 9 References

### 9.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4761] Kompella, K., Y. Rekhter, Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling, RFC 4761, January 2007.
- [RFC4385] Bryant, S., Swallow G., Martini L., D. McPherson, Pseudowire Emulation Edge-to-Edge (PWE3) Control Word, RFC 4385, February 2006.
- [RFC3985] Bryant, S., P. Pate, Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture, RFC3985, March 2005.

Authors' Addresses

Ravi Singh  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US  
EMail: ravis@juniper.net

Kireeti Kompella  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US  
EMail: kireeti@juniper.net

Senad Palislamovic  
Alcatel-Lucent  
EMail: senad.palislamovic@alcatel-lucent.com