

L3VPN Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 17, 2014

H. Jeng  
AT&T  
L. Jalil  
Verizon  
R. Bonica  
Y. Rekhter  
Juniper Networks  
K. Patel  
Cisco Systems  
L. Yong  
Huawei Technologies  
February 13, 2014

Covering Prefixes Outbound Route Filter for BGP-4  
draft-bonica-l3vpn-orf-covering-prefixes-01

Abstract

This document defines a new ORF-type, called the "Covering Prefixes ORF (CP-ORF)". CP-ORF is applicable in Virtual Hub-and-Spoke VPNs. It also is applicable in BGP/MPLS Ethernet VPN (EVPN) Networks.

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 August 17, 2014.

## 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 . . . . .	2
1.1. Terminology . . . . .	2
2. CP-ORF Encoding . . . . .	3
3. Processing Rules . . . . .	5
4. Applicability In Virtual Hub-and-Spoke VPNs . . . . .	8
5. Applicability In Network Virtualization Overlays . . . . .	10
6. Clean-up . . . . .	12
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	12
9. Acknowledgements . . . . .	13
10. Normative References . . . . .	13
Authors' Addresses . . . . .	14

## 1. Problem Statement

A BGP [RFC4271] speaker can send Outbound Route Filters (ORF) [RFC5291] to a peer. The peer uses ORFs to filter routing updates that it sends to the BGP speaker. Using ORF, a speaker can realize a "route pull" paradigm in BGP, in which the speaker, on demand, pulls certain routes from the peer.

This document defines a new ORF-type, called the "Covering Prefixes ORF (CP-ORF)". CP-ORF is applicable in Virtual Hub-and-Spoke VPNs [RFC7024] [RFC4364]. It also is applicable BGP/MPLS Ethernet VPN (EVPN) [I-D.ietf-l2vpn-evpn] Networks.

## 1.1. Terminology

This document uses the following terms:

- o Address Family Indicator (AFI) - defined in [RFC4760]

- o Subsequent Address Family Indicator (SAFI) - defined in [RFC4760]
- o VPN IP Default Route - defined in [RFC7024].
- o V-Hub - defined in [RFC7024].
- o V-Spoke - defined in [RFC7024].
- o BGP/MPLS Ethernet VPN (EVPN) - defined in [I-D.ietf-l2vpn-evpn]
- o EVPN Instance (EVI) - defined in [I-D.ietf-l2vpn-evpn]
- o Default MAC Route (DMR) - An EVPN Route with MAC Address length equal to 0. See Section 10.2.1 of [I-D.ietf-l2vpn-evpn] for details.
- o Default Gateway (DMG) - An EVPN PE that advertises a DMR

## 2. CP-ORF Encoding

[RFC5291] augments the BGP ROUTE-REFRESH message so that it can carry ORF entries. When the ROUTE-REFRESH message carries ORF entries, it includes the following fields:

- o AFI [IANA.AFI]
- o SAFI [IANA.SAFI]
- o When-to-refresh (IMMEDIATE or DEFERRED)
- o ORF Type
- o Length (of ORF entries)

The ROUTE-REFRESH message also contains a list of ORF entries. Each ORF entry contains the following fields:

- o Action (ADD, REMOVE, or REMOVE-ALL)
- o Match (PERMIT or DENY)

The ORF entry may also contain Type-specific information. Type-specific information is present only when the Action is equal to ADD or REMOVE. It is not present when the Action is equal to REMOVE-ALL.

When the BGP ROUTE-REFRESH message carries CP-ORF entries, the following conditions MUST be true:

- o ORF Type MUST be equal to CP-ORF. (The value of CP-ORF is TBD. See Section 7 for details.)
- o The AFI MUST be equal to IPv4, IPv6 or L2VPN
- o If the AFI is equal to IPv4 or IPv6, SAFI MUST be equal to MPLS-labeled VPN address
- o If the AFI is equal to L2VPN, the SAFI MUST be equal to BGP EVPN
- o Match field MUST be equal to PERMIT

Figure 1 depicts the encoding of the CP-ORF type-specific information.

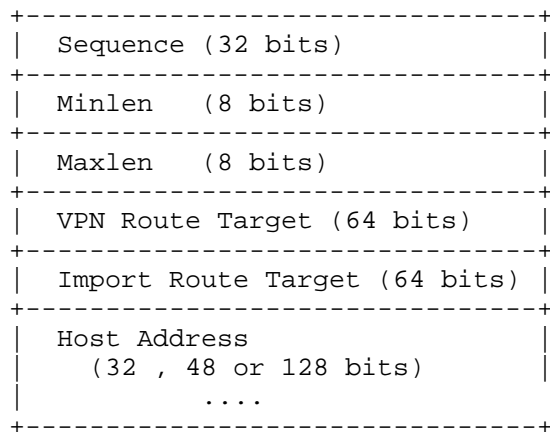


Figure 1: CP-ORF Type-specific Encoding

The Sequence field specifies the relative ordering of the entry among all CP-ORF entries.

The CP-ORF recipient uses the following fields to identify routes that match the CP-ORF:

- o Minlen
- o Maxlen
- o VPN Route Target
- o Host Address

See Section 3 for details.

The CP-ORF recipient marks routes that match CP-ORF with the Import Route Target before advertising those routes to the CP-ORF originator. See Section 3 for details.

If the ROUTE-REFRESH AFI is equal to IPv4, the length of the Host Address field is 32 bits. If the ROUTE-REFRESH AFI is equal to IPv6, the length of the Host Address field is 128 bits. If the ROUTE-REFRESH AFI is equal to L2VPN, the length of the Host Address field is 48 bits.

If the ROUTE-REFRESH AFI value is equal to IPv4 or IPv6, the following rules apply:

- o The value of Minlen MUST be less than or equal to the length of the Host Address field
- o The value of Maxlen MUST be less than or equal to the length of the Host Address field
- o The value of Minlen MUST be less than or equal to the value of Maxlen

If the ROUTE-REFRESH AFI is equal to L2VPN, the following rules apply:

- o The value of Minlen MUST be 48
- o The value of Maxlen MUST be 48

### 3. Processing Rules

According to [RFC4271], every BGP speaker maintains a single Loc-RIB. For each of its peers, the BGP speaker also maintains an Outbound Filter and an Adj-RIB-Out. The Outbound Filter defines policy that determines which Loc-RIB entries are processed into the corresponding Adj-RIB-Out. Mechanisms such as RT-Constraint [RFC4684] and ORF [RFC5291] enable a router's peer to influence the Outbound Filter. Therefore, the Outbound Filter for a given peer is constructed using a combination of the locally configured policy and the information received via RT-Constraint and ORF from the peer.

Using this model we can describe the operations of CP-ORF as follows:

When a BGP speaker receives a ROUTE-REFRESH message that contains a CP-ORF, and that ROUTE-REFRESH message that violates any of the encoding rules specified in Section 2, the BGP speaker MUST log the event and ignore the entire ROUTE-REFRESH message.

Otherwise, the BGP speaker processes each CP-ORF entry as indicated by the Action field. If the Action is equal to ADD, the BGP speaker adds the CP-ORF entry to the Outbound Filter associated with the peer in the position specified by the Sequence field. If the Action is equal to REMOVE, the BGP speaker removes the CP-ORF entry from the Outbound Filter. If the Action is equal to REMOVE-ALL, the BGP speaker removes all CP-ORF entries from the Outbound Filter.

Whenever the BGP speaker applies an Outbound Filter to a route contained by its Loc-RIB, it evaluates the route in terms of the CP-ORF entries first. It then evaluates the route in terms of the remaining, non-CP-ORF entries. The rules for the former are described below. The rules for the latter are outside the scope of this document.

The following route types can match a CP-ORF:

- o IPv4-VPN
- o IPv6-VPN
- o L2VPN (L2VPN MAC Advertisement only. See Section 8.2 of [I-D.ietf-l2vpn-evpn] for details.)

In order for an IPv4-VPN route or IPv6-VPN route to match a CP-ORF, all of the following conditions MUST be true:

- o the route carries an RT whose value is the same as the CP-ORF VPN Route Target
- o the route prefix length is greater than or equal to the CP-ORF Minlen plus 64 (i.e., the length of a VPN Route Distinguisher)
- o the route prefix length is less than or equal to the CP-ORF Maxlen plus 64 (i.e., the length of a VPN Route Distinguisher)
- o ignoring the Route Distinguisher, the leading bits of the route prefix are identical to the leading bits of the CP-ORF Host Address. CP-ORF Minlen defines the number of bits that must be identical.

The BGP speaker ignores Route Distinguishers when determining whether a prefix covers a host address. For example, assume that a CP-ORF carries the following information:

- o Minlen equal to 1
- o Maxlen equal to 32

- o Host Address equal to 192.0.2.1

Assume also that Loc-RIB contains routes for the following IPv4-VPN prefixes, and that all of these routes carry an RT whose value is the same as the CP-ORF VPN Route Target:

- o 1:0.0.0.0/64.
- o 2:192.0.2.0/88
- o 3:192.0.2.0/89

For the purposes of this evaluation, 2:192.0.2.0/88 and 3:192.0.2.0/89 cover 192.0.2.1. This is because the search algorithm ignores Route Distinguishers. However, 1:0.0.0.0/64 does not cover 192.0.2.1, because its length (64) is less than the CP-ORF Minlen (1) plus the length of an L3VPN Route Distinguisher (64).

In order for an EVPN route match a CP-ORF, all of the following conditions MUST be true:

- o the route carries an RT whose value is the same as the CP-ORF VPN Route Target
- o the final 48 bits of the EVPN MAC Address are identical to the CP-ORF Host Address

If a route matches the selection criteria of a CP-ORF entry, and it does not violate any subsequent rule specified by the Outbound Filter (e.g., rules that reflect local policy, or rules that are due to RT-Constrains), the BGP speaker places the route into the Adj-RIB-Out. In Adj-RIB-Out, the BGP speaker adds the CP-ORF Import Route Target to the list of Route Targets that the route already carries. As a result of being placed in Adj-RIB-Out, the route is advertised to the peer associated with the Adj-RIB-Out.

Receiving CP-ORF entries with REMOVE or REMOVE-ALL Actions may cause a route that has previously been installed in a particular Adj-RIB-Out be excluded from that Adj-RIB-Out. In this case, as specified in [RFC4271], "the previously advertised route in that Adj-RIB-Out MUST be withdrawn from service by means of an UPDATE message".

[RFC5291] states that a BGP speaker should respond to a ROUTE REFRESH message as follows:

"If the When-to-refresh indicates IMMEDIATE, then after processing all the ORF entries carried in the message the speaker re-advertises to the peer routes from the Adj-RIB-Out associated with the peer that

have the same AFI/SAFI as what is carried in the message, and taking into account all the ORF entries for that AFI/SAFI received from the peer. The speaker MUST re-advertise all the routes that have been affected by the ORF entries carried in the message, but MAY also re-advertise the routes that have not been affected by the ORF entries carried in the message."

When the ROUTE-REFRESH message includes one or more CP-ORF entries, the BGP speaker MUST re-advertise routes that have been affected by ORF entries carried by the message. While the speaker MAY also re-advertise the routes that have not been affected by the ORF entries carried in the message, this memo RECOMMENDS not to re-advertise the routes that have not been affected.

#### 4. Applicability In Virtual Hub-and-Spoke VPNs

In a Virtual Hub-and-Spoke environment, VPN sites are attached to Provider Edge (PE) routers, where for a given VPN some of these PEs may act as V-hubs, while others as V-spokes. This memo assumes that PEs exchange VPN-IP routes using Route Reflectors (RRs).

This memo also assumes that RED-VPN sites are attached to PE routers, V-hub1 and V-spoke1. All of these devices advertise RED-VPN routes to a RR. They mark these routes with a route target, which we will call RT-RED.

V-hub1 serves the RED-VPN. Therefore, V-hub1 advertises a VPN IP default route for the RED-VPN to the RR, carrying the route target RT-RED-FROM-HUB1.

V-spoke1 establishes a BGP session with the RR, negotiating the CP-ORF capability, as well as the Multiprotocol Extensions Capability [RFC2858]. Upon establishment of the BGP session, the RR does not advertise any routes to V-spoke1. The RR will not advertise any routes until it receives either a ROUTE-REFRESH message or a BGP UPDATE message containing a Route Target Membership NLRI [RFC4684].

Immediately after the BGP session is established, V-spoke1 sends the RR a BGP UPDATE message containing a Route Target Membership NLRI. The Route Target Membership NLRI specifies RT-RED-FROM-HUB1 as its route target. In response to the BGP-UPDATE message, the RR advertises the VPN IP default route for the RED-VPN to V-spoke1. This route carries the route target RT-RED-FROM-HUB1. V-spoke1 subjects this route to its import policy and accepts it because it carries the route target RT-RED-FROM-HUB1.

Now, V-spoke1 begins normal operation, sending all of its RED-VPN traffic through V-hub1. At some point, V-spoke1 determines that it



might benefit from a more direct route to a destination. (Criteria by which V-spoke1 determines that it needs a more direct route are beyond the scope of this document.)

In order to discover a more direct route, V-spoke1 assigns a unique numeric identifier to the destination. V-spoke1 then sends a ROUTE-REFRESH message to the RR, containing the following information:

- o AFI is equal to IPv4 or IPv6, as appropriate
- o SAFI is equal to "MPLS-labeled VPN address"
- o When-to-refresh is equal IMMEDIATE
- o Action is equal to ADD
- o Match is equal to PERMIT
- o ORF Type is equal to CP-ORF
- o CP-ORF Sequence is equal to the identifier associated with the destination
- o CP-ORF Minlen is equal to 1
- o CP-ORF Maxlen is equal to 32 or 128, as appropriate
- o CP-ORF VPN Route Target is equal to RT-RED
- o CP-ORF Host Address is equal the destination address
- o CP-ORF Import Route Target is equal to RT-RED-FROM-HUB1

Upon receipt of the ROUTE-REFRESH message, the RR must ensure that it carries all routes belonging to the RED-VPN. In at least one special case, where all of the RR clients are V-spokes and none of the RR clients are V-hubs, the RR will lack some or all of the required RED-VPN routes. So, the RR sends a BGP UPDATE message containing a Route Target Membership NLRI for VPN-RED to all of its peers. This causes the peers to advertise VPN-RED routes to the RR, if they have not done so already.

Next, the RR adds the received CP-ORF to the Outbound Filter associated with V-spoke1. Using the procedures in Section 3, the RR determines whether any of the routes in its Loc-RIB satisfy the selection criteria of the newly updated Outbound Filter. If any routes satisfy the match criteria, they are added to the Adj-RIB-Out associated with V-spoke1. In Adj-RIB-Out, the RR adds RT-RED-FROM-

HUB1 to the list of Route Targets that the route already carries. Finally, RR advertises the newly added routes to V-spoke1. The advertised routes may specify either V-hub1 or any other node as the NEXT-HOP.

V-spoke1 subjects the advertised routes to its import policy and accepts them because they carry the route target RT-RED-FROM-HUB1.

V-spoke1 may repeat this process whenever it discovers another flow that might benefit from a more direct route to its destination.

## 5. Applicability In Network Virtualization Overlays

In an EVPN environment, Layer 2 networks are connected to Provider Edge (PE) devices. PE devices can be real or virtualized. Within a given EVPN, one or more EVPN Instances (EVI) can serve as a Default MAC Gateway (DMG). Each DMG advertises a Default MAC Route (DMR) to the rest of the EVIs in the EVPN. EVIs use the DMR to forward traffic destined to MAC addresses for which they do not have a corresponding MAC Advertisement Route.

For the purposes of example, assume the following:

- o Layer 2 Networks belonging to the RED-VPN are attached to PEs that support EVPN.
- o At any given point in time, an end-system that belongs to the RED-VPN communicates with only a small subset of other end-systems that belong to the RED-VPN. Therefore, at any given point in time, most of the PEs that serve the RED-VPN use only a small subset of the MAC Advertisement Routes in the RED-VPN.
- o One PE device serves as a DMG for the RED-VPN. We will call this device DMG 1. The RED-VPN EVI on DMG 1 is provisioned with RT-RED-FROM-HUB1 as its export RT, and RT-RED as its import RT.
- o Another PE device that hosts an EVI of the RED-VPN can not accommodate all RED-VPN MAC Advertisement routes. We will call this device Spoke 1. This EVI is provisioned with RT-RED as its export RT, and RT-RED-FROM-HUB1 as its import RT.
- o All PE devices that have EVIs of the RED-VPN advertise various EVPN routes, including MAC Advertisement Routes to one or more RRs.

DMG 1 serves the RED-VPN. Therefore, DMG 1 advertises a DMR for the RED-VPN to the RR, carrying the route target RT-RED-FROM-HUB1.

Spoke 1 establishes a BGP session with the RR, negotiating the CP-ORF capability, as well as the Multiprotocol Extensions Capability [RFC2858]. Upon establishment of the BGP session, the RR does not advertise any routes to Spoke 1. The RR will not advertise any routes until it receives either a ROUTE-REFRESH message or a BGP UPDATE message containing a Route Target Membership NLRI [RFC4684].

Immediately after the BGP session is established, Spoke 1 sends the RR a BGP UPDATE message containing a Route Target Membership NLRI. The Route Target Membership NLRI specifies RT-RED-FROM-HUB1 as its route target. In response to the BGP-UPDATE message, the RR advertises the DMR for the RED-VPN to Spoke 1. This route carries the route target RT-RED-FROM-HUB1. Spoke 1 subjects this route to its import policy and accepts it because it carries the route target RT-RED-FROM-HUB1.

Now, Spoke 1 begins normal operation, sending all of its RED-VPN traffic through DMG 1. At some point, Spoke 1 determines that it might benefit from a more direct route to a destination. (Criteria by which V-spoke1 determines that it needs a more direct route are beyond the scope of this document.)

In order to discover a more direct route, Spoke 1 assigns a unique numeric identifier to the destination. Spoke 1 then sends a ROUTE-REFRESH message to the RR, containing the following information:

- o AFI is equal to L2VPN
- o SAFI is equal to BGP EVPN
- o When-to-refresh is equal IMMEDIATE
- o Action is equal to ADD
- o Match is equal to PERMIT
- o ORF Type is equal to CP-ORF
- o CP-ORF Sequence is equal to the identifier associated with the destination
- o CP-ORF Minlen is equal to 48
- o CP-ORF Maxlen is equal to 48
- o CP-ORF VPN Route Target is equal to RT-RED
- o CP-ORF Host Address is equal the destination address

- o CP-ORF Import Route Target is equal to RT-RED-FROM-HUB1

Next, the RR adds the received CP-ORF to the Outbound Filter associated with Spoke 1. Using the procedures in Section 3, the RR determines whether any of the MAC Advertisement routes in its Loc-RIB satisfy the selection criteria of the newly updated Outbound Filter. If any of these routes satisfy the match criteria, they are added to the Adj-RIB-Out associated with Spoke 1. In Adj-RIB-Out, the RR adds RT-RED-FROM-HUB1 to the list of Route Targets that the route already carries. Finally, RR advertises the newly added routes to Spoke 1. The advertised routes carry as their NEXT-HOP the address of the PE device from which the routes have been originated.

Spoke 1 subjects the the MAC Advertisement Routes received from RR to its import policy and accepts them because they carry the route target RT-RED-FROM-HUB1.

Spoke 1 may repeat this process whenever it discovers another flow that might benefit from a more direct route to its destination.

## 6. Clean-up

Each CP-ORF consumes memory and compute resources on the device that supports it. Therefore, in order to obtain optimal performance, BGP speakers periodically evaluate all CP-ORFs that they have originated and remove unneeded CP-ORFs. The criteria by which a BGP speaker identifies unneeded CP-ORF entries is a matter of local policy, and is beyond the scope of this document.

## 7. IANA Considerations

IANA is requested to assign a All Covering Prefixes ORF Type from the BGP Outbound Route Filtering (ORF) Types Registry.

## 8. Security Considerations

Each CP-ORF consumes memory and compute resources on the device that supports it. Therefore, a device supporting CP-ORF take the following steps to protect itself from oversubscription:

- o When negotiating the ORF capability, advertise willingness to receive the CP-ORF only to known, trusted iBGP peers
- o Enforce a per-peer limit on the number of CP-ORFs that can be installed at any given time. Ignore all requests to add CP-ORFs beyond that limit

## 9. Acknowledgements

The authors wish to acknowledge Han Nguyen and James Uttaro for their comments and contributions.

## 10. Normative References

- [I-D.ietf-l2vpn-evpn]  
Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04 (work in progress), July 2013.
- [IANA.AFI]  
IANA, "abbrev="Address Family Numbers"", <<http://www.iana.org/assignments/address-family-numbers/address-family-numbers.xhtml>>.
- [IANA.SAFI]  
IANA, "abbrev="Subsequent Address Family Identifiers (SAFI) Parameters"", <<http://www.iana.org/assignments/safi-namespace/safi-namespace.xhtml#safi-namespace-2>>.
- [RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [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.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

- [RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, August 2008.
- [RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, August 2008.
- [RFC7024] Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", RFC 7024, October 2013.

## Authors' Addresses

Huajin Jeng  
AT&T

Email: [hj2387@att.com](mailto:hj2387@att.com)

Luay Jalil  
Verizon

Email: [luay.jalil@verizon.com](mailto:luay.jalil@verizon.com)

Ron Bonica  
Juniper Networks  
2251 Corporate Park Drive  
Herndon, Virginia 20170  
USA

Email: [rbonica@juniper.net](mailto:rbonica@juniper.net)

Yakov Rekhter  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, California 94089  
USA

Email: [yakov@juniper.net](mailto:yakov@juniper.net)

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

Email: keyupate@cisco.com

Lucy Yong  
Huawei Technologies

Email: lucy.yong@huawei.com

Network Working Group  
INTERNET-DRAFT  
Category: Standards Track

A. Sajassi  
Cisco

N. Bitar  
Verizon

R. Aggarwal  
Arktan

J. Drake  
Juniper Networks

W. Henderickx  
Alcatel-Lucent

Aldrin Isaac  
Bloomberg

J. Uttaro  
AT&T

Expires: August 12, 2014

February 12, 2014

BGP MPLS Based Ethernet VPN  
draft-ietf-l2vpn-evpn-05

Status of this Memo

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

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

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

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

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

Copyright and License Notice

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



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

## Abstract

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

## Table of Contents

1. Specification of requirements . . . . .	5
2. Terminology . . . . .	5
3. Introduction . . . . .	6
4. BGP MPLS Based EVPN Overview . . . . .	6
5. Ethernet Segment . . . . .	7
6. Ethernet Tag . . . . .	10
6.1 VLAN Based Service Interface . . . . .	10
6.2 VLAN Bundle Service Interface . . . . .	11
6.2.1 Port Based Service Interface . . . . .	11
6.3 VLAN Aware Bundle Service Interface . . . . .	11
6.3.1 Port Based VLAN Aware Service Interface . . . . .	11
7. BGP EVPN NLRI . . . . .	12
7.1. Ethernet Auto-Discovery Route . . . . .	12
7.2. MAC/IP Advertisement Route . . . . .	13
7.3. Inclusive Multicast Ethernet Tag Route . . . . .	14
7.4 Ethernet Segment Route . . . . .	14
7.5 ESI Label Extended Community . . . . .	15
7.6 ES-Import Route Target . . . . .	15
7.7 MAC Mobility Extended Community . . . . .	16
7.8 Default Gateway Extended Community . . . . .	16
8. Multi-homing Functions . . . . .	16
8.1 Multi-homed Ethernet Segment Auto-Discovery . . . . .	17
8.1.1 Constructing the Ethernet Segment Route . . . . .	17
8.2 Fast Convergence . . . . .	17
8.2.1 Constructing the Ethernet A-D per Ethernet Segment (ES) Route . . . . .	18
8.2.1.1. Ethernet A-D Route Targets . . . . .	18
8.3 Split Horizon . . . . .	19
8.3.1 ESI Label Assignment . . . . .	19
8.3.1.1 Ingress Replication . . . . .	19
8.3.1.2. P2MP MPLS LSPs . . . . .	20

8.4 Aliasing and Backup-Path . . . . .	21
8.4.1 Constructing the Ethernet A-D per EVPN Instance (EVI)	
Route . . . . .	22
8.4.1.1 Ethernet A-D Route Targets . . . . .	23
8.5 Designated Forwarder Election . . . . .	23
8.6. Interoperability with Single-homing PEs . . . . .	25
9. Determining Reachability to Unicast MAC Addresses . . . . .	26
9.1. Local Learning . . . . .	26
9.2. Remote learning . . . . .	27
9.2.1. Constructing the BGP EVPN MAC/IP Address	
Advertisement . . . . .	27
9.2.2 Route Resolution . . . . .	29
10. ARP and ND . . . . .	30
10.1 Default Gateway . . . . .	30
11. Handling of Multi-Destination Traffic . . . . .	32
11.1. Construction of the Inclusive Multicast Ethernet Tag	
Route . . . . .	32
11.2. P-Tunnel Identification . . . . .	32
12. Processing of Unknown Unicast Packets . . . . .	33
12.1. Ingress Replication . . . . .	34
12.2. P2MP MPLS LSPs . . . . .	34
13. Forwarding Unicast Packets . . . . .	35
13.1. Forwarding packets received from a CE . . . . .	35
13.2. Forwarding packets received from a remote PE . . . . .	36
13.2.1. Unknown Unicast Forwarding . . . . .	36
13.2.2. Known Unicast Forwarding . . . . .	36
14. Load Balancing of Unicast Frames . . . . .	37
14.1. Load balancing of traffic from an PE to remote CEs . . . . .	37
14.1.1 Single-Active Redundancy Mode . . . . .	37
14.1.2 All-Active Redundancy Mode . . . . .	38
14.2. Load balancing of traffic between an PE and a local CE . . . . .	39
14.2.1. Data plane learning . . . . .	40
14.2.2. Control plane learning . . . . .	40
15. MAC Mobility . . . . .	40
15.1. MAC Duplication Issue . . . . .	42
15.2. Sticky MAC addresses . . . . .	42
16. Multicast & Broadcast . . . . .	42
16.1. Ingress Replication . . . . .	43
16.2. P2MP LSPs . . . . .	43
16.2.1. Inclusive Trees . . . . .	43
17. Convergence . . . . .	44
17.1. Transit Link and Node Failures between PEs . . . . .	44
17.2. PE Failures . . . . .	44
17.3. PE to CE Network Failures . . . . .	44
18. Frame Ordering . . . . .	45
19. Acknowledgements . . . . .	45
20. Security Considerations . . . . .	46
21. Contributors . . . . .	47

22. IANA Considerations . . . . .	47
23. References . . . . .	47
23.1 Normative References . . . . .	48
23.2 Informative References . . . . .	48
24. Author's Address . . . . .	48

## 1. Specification of requirements

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

## 2. Terminology

Bridge Domain:

Broadcast Domain:

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

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

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

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

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

LACP: Link Aggregation Control Protocol

MP2MP: Multipoint to Multipoint

P2MP: Point to Multipoint

P2P: Point to Point

Single-Active Redundancy Mode: When only a single PE, among a group of PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet Segment, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode: When all PEs attached to an Ethernet

segment are allowed to forward traffic to/from that Ethernet Segment, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

### 3. Introduction

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

### 4. BGP MPLS Based EVPN Overview

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

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

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

CEs to connect to multiple active points of attachment. It also improves convergence times in the event of certain network failures.

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

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

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

## 5. Ethernet Segment

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

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

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

In a network with managed and not-managed CEs, the ESI has the

following format:

```

+---+---+---+---+---+---+---+---+---+---+
| T |           ESI Value           |
+---+---+---+---+---+---+---+---+---+---+

```

Where:

T (ESI Type) is a 1-byte field (most significant octet) that specifies the format of the remaining nine bytes (ESI Value). The following 6 ESI types can be used:

- Type 0 (T=0x00) - This type indicates an arbitrary nine-octet ESI value, which is managed and configured by the operator.

- Type 1 (T=0x01) - When IEEE 802.1AX LACP is used between the PEs and CEs, this ESI type indicates an auto-generated ESI value determined from LACP by concatenating the following parameters:

- + CE LACP six octets System MAC address. The CE LACP System MAC address MUST be encoded in the high order six octets of the ESI Value field.
- + CE LACP two octets Port Key. The CE LACP port key MUST be encoded in the two octets next to the System MAC address.
- + The remaining octet will be set to 0x00.

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

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

- Type 2 (T=0x02) - This type is used in the case of indirectly connected hosts via a bridged LAN between the CEs and the PEs. The ESI Value is auto-generated and determined based on the Layer 2 bridge protocol as follows: If MST is used in the bridged LAN then the value of the ESI is derived by listening to BPDUs on the Ethernet segment. To achieve this the PE is not required to run MST. However the PE must learn the Root Bridge MAC address and Bridge Priority of the root of the Internal Spanning Tree (IST) by listening to the BPDUs. The ESI Value is constructed as follows:

- + Root Bridge six octets MAC address. The Root Bridge MAC address MUST be encoded in the high order six octets of the

ESI Value field.

- + Root Bridge two octets Priority. The CE LACP port key MUST be encoded in the two octets next to the Root Bridge MAC address.
- + The remaining octet will be set to 0x00.

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

- Type 3 (T=0x03) - This type indicates a MAC-based ESI Value that can be auto-generated or configured by the operator. The ESI Value is constructed as follows:

- + System MAC address (six octets). The System MAC address MUST be encoded in the high order six octets of the ESI Value field.
- + Local Discriminator value (three octets). The Local Discriminator MUST be encoded in the low order three octets of the ESI Value.

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

- Type 4 (T=0x04) - This type indicates an IP-based ESI Value that can be auto-generated or configured by the operator. The ESI Value is constructed as follows:

- + IP address (four octets). This is an IPv4 address owned by the system and MUST be encoded in the high order four octets of the ESI Value field.
- + Local Discriminator value (four octets). The Local Discriminator MUST be encoded in the four octets next to the IP address.
- + The low order octet of the ESI Value will be set to 0x00.

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

- Type 5 (T=0x05) - This type indicates an AS-based ESI Value that can be auto-generated or configured by the operator. The ESI Value is constructed as follows:

- + AS number (four octets). This is an AS number owned by the system and MUST be encoded in the high order four octets of the ESI Value field. If a two-octet AS number is used, the high order extra two bytes will be 0x0000.



- + Local Discriminator value (four octets). The Local Discriminator MUST be encoded in the four octets next to the AS number.

- + The low order octet of the ESI Value will be set to 0x00.

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

## 6. Ethernet Tag

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

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

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

The following Ethernet Tag value is reserved:

- Ethernet Tag {0xFFFFFFFF} is known as MAX-ET

### 6.1 VLAN Based Service Interface

With this service interface, an EVPN instance consists of only a single broadcast domain (e.g., a single VLAN). Therefore, there is a one to one mapping between a VID on this interface and a MAC-VRF. Since a MAC-VRF corresponds to a single VLAN, it consists of a single bridge domain corresponding to that VLAN. If the VLAN is represented by different VIDs on different PEs, then each PE needs to perform VID

translation for frames destined to its attached CEs. In such scenarios, the Ethernet frames transported over MPLS/IP network SHOULD remain tagged with the originating VID and a VID translation MUST be supported in the data path and MUST be performed on the disposition PE. The Ethernet Tag Identifier in all EVPN routes MUST be set to 0.

## 6.2 VLAN Bundle Service Interface

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

### 6.2.1 Port Based Service Interface

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

## 6.3 VLAN Aware Bundle Service Interface

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

### 6.3.1 Port Based VLAN Aware Service Interface

This service interface is a special case of the VLAN Aware Bundle

service interface, where all of the VLANs on the port are part of the same service and map to the same bundle. The procedures are identical to those described in section 7.3.

## 7. BGP EVPN NLRI

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

Following is the format of the EVPN NLRI:

```

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

```

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

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

This document defines the following Route Types:

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

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

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

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

### 7.1. Ethernet Auto-Discovery Route

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

following:

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

For the purpose of BGP route key processing, only the Ethernet Segment ID and the Ethernet Tag ID are considered to be part of the prefix in the NLRI. The MPLS Label field is to be treated as a route attribute as opposed to being part of the route.

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

## 7.2. MAC/IP Advertisement Route

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

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octet)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0 or 4 or 16 octets)
MPLS Label1 (3 octets)
MPLS Label2 (0 or 3 octets)

For the purpose of BGP route key processing, only the Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, and IP

Address Address fields are considered to be part of the prefix in the NLRI. The Ethernet Segment Identifier and MPLS Label fields are to be treated as route attributes as opposed to being part of the "route".

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

### 7.3. Inclusive Multicast Ethernet Tag Route

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

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

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

### 7.4 Ethernet Segment Route

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

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

For procedures and usage of this route please see section 9.5 "Designated Forwarder Election". The IP address length is in bits.

## 7.5 ESI Label Extended Community

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

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

```

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

```

The low order bit of the flags octet is defined as the "Single-Active" bit. A value of 0 means that the multi-homed site is operating in All-Active redundancy mode and a value of 1 means that the multi-homed site is operating in Single-Active redundancy mode.

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

## 7.6 ES-Import Route Target

This is a new transitive Route Target extended community carried with the Ethernet Segment route. When used, it enables all the PEs connected to the same multi-homed site to import the Ethernet Segment routes. The value is derived automatically from the ESI by encoding the high order 6-byte portion of the 9-byte ESI Value in the ES-Import Route Target. The format of this extended community is as follows:

```

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

```

This document expands the definition of the Route Target extended community to allow the value of high order octet (Type field) to be 0x06 (in addition to the values specified in rfc4360). The value of

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

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

## 7.7 MAC Mobility Extended Community

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

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

```

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

```

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

## 7.8 Default Gateway Extended Community

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

## 8. Multi-homing Functions

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

## 8.1 Multi-homed Ethernet Segment Auto-Discovery

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

### 8.1.1 Constructing the Ethernet Segment Route

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

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

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

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

## 8.2 Fast Convergence

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

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



Otherwise, the PE updates the next-hop adjacencies to point to the backup PE(s).

#### 8.2.1 Constructing the Ethernet A-D per Ethernet Segment (ES) Route

This section describes the procedures used to construct the Ethernet A-D per ES route, which is used for fast convergence (as discussed above) and for advertising the ESI label used for split-horizon filtering (as discussed in section 9.3). Support of this route is MANDATORY.

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

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

The Ethernet Tag ID MUST be set to MAX-ET.

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

The "ESI Label Extended Community" MUST be included in the route. If All-Active redundancy mode is desired, then the "Single-Active" bit in the flags of the ESI Label Extended Community MUST be set to 0 and the MPLS label in that extended community MUST be set to a valid MPLS label value. The MPLS label in this Extended Community is referred to as the ESI label and MUST have the same value in each Ethernet A-D per ES route advertised for the ES. This label MUST be a downstream assigned MPLS label if the advertising PE is using ingress replication for receiving multicast, broadcast or unknown unicast traffic from other PEs. If the advertising PE is using P2MP MPLS LSPs for sending multicast, broadcast or unknown unicast traffic, then this label MUST be an upstream assigned MPLS label. The usage of this label is described in section 9.3.

If Single-Active redundancy mode is desired, then the "Single-Active" bit in the flags of the ESI Label Extended Community MUST be set to 1 and the ESI label MUST be set to zero.

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

Each Ethernet A-D per ES route MUST carry one or more Route Target (RT) attributes. The set of Ethernet A-D routes per ES MUST carry the entire set of RTs for all the EVPN instances to which the Ethernet Segment belongs.

### 8.3 Split Horizon

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

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

#### 8.3.1 ESI Label Assignment

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

##### 8.3.1.1 Ingress Replication

The non-DF PEs attached to a given ES that is operating in All-Active redundancy mode and that use ingress replication to receive BUM traffic advertise a downstream assigned ESI label in the set of Ethernet A-D per ES routes for that ES. This label MUST be programmed in the platform label space by the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for.

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

#### 8.3.1.2. P2MP MPLS LSPs

The non-DF PE attached to a given ES that is operating in All-Active redundancy mode and that use P2MP LSPs to send BUM traffic advertise an upstream assigned ESI label in the set of Ethernet A-D per ES routes for that ES. This label is upstream assigned by the PE that advertises the route. This label MUST be programmed by the other PEs, that are connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for. This label MUST also be programmed by the other PEs, that import the route but are not connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must be a POP with no other associated action.

Consider PE1 and PE2 that are multi-homed to CE1 on ES1 and operating in All-Active multi-homing mode. Also consider PE3 belongs to one of the EVPN instances of ES1. Further, assume that PE1 which is the non-DF, using P2MP MPLS LSPs to send BUM packets. When PE1 sends a BUM packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that it has assigned for the ESI that the packet was received on. The resulting packet is further encapsulated in the P2MP MPLS label stack necessary to transmit the

packet to the other PEs. Penultimate hop popping MUST be disabled on the P2MP LSPs used in the MPLS transport infrastructure for EVPN. When PE2 receives this packet, it de-capsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1, then PE2 MUST NOT forward the packet onto ES1. When PE3 receives this packet, it de-capsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1 and PE3 is not connected to ES1, then PE3 MUST pop the label and flood the packet over all local ESIs in that EVPN instance. It should be noted that when PE2 sends a BUM frame over a P2MP LSP, it does not need to encapsulate the frame with an ESI label because it is the DF for that VLAN.

#### 8.4 Aliasing and Backup-Path

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

To address this issue, EVPN introduces the concept of 'Aliasing' which is the ability of a PE to signal that it has reachability to an EVPN instance on a given ES even when it has learnt no MAC addresses from that EVI/ES. The Ethernet A-D per EVI route is used for this purpose. A remote PE that receives a MAC advertisement route with non-reserved ESI SHOULD consider the advertised MAC address to be reachable via all PEs that have advertised reachability to that MAC address' EVI/ES via the combination of an Ethernet A-D per EVI route for that EVI/ES (and Ethernet Tag if applicable) AND Ethernet A-D per ES routes for that ES with the 'Single-Active' bit in the flags of the ESI Label Extended Community set to 0.

Note that the Ethernet A-D per EVI route may be received by a remote PE before it receives the set of Ethernet A-D per ES routes. Therefore, in order to handle corner cases and race conditions, the

Ethernet A-D per EVI route MUST NOT be used for traffic forwarding by a remote PE until it also receives the associated set of Ethernet A-D per ES routes.

Backup-path is a closely related function, but it is used in Single-Active redundancy mode. In this case a PE also advertises that it has reachability to a given EVI/ES using same combination of Ethernet A-D per EVI route and Ethernet A-D per ES route as above, but with the 'Single-Active' bit in the flags of the ESI Label Extended Community set to 1. A remote PE that receives a MAC advertisement route with non-reserved ESI SHOULD consider the advertised MAC address to be reachable via any PE that has advertised this combination of Ethernet A-D routes and it SHOULD install a backup-path for that MAC address.

#### 8.4.1 Constructing the Ethernet A-D per EVPN Instance (EVI) Route

This section describes the procedures used to construct the Ethernet A-D per EVPN Instance (EVI) route, which is used for aliasing (as discussed above). Support of this route is OPTIONAL.

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

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

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

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

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

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

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

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

#### 8.4.1.1 Ethernet A-D Route Targets

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

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

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

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

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

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

#### 8.5 Designated Forwarder Election

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

Forwarder (DF), is responsible for certain actions:

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

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

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

If a bridged network is multi-homed to more than one PE in an EVPN network via switches, then the support of All-Active redundancy mode requires the bridge network to be connected to two or more PEs using a LAG.

If a bridged network does not connect to the PEs using LAG, then only one of the links between the switched bridged network and the PEs must be the active link for a given EVPN instance. In this case, the set of Ethernet A-D per ES routes advertised by each PE MUST have the 'Single-Active' bit in the flags of the ESI Label Extended Community set to 1.

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

1. When a PE discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute.
2. The PE then starts a timer (default value = 3 seconds) to allow

the reception of Ethernet Segment routes from other PE nodes connected to the same Ethernet Segment. This timer value MUST be same across all PEs connected to the same Ethernet Segment.

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

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

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

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

#### 8.6. Interoperability with Single-homing PEs

Let's refer to PEs that only support single-homed CE devices as single-homing PEs. For single-homing PEs, all the above multi-homing



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

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

## 9. Determining Reachability to Unicast MAC Addresses

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

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

### 9.1. Local Learning

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

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

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

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

There are applications where a MAC address that is reachable via a

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

## 9.2. Remote learning

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

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

### 9.2.1. Constructing the BGP EVPN MAC/IP Address Advertisement

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

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

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

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

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

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

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

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

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

The MPLS label1 field is encoded as 3 octets, where the high-order 20 bits contain the label value. The MPLS label1 MUST be downstream assigned and it is associated with the MAC address being advertised by the advertising PE. The advertising PE uses this label when it receives an MPLS-encapsulated packet to perform forwarding based on the destination MAC address. The forwarding procedures are specified in section "Forwarding Unicast Packets" and "Load Balancing of Unicast Packets".

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

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

ID translation on egress to the CE.

The MPLS label2 field is an optional field and if it is present, then it is encoded as 3 octets, where the high-order 20 bits contain the label value. The use of MPLS label2 is for further study.

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

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

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

#### 9.2.2 Route Resolution

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

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

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

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

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

T2- If after T1, PE1 withdraws its set of Ethernet A-D per ES routes, then PE3 forwards traffic destined to M1 to PE2 only.

T3- If after T1, PE2 withdraws its set of Ethernet A-D per ES routes, then PE3 forwards traffic destined to M1 to PE1 only.

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

## 10. ARP and ND

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

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

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

### 10.1 Default Gateway

When a PE needs to perform inter-subnet forwarding where each subnet is represented by a different broadcast domain (e.g., different VLAN) the inter-subnet forwarding is performed at layer 3 and the PE that performs such function is called the default gateway. In this case

when the PE receives an ARP Request for the IP address of the default gateway, the PE originates an ARP Reply.

Each PE that acts as a default gateway for a given EVPN instance MAY advertise in the EVPN control plane its default gateway MAC address using the MAC advertisement route, and indicates that such route is associated with the default gateway. This is accomplished by requiring the route to carry the Default Gateway extended community defined in [Section 8.8 Default Gateway Extended Community]. The ESI field is set to zero when advertising the MAC route with the Default Gateway extended community.

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

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

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

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

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

## 11. Handling of Multi-Destination Traffic

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

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

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

### 11.1. Construction of the Inclusive Multicast Ethernet Tag Route

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

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

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

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

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

### 11.2. P-Tunnel Identification

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

Depending on the technology used for the P-tunnel for the EVPN instance on the PE, the PMSI Tunnel attribute of the Inclusive Multicast Ethernet Tag route is constructed as follows.

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

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

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

## 12. Processing of Unknown Unicast Packets

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

However, if a destination MAC address of a received packet is not known by the PE, the PE may have to flood the packet. When flooding, one must take into account "split horizon forwarding" as follows: The



principles behind the following procedures are borrowed from the split horizon forwarding rules in VPLS solutions [RFC4761] and [RFC4762]. When an PE capable of flooding (say PEx) receives an unknown destination MAC address, it floods the frame. If the frame arrived from an attached CE, PEx must send a copy of the frame to every other attached CE participating in that EVPN instance, on a different ESI than the one it received the frame on, as long as the PE is the DF for the egress ESI. In addition, the PE must flood the frame to all other PEs participating in that EVPN instance. If, on the other hand, the frame arrived from another PE (say PEy), PEx must send a copy of the packet only to attached CEs as long as it is the DF for the egress ESI. PEx MUST NOT send the frame to other PEs, since PEy would have already done so. Split horizon forwarding rules apply to unknown MAC addresses.

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

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

#### 12.1. Ingress Replication

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

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

#### 12.2. P2MP MPLS LSPs

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

The P-Tunnel attribute specifies the P2MP LSP identifier. This is the equivalent of an Inclusive tree in [VPLS-MCAST]. Note that multiple

Ethernet Tags, which may be in different EVPN instances, may use the same P2MP LSP, using upstream labels [VPLS-MCAST]. This is the equivalent of an Aggregate Inclusive tree in [VPLS-MCAST]. When P2MP LSPs are used for flooding unknown unicast traffic, packet re-ordering is possible.

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

### 13. Forwarding Unicast Packets

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

#### 13.1. Forwarding packets received from a CE

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

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

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

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

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

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

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

- The PE MUST drop the packet.

### 13.2. Forwarding packets received from a remote PE

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

#### 13.2.1. Unknown Unicast Forwarding

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

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

#### 13.2.2. Known Unicast Forwarding

If the top MPLS label ends up being an EVPN label that was advertised

in the unicast MAC advertisements, then the PE either forwards the packet based on CE next-hop forwarding information associated with the label or does a destination MAC address lookup to forward the packet to a CE.

#### 14. Load Balancing of Unicast Frames

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

##### 14.1. Load balancing of traffic from an PE to remote CEs

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

###### 14.1.1 Single-Active Redundancy Mode

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

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

If there is only one backup PE for a given ES, the remote PE MAY use the primary PE's withdrawal of its set of Ethernet A-D per ES routes as a trigger to update its forwarding entries, for the associated MAC addresses, to point towards the backup PE. As the backup PE starts learning the MAC addresses over its attached ES, it will start sending MAC Advertisement routes while the failed PE withdraws its routes. This mechanism minimizes the flooding of traffic during fail-over events.

If there is more than one backup PE for a given ES, the remote PE MUST use the primary PE's withdrawal of its set of Ethernet A-D per ES routes as a trigger to start flooding traffic for the associated MAC addresses (as long as flooding of unknown unicast is

administratively allowed), as it is not possible to select a single backup PE.

#### 14.1.2 All-Active Redundancy Mode

For a given ES, if the remote PE has imported the set of Ethernet A-D per ES routes from one or more PEs and none of them have the "Single-Active" flag in the ESI Label Extended Community set, then the remote PE MUST deduce that the ES is operating in All-Active redundancy mode. A remote PE that receives a MAC advertisement route with non-reserved ESI SHOULD consider the advertised MAC address to be reachable via all PEs that have advertised reachability to that MAC address' EVI/ES via the combination of an Ethernet A-D per EVI route for that EVI/ES (and Ethernet Tag if applicable) AND an Ethernet A-D per ES route for that ES. The remote PE MUST use received MAC Advertisement routes and Ethernet A-D per EVI/per ES routes to construct the set of next-hops for the advertised MAC address.

The remote PE MUST use the MAC advertisement and eligible Ethernet A-D routes to construct the set of next-hops that it can use to send the packet to the destination MAC. Each next-hop comprises an MPLS label stack that is to be used by the egress PE to forward the packet. This label stack is determined as follows:

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

-If a set of Ethernet A-D per ES routes for that ES AND an Ethernet A-D route per EVI exist, then the label from that latter route must be used.

The following example explains the above.

Consider a CE (CE1) that is dual-homed to two PEs (PE1 and PE2) on a LAG interface (ES1), and is sending packets with MAC address MAC1 on VLAN1 (mapped to EVI1). A remote PE, say PE3, is able to learn that MAC1 is reachable via PE1 and PE2. Both PE1 and PE2 may advertise MAC1 in BGP if they receive packets with MAC1 from CE1. If this is not the case, and if MAC1 is advertised only by PE1, PE3 still

considers MAC1 as reachable via both PE1 and PE2 as both PE1 and PE2 advertise a set of Ethernet A-D per ES routes for ES1 as well as an Ethernet A-D per EVI route for <EVI1, ES1>.

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

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

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

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

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

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

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

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

A CE may be configured with more than one interface connected to different PEs or the same PE for load balancing, using a technology

such as LAG. The PE(s) and the CE can load balance traffic onto these interfaces using one of the following mechanisms.

#### 14.2.1. Data plane learning

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

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

#### 14.2.2. Control plane learning

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

### 15. MAC Mobility

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

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

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

local learning is performed using the control or management planes, these interactions serve as the trigger for these PEs to withdraw their advertisements.

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

An implementation MUST handle the scenarios where the sequence number wraps around to process mobility event correctly.

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

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



homing, is to prevent false detection of MAC move among the PEs attached to the same multi-homed site.

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

#### 15.1. MAC Duplication Issue

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

#### 15.2. Sticky MAC addresses

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

#### 16. Multicast & Broadcast

The PEs in a particular EVPN instance may use ingress replication or

P2MP LSPs to send multicast traffic to other PEs.

#### 16.1. Ingress Replication

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

#### 16.2. P2MP LSPs

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

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

##### 16.2.1. Inclusive Trees

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

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

The procedures for signaling an Inclusive Tree are the same as those in [VPLS-MCAST] with the VPLS-AD route replaced with the Inclusive Multicast Ethernet Tag route. The P-Tunnel attribute [VPLS-MCAST] for an Inclusive tree is advertised in the Inclusive Multicast route as described in section "Handling of Multi-Destination Traffic". Note

that an PE can "aggregate" multiple inclusive trees for different EVPN instances on the same P2MP LSP using upstream labels. The procedures for aggregation are the same as those described in [VPLS-MCAST], with VPLS A-D routes replaced by EVPN Inclusive Multicast routes.

## 17. Convergence

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

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

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

### 17.2. PE Failures

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

### 17.3. PE to CE Network Failures

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

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

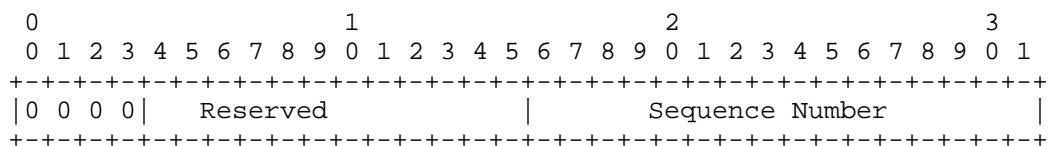
## 18. Frame Ordering

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

In order to avoid any such mis-ordering, the following rules are applied:

- If a network uses deep packet inspection for its ECMP, then the control word SHOULD be used when sending EVPN encapsulated packets over a MP2P LSP.
- If a network uses Entropy label [RFC6790], then the control word SHOULD NOT be used when sending EVPN encapsulated packet over a MP2P LSP.
- When sending EVPN encapsulated packets over a P2MP LSP or TE P2P LSP, then the control word SHOULD NOT be used.

The control word is defined as follows:



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

## 19. Acknowledgements

Special thanks to Yakov Rekhter for reviewing this draft several times and providing valuable comments and for his very engaging discussions on several topics of this draft that helped shape this document. We would also like to thank Pedro Marques, Kaushik Ghosh, Nischal Sheth, Robert Raszuk, Amit Shukla and Nadeem Mohammed for discussions that helped shape this document. We would also like to thank Han Nguyen for his comments and support of this work. We would also like to thank Steve Kensil and Reshad Rahman for their reviews. We would like to thank Jorge Rabadan for his contribution to section 5 of this draft. We like to thank Thomas Morin for his review of this draft and his contribution of section 8.6. Last but not least, many thanks to Jakob Heitz for his help to improve several sections of this draft.

## 20. Security Considerations

Security considerations discussed in [RFC4761] and [RFC4762] apply to this document for MAC learning in data-plane over an Attachment Circuit (AC) and for flooding of unknown unicast and ARP messages over the MPLS/IP core. Security considerations discussed in [RFC4364] apply to this document for MAC learning in control-plane over the MPLS/IP core. This section describes additional considerations.

As mentioned in [RFC4761], there are two aspects to achieving data privacy and protecting against denial-of-service attacks in a VPN: securing the control plane and protecting the forwarding path. Compromise of the control plane could result in a PE sending customer data belonging to some EVPN to another EVPN, or black-holing EVPN customer data, or even sending it to an eavesdropper; none of which are acceptable from a data privacy point of view. In addition, compromise of the control plane could result in black-holing EVPN customer data and could provide opportunities for unauthorized EVPN data usage (e.g., exploiting traffic replication within a multicast tree to amplify a denial-of-service attack based on sending large amounts of traffic).

The mechanisms in this document use BGP for the control plane. Hence, techniques such as in [RFC5925] help authenticate BGP messages, making it harder to spoof updates (which can be used to divert EVPN traffic to the wrong EVPN instance) or withdrawals (denial-of-service attacks). In the multi-AS methods (b) and (c), this also means protecting the inter-AS BGP sessions, between the ASBRs, the PEs, or the Route Reflectors.

Note that [RFC5925] will not help in keeping MPLS labels private -- knowing the labels, one can eavesdrop on EVPN traffic. However, this requires access to the data path within an SP network, which is assumed to be composed of trusted nodes/links.

One of the requirements for protecting the data plane is that the MPLS labels be accepted only from valid interfaces. For a PE, valid interfaces comprise links from other routers in the PE's own AS. For an ASBR, valid interfaces comprise links from other routers in the ASBR's own AS, and links from other ASBRs in ASes that have instances of a given EVPN. It is especially important in the case of multi-AS EVPN instances that one accept EVPN packets only from valid interfaces.

It is also important to help limit malicious traffic into a network for an imposter MAC address. The mechanism described in section 16.1, shows how duplicate MAC addresses can be detected and continuous false MAC mobility can be prevented. The mechanism described in section 16.2, shows how MAC addresses can be pinned to a given Ethernet Segment, such that if they appear behind any other Ethernet Segments, the traffic for those MAC addresses be prevented from entering the EVPN network from the other Ethernet Segments.

## 21. Contributors

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

Samer Salam  
Sami Boutros  
Keyur Patel  
Clarence Filsfils  
Dennis Cai  
Cisco

Ravi Shekhar  
Quaizar Vohra  
Kireeti Kompella  
Apurva Mehta  
Nadeem Mohammad  
Juniper Networks

Florin Balus  
Nuage Networks

## 22. IANA Considerations

This document defines a new NLRI, called "EVPN", to be carried in BGP using multiprotocol extensions. This NLRI uses the existing AFI of 25 (L2VPN). IANA has assigned it a SAFI value of 70.

## 23. References

### 23.1 Normative References

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

### 23.2 Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-04.txt, July 2013.
- [VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-14.txt, July 2013.
- [RT-CONSTRAIN] P. Marques et. al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.
- [RFC6790] K. Kompella et. al, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

### 24. Author's Address

Ali Sajassi  
Cisco  
Email: sajassi@cisco.com

Rahul Aggarwal  
Email: raggarwa\_1@yahoo.com

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

Aldrin Isaac  
Bloomberg  
Email: aisaac71@bloomberg.net

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

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

John Drake  
Juniper Networks  
Email: jdrake@juniper.net



NVO3  
Internet-Draft  
Intended status: Standards Track  
Expires: August 16, 2014

P. Jain  
K. Singh  
F. Balus  
Nuage Networks  
W. Henderickx  
Alcatel-Lucent  
V. Bannai  
PayPal  
R. Shekhar  
A. Lohiya  
Juniper Networks  
February 12, 2014

Generic Overlay OAM and Datapath Failure Detection  
draft-jain-nvo3-overlay-oam-01

Abstract

This proposal describes a mechanism that can be used to detect Data Path Failures of various overlay technologies as VXLAN, NVGRE, MPLSoGRE and MPLSoUDP and verifying/sanity of their Control and Data Plane for given Overlay Segment. This document defines the following for each of the above Overlay Technologies:

- o Encapsulation of OAM Packet, such that it has same Outer and Overlay Header as any End-System's data going over the same Overlay Segment.
- o The mechanism to trace the Underlay that is exercised by any Overlay Segment.
- o Procedure to verify presence of any given Tenant VM or End-System within a given Overlay Segment at Overlay End-Point.

Even though the present proposal addresses Overlay OAM for VXLAN, NVGRE, MPLSoGRE and MPLSoUDP, but the procedures described are generic enough to accommodate OAM for any other Overlay Technology.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 16, 2014.

#### 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 . . . . .	6
2. Terminology . . . . .	8
3. Motivation for Overlay OAM . . . . .	9
4. Approach . . . . .	10
5. Packet Format . . . . .	11
5.1. Overlay OAM Encapsulation in Layer 2 Context . . . . .	11
5.2. Overlay OAM Encapsulation in Layer 3 Context . . . . .	11
5.3. Generic Overlay OAM Packet Format . . . . .	11
5.3.1. TLV Types for various Overlay Ping Models . . . . .	14
5.3.1.1. TLV for VXLAN Ping . . . . .	15
5.3.1.2. TLV for NVGRE Ping . . . . .	16
5.3.1.3. TLV for MPLSoGRE Ping . . . . .	17
5.3.1.4. TLV for MPLSoUDP Ping . . . . .	18
6. Return Codes . . . . .	19
7. Procedure for Overlay Segment Ping . . . . .	20
7.1. Encoding of Inner Header for Echo Request in Layer 2 Context . . . . .	20
7.2. Encoding of Inner Header for Echo Request in Layer 3 Context . . . . .	21
7.3. VXLAN Procedures . . . . .	21
7.3.1. Sending VXLAN Echo Request . . . . .	21
7.3.2. Receiving VXLAN Echo Request . . . . .	22
7.3.3. Sending VXLAN Echo Reply . . . . .	23
7.3.4. Receiving VXLAN Echo Reply . . . . .	23
7.4. NVGRE Procedures . . . . .	23
7.4.1. Sending NVGRE Echo Request . . . . .	23
7.4.2. Receiving NVGRE Echo Request . . . . .	24
7.4.3. Sending NVGRE Echo Reply . . . . .	25
7.4.4. Receiving NVGRE Echo Reply . . . . .	25
7.5. MPLSoGRE Procedures . . . . .	25
7.5.1. Sending MPLSoGRE Echo Request . . . . .	25
7.5.2. Receiving MPLSoGRE Echo Request . . . . .	25
7.5.3. Sending MPLSoGRE Echo Reply . . . . .	26
7.5.4. Receiving MPLSoGRE Echo Reply . . . . .	26
7.6. MPLSoUDP Procedures . . . . .	26
7.6.1. Sending MPLSoUDP Echo Request . . . . .	26
7.6.2. Receiving MPLSoUDP Echo Request . . . . .	27
7.6.3. Sending MPLSoUDP Echo Reply . . . . .	28
7.6.4. Receiving MPLSoUDP Echo Reply . . . . .	28

8. Procedure for Trace . . . . .	29
9. Procedure for End-System Ping . . . . .	30
9.1. Sub-TLV for End-System Ping . . . . .	30
9.1.1. Sub-TLV for Validating End-System MAC Address . . . . .	31
9.1.2. Sub-TLV for Validating End-System IP Address . . . . .	32
9.1.3. Sub-TLV for Validating End-System MAC and IP Address . . . . .	33
9.2. Sending End-System Ping Request . . . . .	34
9.3. Receiving End-System Ping Request . . . . .	34
9.4. Sending End-System Ping Reply . . . . .	35
9.5. Receiving End-System Ping Reply . . . . .	35
10. Security Considerations . . . . .	37
11. Management Considerations . . . . .	38
12. Acknowledgements . . . . .	39
13. IANA Considerations . . . . .	40
14. References . . . . .	41
14.1. Normative References . . . . .	41
14.2. Informative References . . . . .	42
Authors' Addresses . . . . .	43

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 [RFC2119].

When used in lower case, these words convey their typical use in common language, and are not to be interpreted as described in RFC2119 [RFC2119].

## 1. Introduction

VXLAN [I-D.draft-mahalingam-dutt-dcops-vxlan], NVGRE [I-D.draft-sridharan-virtualization-nvgre], MPLSoGRE [RFC4023] and MPLSoUDP [I-D.draft-ietf-mpls-in-udp] are well known technologies and are used as tunneling mechanism to Overlay either Layer 2 networks or Layer 3 networks on top of Layer 3 Underlay networks. For all above Overlay Models there are two Tunnel End Points for a given Overlay Segment. One End Point is where the Overlay Originates, and other where Overlay Terminates. In most cases the Tunnel End Point is intended to be at the edge of the network, typically connecting an access switch to an IP transport network. The access switch could be a physical or a virtual switch located within the hypervisor on the server which is connected to End System which is a VM.

This document describes a mechanism that can be used to detect Data Plane failures and sanity of Overlay Control and Data Plane for a given Overlay Segment, and the method to trace the Underlay path that is exercised by any given Overlay Segment.

The document also defines procedures for validating the presence of any given Tenant VM/End-System/End-System or Flow representing the End-System System within a given Overlay Segment.

The proposal describes:

- o The mechanism to verify Overlay Control Plane and Data Plane consistency at the Overlay End Point(s), by encapsulating the OAM Packet in exact the same way as that of any End System Traffic that is transported over the Overlay Segment.
- o The mechanism to trace the Underlay that is exercised by any Overlay Segment.
- o The mechanism to verify presence of any "End-System" in a given Overlay Segment.

The proposal defines the information to check correct operation of the Data Plane, as well as a mechanism to verify the Data Plane against the Control Plane for a given Overlay Segment.

It is important consideration in this proposal to carry Echo Request along same Data Path that any End System's data using the given Overlay Segment takes.

The tenants VM(s) or End System(s) are not aware of the Overlays and as such the need for the verification of the Data Path MUST solely rest with the Cloud Provider. The use cases where the Tenant VM(s)

need to be aware of the Data Plane failures is beyond the scope of this document.

## 2. Terminology

Terminology used in this document:

OAM: Operations, Administration, and Management

VXLAN: Virtual eXtensible Local Area Network.

NVGRE: Network Virtualization using GRE.

MPLSoGRE: Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)

MPLSoUDP: Encapsulating MPLS in UDP.

Originating End Point: Overlay Segment's Head End or Starting Point of Overlay Tunnel.

Terminating End Point: Overlay Segment's Tail End or Terminating Point of Overlay Tunnel.

VM: Virtual Machine.

VNI: VXLAN Network Identifier (or VXLAN Segment ID)

VSID: Virtual Subnet ID. (for NVGRE)

NVE: Network Virtualized Edge

End System: Could be Tenant VM, Host, Bridge etc. - System whose data is expected to go over Overlay Segment.

Echo Request: Throughout this document, Echo Request packet is expected to be transmitted by Originator Overlay End Point and destined to Overlay Terminating End Point.

Echo Reply: Throughout this document, Echo Reply packet is expected to be transmitted by Terminating Overlay End Point and destined to Overlay Originating End Point.

Other terminologies are as defined in  
[I-D.draft-mahalingam-dutt-dcops-vxlan],  
[I-D.draft-sridharan-virtualization-nvgre], [RFC4023] and  
[I-D.draft-ietf-mpls-in-udp]



### 3. Motivation for Overlay OAM

When any Overlay Segment fails to deliver user traffic, there is a need to provide a tool that would enable users, as Cloud Providers to detect such failures, and a mechanism to isolate faults. It may also be desirable to test the data path before mapping End System traffic to the Overlay Segment.

The basic idea is to facilitate following verifications:-

- o End-System's data that are expected to go over a particular Overlay Segment actually ends up using the Data-Path represented by given Overlay Segment between the two End-Points.
- o To verify the correct value of Overlay Segment Identifier is programmed at Originating and Terminating End Point(s) for a given Overlay Segment. Segment Identifier will be VNI for VXLAN, VSID for NVGRE, MPLS Label for MPLSoGRE and MPLSoUDP.
- o The facilitate mechanism to trace the Underlay that is exercised by any Overlay Segment.
- o The mechanism to verify presence of any "End-System" in a given Overlay Segment.

To facilitate verification of Overlay Segment or any End-System using the Overlay, this document proposes sending of a Packet (called an "Echo Request") along the same data path as other Packets belonging to this Segment. Echo Request also carries information about the Overlay Segment whose Data Path is to be verified. This Echo Request is forwarded just like any other End System Data Packet belonging to that Overlay Segment, as it contains the same Overlay Encapsulation as regular End System's data.

On receiving Echo Request at the end of the Overlay Segment, it is sent to the Control Plane of the Terminating Overlay End Point, which in-turn would respond with Echo Reply.

To facilitate tracing of the Underlay used by any given Overlay Segment, the document proposes Echo Request/Reply encapsulation in "trace mode", which would allow the user or Cloud Provider to gather information of the Underlay network.

#### 4. Approach

The proposal aims at validating Data Plane and its view of Control Plane for a particular Overlay Segment. To achieve this aim, the draft proposes creating an Overlay OAM Packet which MUST be encapsulated with the Overlay Header as that of any End-Point data going over the same Overlay Segment. This would guarantee the data-path for OAM Packet follows the same path as that for any End User data going over the same Overlay Segment.

The draft outlines procedures to encode Overlay Header and Inner Ethernet or IP Header based on the type of payload that Overlay is expected to carry.

## 5. Packet Format

Generic Overlay Echo Request/Reply is a UDP Packet identified by well known UDP Port XXXX. The payload carried by Overlay typically could be either be Layer 2 / Ethernet Frame, or it could be Layer 3 / IP Packet.

### 5.1. Overlay OAM Encapsulation in Layer 2 Context

If the encapsulated payload carried by Overlay is of type Ethernet, then the OAM Echo Request packet would have inner Ethernet Header, followed by IP and UDP Header. The payload of inner UDP would be as described in below section "Generic Overlay OAM Packet Format".

### 5.2. Overlay OAM Encapsulation in Layer 3 Context

If the encapsulated payload carried by Overlay is of type IP, then the OAM Echo Request packet would have inner IP Header, followed by UDP Header. The payload of inner UDP would be as described in below section "Generic Overlay OAM Packet Format".

### 5.3. Generic Overlay OAM Packet Format

Following is the format of UDP payload of Generic Overlay OAM Packet:

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Message Type | Reply mode | Return Code | Return Subcode |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Originator Handle                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Sequence Number                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               TimeStamp Sent (seconds)                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               TimeStamp Sent (microseconds)                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               TimeStamp Received (seconds)                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               TimeStamp Received (microseconds)                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               TLVs ...                               |
.                                                                           .
.                                                                           .
.                                                                           .
+-----+-----+-----+-----+-----+-----+-----+-----+

```

## Generic Overlay OAM Packet

The Message Type is one of the following:-

Value	What it means
1	Echo Request
2	Echo Reply

Reply Mode Values:-

Value	What it means
1	Do not reply
2	Reply via an IPv4/IPv6 UDP Packet
3	Reply via Overlay Segment

Echo Request with 1 (Do not reply) in the Reply Mode field may be used for one-way connectivity tests. The receiving node may log gaps in the Sequence Numbers and/or maintain delay/jitter statistics. For normal operation Echo Request would have 2 (Reply via an IPv4 UDP

Packet) in the Reply Mode field.

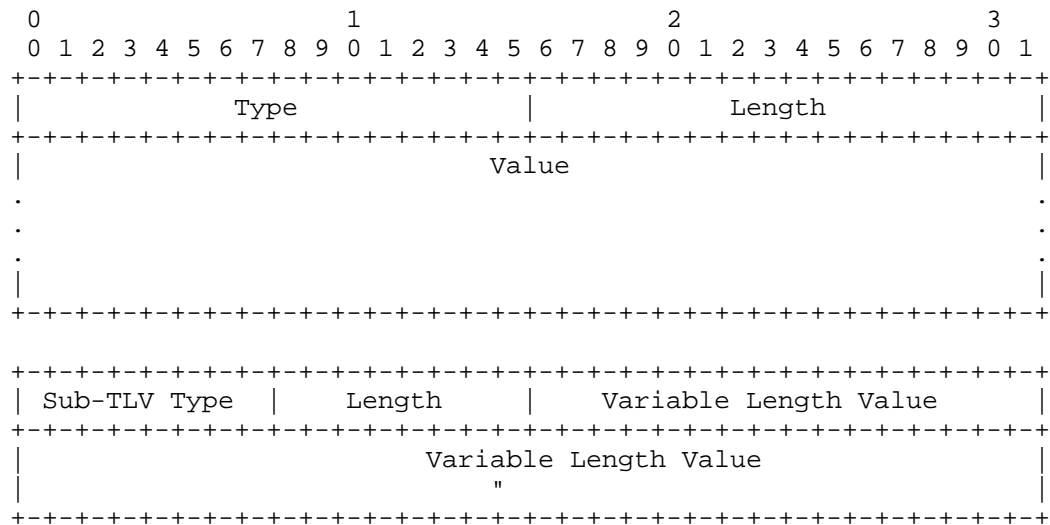
If it is desired that the reply also comes back via Overlay Segment i.e. encapsulated with the Overlay Header, then the Reply Mode field needs to be set to 3 (Reply via Overlay Segment).

The Originator's Handle is filled in by the Originator, and returned unchanged by the receiver in the Echo Reply (if any). The value used for this field can be implementation dependent, this MAY be used by the Originator for matching up requests with replies.

The Sequence Number is assigned by the Originator of Echo Request and can be (for example) used to detect missed replies.

The TimeStamp Sent is the time-of-day (in seconds and microseconds, according to the sender's clock) in NTP format [NTP] when the VXLAN Echo Request is sent. The TimeStamp Received in an Echo Reply is the time-of-day (according to the receiver's clock) in NTP format that the corresponding Echo Request was received.

TLVs (Type-Length-Value tuples) have the following format:



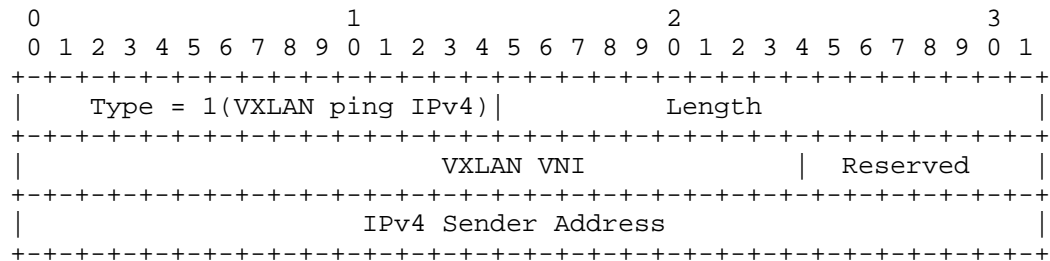
Types are defined below; Length is the length of the Value field in octets. The Value field depends on the Type; it is zero padded to align to a 4-octet boundary. There could be one or many optional Sub-TLV that could be encoded under the TLV.

## 5.3.1. TLV Types for various Overlay Ping Models

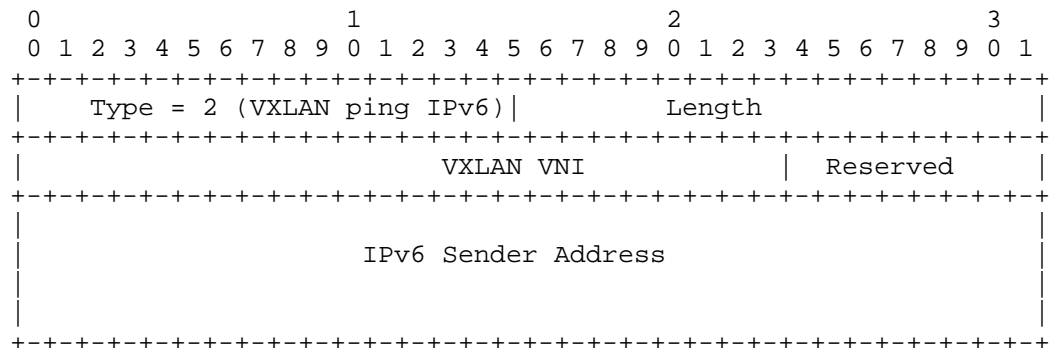
## TLV Types:-

Value	What it means
1	VXLAN Segment Ping for IPv4
2	VXLAN Segment Ping for IPv6
3	NVGRE Segment Ping for IPv4
4	NVGRE Segment Ping for IPv6
5	MPLSoGRE Segment Ping for IPv4
6	MPLSoGRE Segment Ping for IPv6
7	MPLSoUDP Segment Ping for IPv4
8	MPLSoUDP Segment Ping for IPv6

## 5.3.1.1. TLV for VXLAN Ping

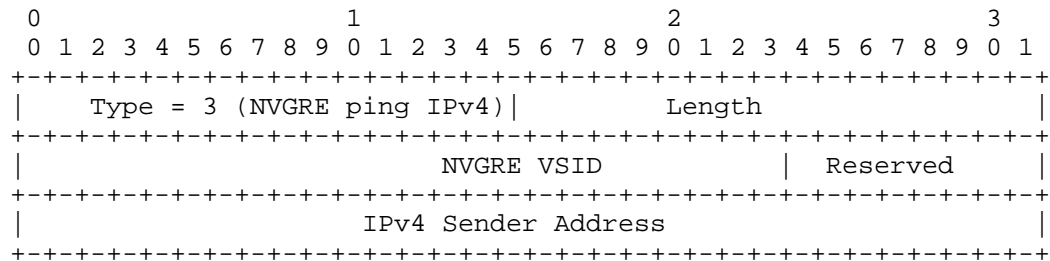


TLV if Sender Address is IPv4

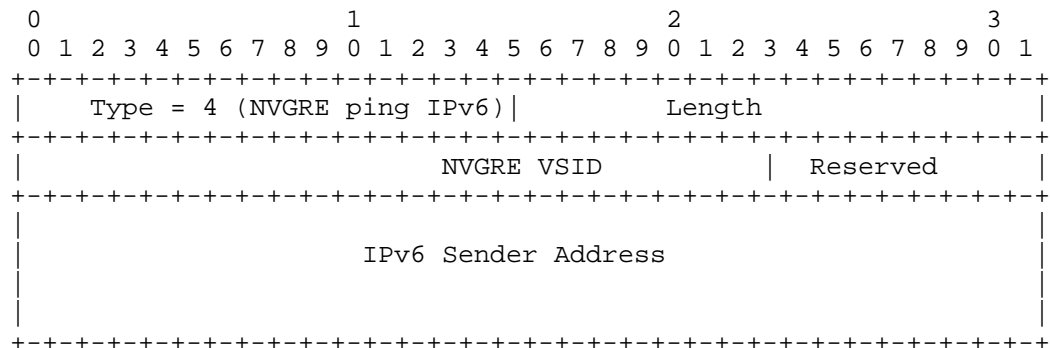


TLV if Sender Address is IPv6

## 5.3.1.2. TLV for NVGRE Ping



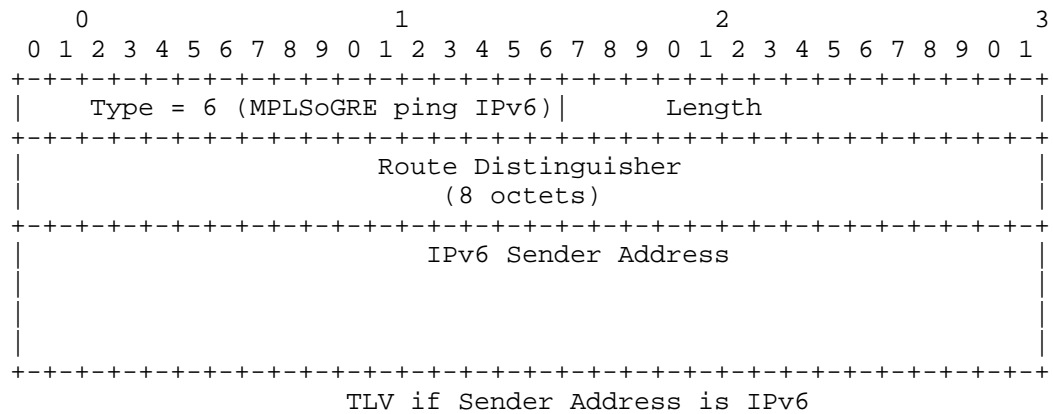
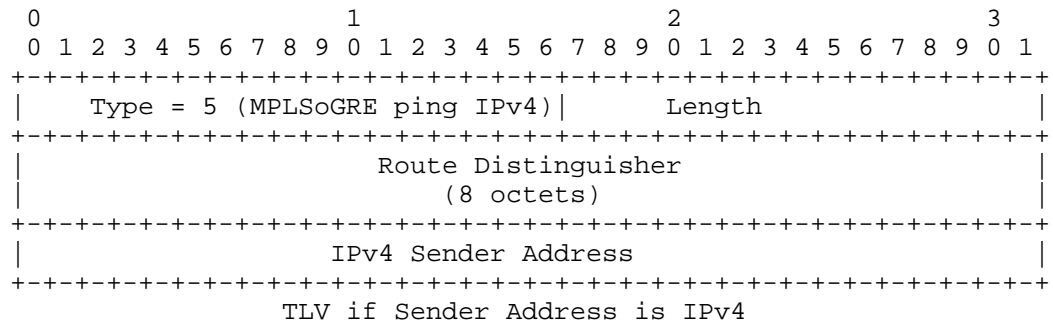
TLV if Sender Address is IPv4



TLV if Sender Address is IPv6

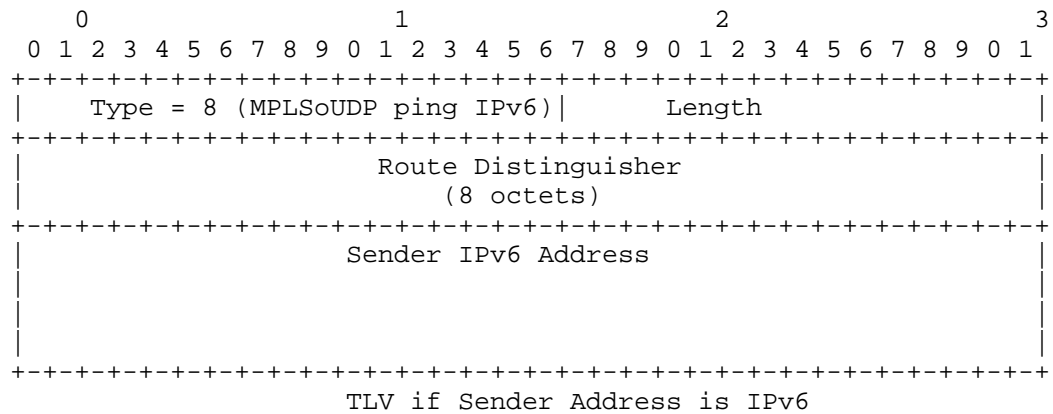
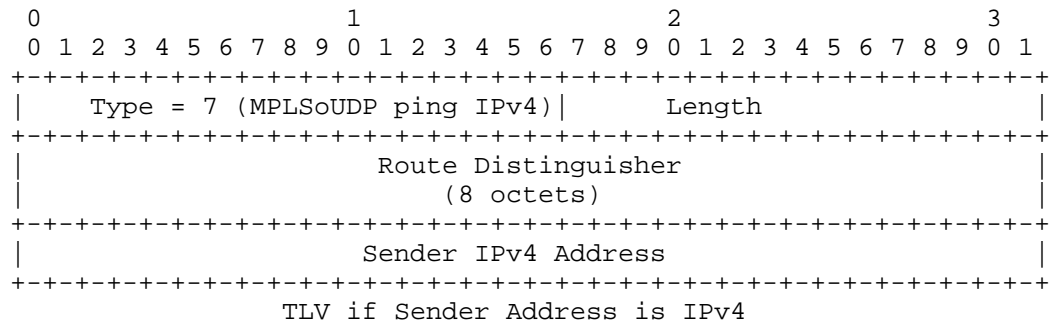


## 5.3.1.3. TLV for MPLSoGRE Ping



Route Distinguisher is defined as part of [RFC4365]

## 5.3.1.4. TLV for MPLSoUDP Ping



Route Distinguisher is defined as part of [RFC4365]

## 6. Return Codes

Sender MUST always set the Return Code set to zero. The receiver can set it to one of the values listed below when replying back to Echo-Request.

Following are the Return Codes (Suggested):-

Value	What it means
-----	-----
0	No return code
1	Malformed Echo Request Received
2	Overlay Segment Not Present
3	Overlay Segment Not Operational
4	Return-Code-OK

## 7. Procedure for Overlay Segment Ping

Echo Request is used to test Data Plane and its view of Control Plane for particular Overlay Segment. The Overlay Segment to be verified is identified differently for various Overlay Technologies. For VXLAN, VNI is used to identify given Overlay Segment. For NVGRE, VSID is used. For MPLSoGRE and MPLSoUDP the MPLS Stack is used to identify a given Overlay Segment.

For the Data Plane verification, the Overlay Echo Request Packet MUST be encapsulated within the Overlay Header, which is same as that of any End-Point data going over the same Overlay Segment. This would guarantee the data-path for OAM Packet follows the same path as that for any End User data going over the same Overlay Segment.

The payload carried by Overlay typically could be either be Layer 2 or Ethernet Frame, or it could be Layer 3 or IP Packet. Based on the type of payload following is the way inner Header(s) of Echo Request would be encoded.

### 7.1. Encoding of Inner Header for Echo Request in Layer 2 Context

If the encapsulated payload carried by Overlay is of type Ethernet, then the OAM Echo Request packet would have inner Ethernet Header, followed by IP and UDP Header. The payload of inner UDP would be as described in below section "Generic Overlay OAM Packet Format".

Inner Ethernet Header for the Echo Request Packet MUST have the Destination Mac set to 00-00-5E-90-XX-XX (to be assigned IANA). The Source Mac should be set to Mac Address of the Originating VTEP. However, it is desired that the Inner Source Mac SHOULD not be learnt in the MAC-Table as this represent Control Packet in context of Overlay OAM.

Inner IP header is set with the Source IP Address which is a routable Address of the sender; the Destination IP Address is a (randomly chosen) IPv4 Address from the range 127/8, IPv6 addresses are chosen from the range 0:0:0:0:0:FFFF:127/104. The IP TTL is set to 255.

The inner Destination UDP port is set to xxxx (assigned by IANA for Overlay OAM).

The "Generic Overlay OAM Packet" will now be encoded, with following information.

The sender chooses a Originator's Handle and a Sequence Number. When sending subsequent Overlay Echo Requests, the sender SHOULD increment the Sequence Number by 1.

The TimeStamp Sent is set to the time-of-day (in seconds and microseconds) that the Echo Request is sent. The TimeStamp Received is set to zero. Also, the Reply Mode must be set to the desired reply mode. The Return Code and Subcode are set to zero.

Next, the TLV is Encoded for desired Overlay Type, as per Section "Types of TLVs defined for various Overlay Ping Models"

## 7.2. Encoding of Inner Header for Echo Request in Layer 3 Context

If the encapsulated payload carried by Overlay is of type IP, then the Encoding of the Echo Request would be same as above Section "Encoding of Inner Header for Echo Request in Layer 2 Context", but without the presence of Inner Ethernet Header.

## 7.3. VXLAN Procedures

### 7.3.1. Sending VXLAN Echo Request

The Outer VxLAN header for the Echo Request packet follows the encapsulation as defined in [I-D.draft-mahalingam-dutt-dcops-vxlan]. The VNI is same as that of the VXLAN Segment that is being verified. This would make sure that OAM Packet takes the same datapath as any other End System data going over this VXLAN Segment.

The VXLAN Router Alert option

[I-D.draft-singh-nvo3-vxlan-router-alert] MUST be set in the VXLAN header as shown below.

VXLAN Header:

```

+++++
|R|R|R|R|I|R|R|RA|          Reserved          |
+++++
|          VXLAN Network Identifier (VNI) |   Reserved   |
+++++

```

RA: Router Alter Bit (Proposed)

Originating VTEP MAY set the I Bit to 0 in VXLAN Header when sending OAM Frame. This would cause dropping of such VXLAN frames on any Terminating VTEP that does not understand Overlay OAM framework, and prevent sending those frames to End-Systems or VMs.

It is desired to choose the Source UDP port (in the outer header), so as to exercise the same Data-Path as that of the traffic carried over the VXLAN Segment and is left to the implementation.

The Encoding of Inner Header(s) and UDP payload of Generic Overlay OAM Packet is as described in above Sub-Section i.e. "Encoding of Inner Header for Echo Request in Layer 2/Layer 3 Context".

### 7.3.2. Receiving VXLAN Echo Request

At the Terminating Overlay End Point or VTEP, since the Overlay OAM Packet is exactly same as that of End-System Packet(s). It is important to send OAM packet to Control Plane and prevent it from sending to the End System. The trapping and sending VXLAN Echo Request to the Control Plane is triggered by one of the following Packet processing exceptions: VXLAN Router Alert option, [I-D.draft-singh-nvo3-vxlan-router-alert] the Inner Destination MAC Address of 00-00-5E-90-XX-XX as defined in above section, and the Destination IP Address in the 127/8 Address range for IPv4 Address, or 0:0:0:0:0:FFFF:127/104 for IPv6 Address.

The Control Plane further identifies the Overlay OAM Application by UDP well know destination port xxxx.

Since the VxLAN Router Alert bit is set in VxLAN Header, which signifies the presence of Control Packet. The terminating VTEP SHOULD not learn the Mac address set in the Inner Mac Header of VxLAN Echo Request Packet.

Once the VXLAN Echo Request Packet is identified at Control Plane, it is processed as follows:-

- o General Packet sanity is verified. If the Packet is not well-formed, VTEP SHOULD send VXLAN Echo Reply with the Return Code set to "Malformed Echo Request received" and the Subcode to zero. The header fields Originator's Handle, Sequence Number, and Timestamp Sent are not examined, but are included in the VXLAN Echo Reply message
- o VNI Validation: If there is no entry for VNI, it indicates that there could be a transient or permanent disconnect between Control Plane and data Plane and VTEP needs to report an error with Return Code of "Overlay Segment Not Present" and a Return Subcode of Zero. If the mapping for VNI Exists, but the state is not Operational, VTEP needs to report an error with Return Code of "Overlay Segment Not Operational" If the mapping exists then send VXLAN Echo Reply with a Return Code of "Return-Code-OK", and a Return Subcode of Zero. The procedures for sending the Echo Reply are found in subsection below section.

### 7.3.3. Sending VXLAN Echo Reply

If the Reply Mode is set to "Reply via an IPv4/IPv6 UDP Packet", the Echo Reply is a UDP Packet. It MUST ONLY be sent in response to Echo Request. The Source IP Address in the Header should be Routable Address of the replier; The Destination IP Address should be IP Address of the Echo Request's Originating End Point or the requester. The destination UDP Port is set to XXXX (assigned by IANA for identifying VXLAN OAM application). The IP TTL is set to 255.

The format of the Echo Reply is the same as the Echo Request. The Originator Handle, the Sequence Number, and TimeStamp Sent are copied from the Echo Request; the TimeStamp Received is set to the time-of-day that the Echo Request is received (note that this information is most useful if the time-of-day clocks on the requester and the replier are synchronized). The replier MUST fill in the Return Code and Subcode, as determined in the previous subsection.

If the Reply Mode is set to "Reply via Overlay Segment", then the Replying Overlay End Point is expected to place Echo Reply packet in-band in the Overlay Segment destined to the Originating Overlay End Point. The detailed encapsulation for this would be covered in next revision of the draft.

### 7.3.4. Receiving VXLAN Echo Reply

An Originating Overlay End Point should only receive Echo Reply in response to an Echo Request that it sent. When the Reply Mode is "Reply via an IPv4/IPv6 UDP Packet", the Echo Reply would be an IP Packet/UDP Packet, and is identified by the destination UDP Port XXXX. The Originating Overlay End Point should parse the Packet to ensure that it is well-formed, then attempt to match up the Echo Reply with an Echo Request that it had previously sent, and the Originator Handle. If no match is found, then it should drop the Echo Reply Packet; otherwise, it checks the Sequence Number to see if it matches.

## 7.4. NVGRE Procedures

### 7.4.1. Sending NVGRE Echo Request

The Outer NVGRE header for the Echo Request packet follows the encapsulation as defined in [I-D.draft-sridharan-virtualization-nvgre]. The VSID is same as that of the NVGRE Segment that is being verified. This would make sure that OAM Packet takes the same datapath as any other End System data going over this NVGRE Segment.

The NVGRE Router Alert option

[I-D.draft-singh-nvo3-nvgre-router-alert] MUST be set in the NVGRE header as shown below.

GRE Header:

```

+++++
|0| |1|0| Reserved0      RA| Ver |   Protocol Type 0x6558   |
+++++
|                               Virtual Subnet ID (VSID)       | Reserved |
+++++

```

RA: Router Alter Bit (Proposed)

The Encoding of Inner Header(s) and UDP payload of Generic Overlay OAM Packet is as described in above Sub-Section i.e. "Encoding of Inner Header for Echo Request in Layer 2/Layer 3 Context".

#### 7.4.2. Receiving NVGRE Echo Request

At the Terminating Overlay End Point, since the Overlay OAM Packet is exactly same as that of End-System Packet(s). It is important to send OAM packet to Control Plane and prevent it from sending to the End System. The trapping and sending NVGRE Echo Request to the Control Plane is triggered by one of the following Packet processing exceptions: NVGRE Router Alert option, [I-D.draft-singh-nvo3-nvgre-router-alert] the Inner Destination MAC Address of 00-00-5E-90-XX-XX as defined in above section, and the Destination IP Address in the 127/8 Address range for IPv4 Address, or 0:0:0:0:0:FFFF:127/104 for IPv6 Address.

The Control Plane further identifies the Overlay OAM Application by UDP well know destination port xxxx.

Since the NVGRE Router Alert bit is set in NVGRE Header, which signifies the presence of Control Packet. The Terminating Overlay End Point SHOULD not learn the Mac address set in the Inner Mac Header of NVGRE Echo Request Packet.

Once the NVGRE Echo Request Packet is identified at Control Plane, it is processed as follows:-

- o General Packet sanity is verified. If the Packet is not well-formed, NVGRE End Point SHOULD send NVGRE Echo Reply with the Return Code set to "Malformed Echo Request received" and the Subcode to zero. The header fields Originator's Handle, Sequence Number, and Timestamp Sent are not examined, but are included in the NVGRE Echo Reply message



- o VSID Validation: If there is no entry for VSID, it indicates that there could be a transient or permanent disconnect between Control Plane and data Plane and NVGRE End Point needs to report an error with Return Code of "Overlay Segment Not Present" and a Return Subcode of Zero. If the mapping for VSID Exists, but the state is not Operational, NVGRE End Point needs to report an error with Return Code of "Overlay Segment Not Operational" If the mapping exists then send NVGRE Echo Reply with a Return Code of "Return-Code-OK", and a Return Subcode of Zero. The procedures for sending the Echo Reply are found in subsection below section.

#### 7.4.3. Sending NVGRE Echo Reply

The procedure for sending NVGRE Echo Reply are exactly same as defined in above section "Sending VXLAN Echo Reply".

#### 7.4.4. Receiving NVGRE Echo Reply

The procedure for Receiving NVGRE Echo Reply are exactly same as defined in above section "Receiving VXLAN Echo Reply".

### 7.5. MPLSoGRE Procedures

#### 7.5.1. Sending MPLSoGRE Echo Request

The Outer header of MPLSoGRE for the Echo Request packet follows the encapsulation as defined in [RFC4023]. The MPLS Stack is same as that of the MPLSoGRE Segment that is being verified. This would make sure that OAM Packet takes the same datapath as any other End System data going over this MPLSoGRE Segment.

However, the bottommost Label in MPLS Stack MUST be MPLS Router Alert Label [RFC3032]. This would indicate the Overlay Terminating End Point that the payload is a Control Packet and needs to be delivered to Control Plane.

The Encoding of Inner Header(s) and UDP payload of Generic Overlay OAM Packet is as described in above Sub-Section i.e. "Encoding of Inner Header for Echo Request in Layer 2/Layer 3 Context".

#### 7.5.2. Receiving MPLSoGRE Echo Request

At the Terminating Overlay End Point, since the Overlay OAM Packet is exactly same as that of End-System Packet(s). It is important to send OAM packet to Control Plane and prevent it from sending to the End System. The trapping and sending MPLSoGRE Echo Request to the Control Plane is triggered by one of the following Packet processing exceptions: MPLS Router Alert Label, and the Destination IP Address

in the 127/8 Address range for IPv4 Address, or 0:0:0:0:0:FFFF:127/104 for IPv6 Address.

The Control Plane further identifies the Overlay OAM Application by UDP well know destination port xxxx.

Once the MPLSoGRE Echo Request Packet is identified at Control Plane, it is processed as follows:-

- o General Packet sanity is verified. If the Packet is not well-formed, MPLSoGRE End Point SHOULD send MPLSoGRE Echo Reply with the Return Code set to "Malformed Echo Request received" and the Subcode to zero. The header fields Originator's Handle, Sequence Number, and Timestamp Sent are not examined, but are included in the MPLSoGRE Echo Reply message
- o Segment Validation: If there is no entry for service represented by given Route Distinguisher for the MPLSoGRE Segment, it indicates that there could be a transient or permanent disconnect between Control Plane and Data Plane and MPLSoGRE End Point needs to report an error with Return Code of "Overlay Segment Not Present" and a Return Subcode of Zero. If the entry for service represented by given Route Distinguisher for the MPLSoGRE Segment is present, but is Operationally Down. The End Point needs to report an error with Return Code of "Overlay Segment Not Operational" If the mapping of service represented by given Route Distinguisher for the MPLSoGRE Segment is present and Active, then send MPLSoGRE Echo Reply with a Return Code of "Return-Code-OK".

#### 7.5.3. Sending MPLSoGRE Echo Reply

The procedure for sending MPLSoGRE Echo Reply are exactly same as defined in above section "Sending VXLAN Echo Reply".

#### 7.5.4. Receiving MPLSoGRE Echo Reply

The procedure for Receiving MPLSoGRE Echo Reply are exactly same as defined in above section "Receiving VXLAN Echo Reply".

### 7.6. MPLSoUDP Procedures

#### 7.6.1. Sending MPLSoUDP Echo Request

The Outer header of MPLSoUDP for the Echo Request packet follows the encapsulation as defined in [I-D.draft-ietf-mpls-in-udp]. The MPLS Stack is same as that of the MPLSoUDP Segment that is being verified. This would make sure that OAM Packet takes the same datapath as any other End System data going over this MPLSoUDP Segment.

However, the bottommost Label in MPLS Stack MUST be MPLS Router Alert Label [RFC3032]. This would indicate the Overlay Terminating End Point that the payload is a Control Packet and needs to be delivered to Control Plane.

It is desired to choose the Source UDP port (in the outer header), so as to exercise the same Data-Path as that of the traffic carried over the MPLSoUDP Segment and is left to the implementation.

The Encoding of Inner Header(s) and UDP payload of Generic Overlay OAM Packet is as described in above Sub-Section i.e. "Encoding of Inner Header for Echo Request in Layer 2/Layer 3 Context".

#### 7.6.2. Receiving MPLSoUDP Echo Request

At the Terminating Overlay End Point, since the Overlay OAM Packet is exactly same as that of End-System Packet(s). It is important to send OAM packet to Control Plane and prevent it from sending to the End System. The trapping and sending MPLSoGRE Echo Request to the Control Plane is triggered by one of the following Packet processing exceptions: MPLS Router Alert Label, and the Destination IP Address in the 127/8 Address range for IPv4 Address, or 0:0:0:0:0:FFFF:127/104 for IPv6 Address.

The Control Plane further identifies the Overlay OAM Application by UDP well know destination port xxxx.

Once the MPLSoUDP Echo Request Packet is identified at Control Plane, it is processed as follows:-

- o General Packet sanity is verified. If the Packet is not well-formed, MPLSoUDP End Point SHOULD send MPLSoUDP Echo Reply with the Return Code set to "Malformed Echo Request received" and the Subcode to zero. The header fields Originator's Handle, Sequence Number, and Timestamp Sent are not examined, but are included in the MPLSoUDP Echo Reply message
- o Segment Validation: If there is no entry for service represented by given Route Distinguisher for the MPLSoUDP Segment, it indicates that there could be a transient or permanent disconnect between Control Plane and data Plane and MPLSoUDP End Point needs to report an error with Return Code of "Overlay Segment Not Present" and a Return Subcode of Zero. If the entry for service represented by given Route Distinguisher for the MPLSoUDP Segment is present, but is Operationally Down. The End Point needs to report an error with Return Code of "Overlay Segment Not Operational" If the mapping of service represented by given Route Distinguisher for the MPLSoUDP Segment is present and Active, then

send MPLSoUDP Echo Reply with a Return Code of "Return-Code-OK".

#### 7.6.3. Sending MPLSoUDP Echo Reply

The procedure for sending MPLSoGRE Echo Reply are exactly same as defined in above section "Sending VXLAN Echo Reply".

#### 7.6.4. Receiving MPLSoUDP Echo Reply

The procedure for Receiving MPLSoGRE Echo Reply are exactly same as defined in above section "Receiving VXLAN Echo Reply".

## 8. Procedure for Trace

In order to be able to trace the Path that a particular flow in the Overlay takes through the Underlay Network, following mechanism can be used - An overlay Echo Request packet is built and sent using the mechanisms described in the Section "Procedure for Overlay Segment Ping" so that the overlay traceroute follows the same path as the data packet for the overlay segment being traced.

The Echo Request packet in the traceroute mode is sent with the initial TTL set to 1 in the Outer IP header and thereafter incremented by 1 in each successive request. At each transit hop where the TTL expires, an exception is created. Because of this exception, the packet gets delivered to the Control Plane. Control plane can further deliver the packet to the OAM application based on the TTL exception and the specific UDP port XXXX in the incoming overlay echo request packet. If the transit node has the IP reachability to the destination IP address in the outer IP header, it sends back an overlay echo reply response otherwise the Overlay Echo Request is discarded by the Overlay OAM module on the transit nodes. If the transit node does not support overlay OAM functionality, it will simply generate a regular ICMP TTL exceeded response. This could result into "false negatives". The originating Overlay node that generated the OAM echo request SHOULD try sending the echo request with TTL=n+1, n+2, ... to probe the nodes further down the path to the terminating overlay End-point.

At the originating node, when the Echo Reply from the transit node corresponding to the traceroute query is received, it can correlate the incoming Echo Reply with the traceroute query by matching on the sequence numbers in the Overlay Echo Request/Reply packets.

Current revision of this draft limits overlay traceroute capability to fault isolation only. A subsequent version of the draft will include mechanisms to trace all possible paths in the underlay that can be used to carry overlay tunnel traffic.

## 9. Procedure for End-System Ping

In typical Overlay deployment scenarios there is a desired to check the presence of any given Tenant VM/End-System or Flow representing the End-System System within a given Overlay Segment. This draft proposes the way to achieve it via End-System Ping.

The End-System can be identified at Overlay End Point by either its IP Address, Ethernet MAC Address or combination of IP/MAC Address.

In that case, it would be important to verify the End-System connectivity by procedure which goes over the Overlay Segment from Originating Overlay End-Point and verifies the presence of the End-System at the Terminating Overlay End-Point.

The scope of End-System Ping is solely with the Cloud Provider which owns control of the Overlay End Point(s). It is expected that the Overlay End Point traps this request and checks the Presence of the End-System via its MAC Address, Route or Flow information and replies back. There SHOULD not be a case where the End-System Ping is delivered to the actual End-Point.

### 9.1. Sub-TLV for End-System Ping

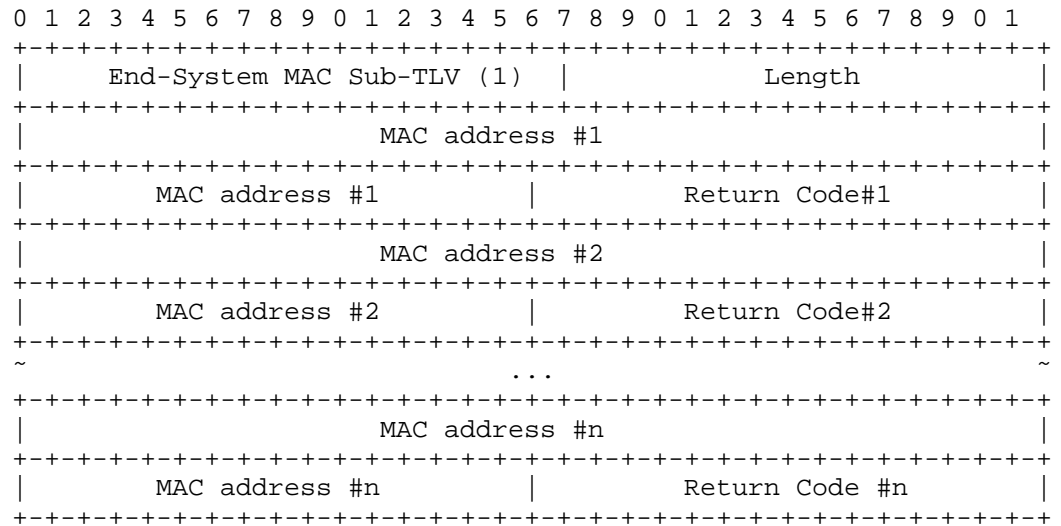
#### Sub-TLV Types:-

Value	What it means
1	End-System MAC Sub-TLV
2	End-System IPv4 Sub-TLV
3	End-System IPv6 Sub-TLV
4	End-System MAC/IPv4 Sub-TLV
6	End-System MAC/IPv6 Sub-TLV

#### End-System Return Code:-

Value	What it means
1	End-System Present
2	End-System Not Present

## 9.1.1.1. Sub-TLV for Validating End-System MAC Address



MAC Address: MAC Address of the End-System, that user is interested to validate.

Return Code: Return Code specifying status of End-System at Overlay End Point

## 9.1.2. Sub-TLV for Validating End-System IP Address

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      End-System IPv4 Sub-TLV (2) |      Length      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     IP address #1      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Return Code #1      |      IP address #2      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      IP address #2      |      Return Code #2      |
+-----+-----+-----+-----+-----+-----+-----+-----+
~                                     ~
...
+-----+-----+-----+-----+-----+-----+-----+-----+
|      IP address #n      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Return Code #n      |
+-----+-----+-----+-----+-----+-----+-----+

```

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      End-System IPv6 Sub-TLV (3) |      Length      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     IPv6 Address #1    |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Return Code #1      |      IPv6 Address #2...  |
+-----+-----+-----+-----+-----+-----+-----+-----+
~                                     ~
...
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     IPv6 Address #n    |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Return Code #n      |
+-----+-----+-----+-----+-----+-----+-----+

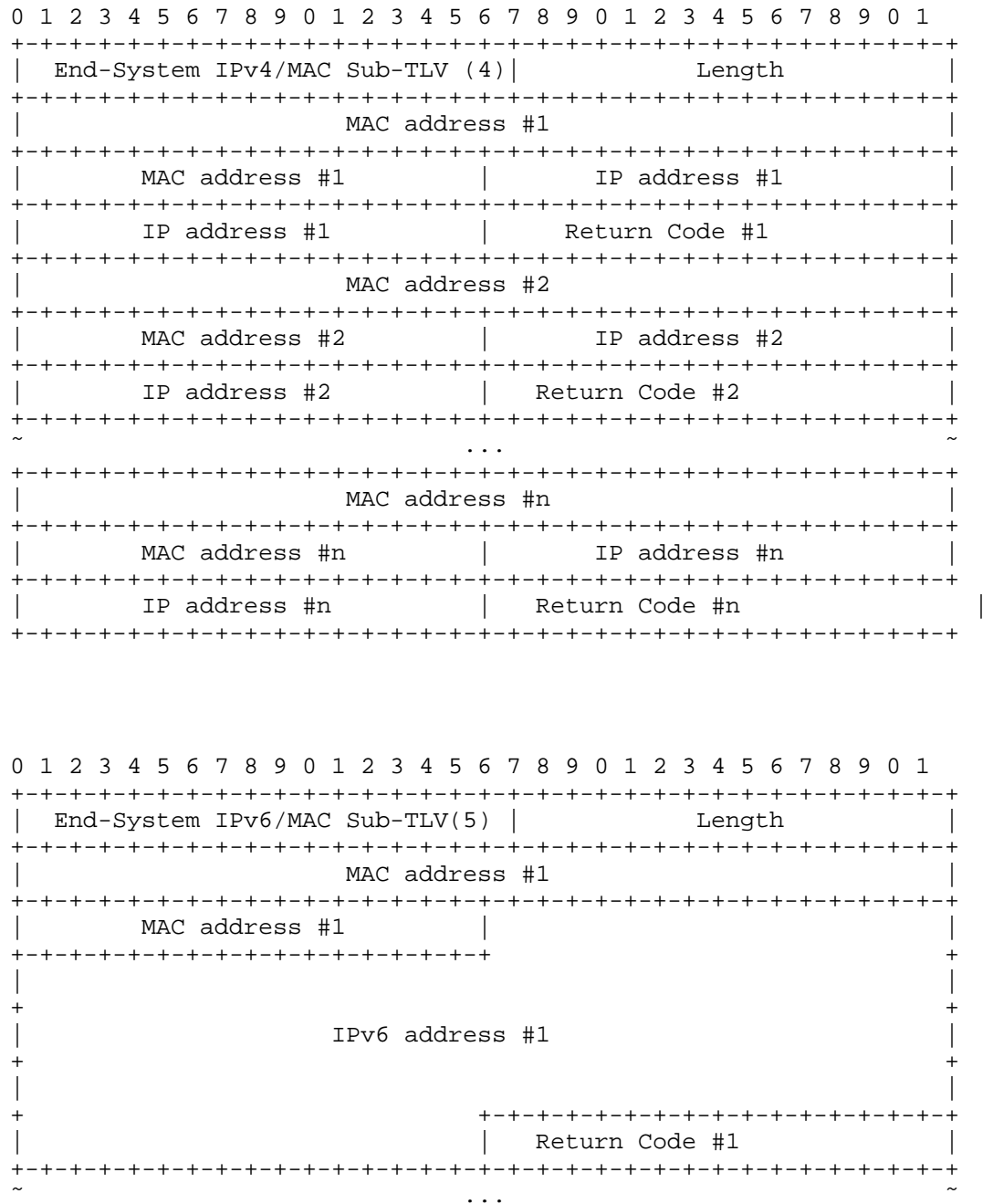
```

IP Address : IP Address of the End-System, that user is interested to validate.

Return Code: Return Code specifying status of End-System at Overlay End Point



## 9.1.3. Sub-TLV for Validating End-System MAC and IP Address



```

+-----+
|                                     MAC address #n                                     |
+-----+
|             MAC address #n             |
+-----+
|
+
|             IPv6 address #1             |
+
+
|                                     +-----+
|                                     | Return Code #1 |
|                                     +-----+
+-----+

```

IP Address : IP Address of the End-System, that user is interested to validate.

MAC Address: MAC Address of the End-System, that user is interested to validate.

Return Code: Return Code specifying status of End-System at Overlay End Point

## 9.2. Sending End-System Ping Request

When it is desired to check presence of a given End-System, the Echo Request Message is prepared as described in above Section "Procedure for Overlay Segment Ping". This packet should compose of Outer Header, Overlay Header, Inner Header, Generic Overlay Header with TLV representing desired Overlay Type (VXLAN, NVGRE, MPLSoGRE or MPLSoUDP). Apart from this the packet should also have one of the Sub-TLV's as defined in above section "Sub-TLV for End-System Ping" to identify the type of End-System Ping that user is interested in.

Because of the above mentioned encapsulation, it would be guaranteed that the packet follows the same Data Path as that of any End-User data going over the given Overlay Segment.

User need to fill in MAC, IP or MAC/IP combination for the End-System(s) that needs to be validated at the Overlay End Point in the respective Sub-TLV for End-System Ping.

## 9.3. Receiving End-System Ping Request

On receiving the End-System Ping Request the processing to trap this Packet, and sent it to Control Plane is done by Overlay Terminating End-System as define in above Section "Procedure for Overlay Segment Ping". Once the OAM Packet reaches OAM Application, it is identified as End-System Ping Request by virtue of presence any of the Sub-TLV's as defined in Section "Sub-TLV for End-System Ping".

If the Sub-TLV is of Type "End-System MAC Sub-TLV", the Overlay End Point should iterate through the list of MAC Addresses and verify the presence of individual MAC Address in its Flow Table or MAC Table for the given Overlay Segment.

If the MAC Address is present, it should set the respective End-System's Return Code field in the Sub-TLV to 1 "End-System-Present".

If the MAC Address is not present, it should set respective the End-System's Return Code field in the Sub-TLV to 2 "End-System-Not-Present".

If the Sub-TLV is of Type "End-System IP Sub-TLV", the Overlay End Point should iterate through the list of IP Addresses and verify the presence of individual IP Address in its Flow Table or Route Table for the given Overlay Segment.

If the IP Address is present, it should set the respective End-System's Return Code field in the Sub-TLV to 1 "End-System-Present".

If the IP Address is not present, it should set respective the End-System's Return Code field in the Sub-TLV to 2 "End-System-Not-Present".

If the Sub-TLV is of Type "End-System MAC and IP Sub-TLV", the Overlay End Point should iterate through the list of MAC/IP Addresses and verify the presence of individual MAC/IP Combination in its Flow Table or MAC and IP Table for the given Overlay Segment.

If the IP and MAC Address is present, it should set the respective End-System's Return Code field in the Sub-TLV to 1 "End-System-Present".

If the IP and MAC Address is not present, it should set respective the End-System's Return Code field in the Sub-TLV to 2 "End-System-Not-Present".

#### 9.4. Sending End-System Ping Reply

The procedure for sending End-System Echo Reply is same as defined in above section "Sending VXLAN Echo Reply". The replier MUST fill Sub-TLV with proper Return Code for each element in the End-System Sub-TLV.

#### 9.5. Receiving End-System Ping Reply

An Originating Overlay End Point should only receive Echo Reply for End-System Ping, in response to an Echo Request that it sent. By

virtue of presence of End-System Sub-TLV it would identify the status of respective End-System, and report it to the user. The other part of the handling is similar to section "Receiving VXLAN Echo Reply"

## 10. Security Considerations

TBD

## 11. Management Considerations

None

## 12. Acknowledgements

This document is the outcome of many discussions among many people, including Saurabh Shrivastava, Krishna Ram Kuttuva Jeyaram and Suresh Boddapati of Nuage Networks, Jorge Rabadan of Alcatel-Lucent, Inc and Rahul Kasralikar of Juniper Networks, Inc.

### 13. IANA Considerations

Action-1: This specification reserves a IANA UDP Port Number to be used when sending the Overlay OAM Packet

Action-2: This specification reserves a IANA Ethernet unicast Address for VXLAN/NVGRE Exception handling. This Address needs to be reserved from the block. "IANA Ethernet Address block - Unicast Use"



## 14. References

### 14.1. Normative References

- [I-D.draft-ietf-mpls-in-udp]  
Xu, Sheth, Yong, Pignataro, Yongbing , and Li,  
"Encapsulating MPLS in UDP", May 2013.
- [I-D.draft-lasserre-nvo3-framework]  
Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y.  
Rekhter, "Framework for DC Network Virtualization",  
September 2011.
- [I-D.draft-mahalingam-dutt-dcops-vxlan]  
Mahalingam, M., Dutt, D., Agarwal, P., Kreeger, L.,  
Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A  
Framework for Overlaying Virtualized Layer 2 Networks  
over Layer 3 Networks", May 2013.
- [I-D.draft-singh-nvo3-nvgre-router-alert]  
Singh, K., Jain, P., Balus, F., and W. Henderickx, "NVGRE  
Router Alert Option", May 2013.
- [I-D.draft-singh-nvo3-vxlan-router-alert]  
Singh, K., Jain, P., Balus, F., and W. Henderickx, "VxLAN  
Router Alert Option", May 2013.
- [I-D.draft-sridharan-virtualization-nvgre]  
Sridharan, M., Duda, K., Greenberg, A., Lin, G., Pearson,  
M., Thaler, P., Tumuluri, C., Venkataramiah, N., and Y.  
Wang, "NVGRE: Network Virtualization using Generic Routing  
Encapsulation", February 2013.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y.,  
Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack  
Encoding", RFC 3032, January 2001.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating  
MPLS in IP or Generic Routing Encapsulation (GRE)",  
RFC 4023, March 2005.
- [RFC4365] Rosen, E., "Applicability Statement for BGP/MPLS IP  
Virtual Private Networks (VPNs)", RFC 4365, February 2006.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol  
Label Switched (MPLS) Data Plane Failures", RFC 4379,  
February 2006.

## 14.2. Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4330] Mills, D., "Simple Network Time Protocol (SNTP) Version 4 for IPv4, IPv6 and OSI", RFC 4330, January 2006.

## Authors' Addresses

Pradeep Jain  
Nuage Networks  
755 Ravendale Drive  
Mountain View, CA 94043  
USA

Email: [pradeep@nuagenetworks.net](mailto:pradeep@nuagenetworks.net)

Kanwar Singh  
Nuage Networks  
755 Ravendale Drive  
Mountain View, CA 94043  
USA

Email: [kanwar@nuagenetworks.net](mailto:kanwar@nuagenetworks.net)

Florin Balus  
Nuage Networks  
755 Ravendale Drive  
Mountain View, CA 94043  
USA

Email: [florin@nuagenetworks.net](mailto:florin@nuagenetworks.net)

Wim Henderickx  
Alcatel-Lucent  
Copernicuslaan 50  
Antwerp 2018  
Belgium

Email: [wim.henderickx@alcatel-lucent.be](mailto:wim.henderickx@alcatel-lucent.be)

Vinay Bannai  
PayPal  
2211 N. First St,  
San Jose 95131  
USA

Email: [vbannai@paypal.com](mailto:vbannai@paypal.com)

Ravi Shekhar  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, CA 94089  
USA

Email: [rshekhar@juniper.net](mailto:rshekhar@juniper.net)

Anil Lohiya  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, CA 94089  
USA

Email: [alohiya@juniper.net](mailto:alohiya@juniper.net)



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 18, 2014

Z. Li  
L. Yong  
J. Zhang  
Huawei Technologies  
February 14, 2014

Segment-Based EVPN(S-EVPN)  
draft-li-l2vpn-segment-evpn-01

Abstract

This document proposes an enhanced EVPN mechanism, segment-based EVPN (S-EVPN). It satisfies the requirements of PBB-EVPN but does not require PBB implementation on PE. The solution uses a global label for each Ethernet Segment (ES) in an EVPN. It inserts the source ES label into packets at ingress PE and learns C-MAC and source ES label binding at egress PE. The solution makes the implementation easier and closer to E-VPN compared to PBB-EVPN but has the PBB-EVPN benefits.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on August 18, 2014.

## 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 . . . . .	3
3. Challenges of PBB-EVPN Implementation . . . . .	4
4. Architecture of S-EVPN . . . . .	6
4.1. C-MAC Learning . . . . .	6
4.2. ES Global Label Assignment . . . . .	7
4.3. Ethernet A-D Route Per EVI . . . . .	9
4.4. Ethernet A-D Route Per ES . . . . .	9
5. Improvement on EVPN . . . . .	9
5.1. Split Horizon . . . . .	9
5.2. Unifying MPLS Forwarding . . . . .	10
6. BGP E-VPN NLRI Extensions . . . . .	11
6.1. ES Global Label Request Extended Community . . . . .	11
6.2. ES Global Label Mapping Route . . . . .	11
7. Operations . . . . .	12
7.1. ES Global Label Request . . . . .	12
7.2. ES Global Label Allocation . . . . .	13
8. Solution Advantages . . . . .	13
9. IANA Considerations . . . . .	14
10. Security Considerations . . . . .	14
11. References . . . . .	14
11.1. Normative References . . . . .	14
11.2. References . . . . .	14
Authors' Addresses . . . . .	15

## 1. Introduction

E-VPN [I-D.ietf-l2vpn-evpn] introduces a solution for multipoint L2VPN services. It has multi-homing capability and uses BGP for distributing customer/client MAC address reachability information

over the core MPLS/IP network. PBB-EVPN [I-D.ietf-l2vpn-pbb-evpn] integrates PBB and E-VPN to achieves following objects:

1. reduce the number of MAC advertisement routes in BGP;
2. provide client MAC address mobility;
3. confine the scope of C-MAC learning to only active flows;
4. offer per site policies and avoid C-MAC address flushing on topology changes.

This document discusses the challenges faced by PBB-EVPN in the implementation and operation. It proposes an enhanced E-VPN mechanism, i.e. segment-based EVPN (S-EVPN), that provides the same benefits as of PBB-EVPN but does not require implementing PBB function on PE. S-EVPN mechanism allocates a global label for each Ethernet Segments in E-VPN, inserts the source ES label into the packet at ingress PE, and learns C-MAC and source ES label binding at egress PE. As a result it is not necessary to determine the source of C-MAC according to the B-MAC encapsulation which is required in PBB-EVPN. S-EVPN has simpler operation and management of EVPN and better encapsulation efficiency of packets compared to PBB-EVPN. In addition, it is easy to enhance the E-VPN to support S-EVPN and S-EVPN can unify the unicast traffic forwarding no matter C-MACs are learned by control plane or data plane.

## 2. Terminology

BEB: Backbone Edge Bridge

B-MAC: Backbone MAC Address

CE: Customer Edge

C-MAC: Customer/Client MAC Address

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

E-VPN: Ethernet VPN

EVI: Ethernet VPN Instance

LACP: Link Aggregation Control Protocol

P2P: Point to Point



PBB: Provider Backbone Bridge

PE: Provider Edge

S-EVPN: Segment-based EVPN

### 3. Challenges of PBB-EVPN Implementation

PBB-EVPN has advantages in the following aspects as [I-D.ietf-l2vpn-pbb-evpn]:

- MAC Advertisement Route Scalability
- C-MAC Mobility with MAC Sub-netting
- C-MAC Address Learning and Confinement
- Seamless Interworking with TRILL and 802.1aq Access Networks
- Per Site Policy Support
- Avoiding C-MAC Address Flushing

However, there are some challenges to implement PBB-EVPN.

#### 1. Creation and Management B-MAC

For PBB-EVPN, the choice of B-MAC address(es) for the PE nodes must be examined carefully as it has implications on the proper operation of multi-homing. These addresses are usually locally administered by the Service Provider which involves a lot of operation and management such as design, configuration and checking. Automating B-MAC Address Assignment can be applied, but for some scenarios the method cannot work and manual provision is inevitable. A more general automated solution can be proposed to reduce manual intervention.

#### 2. Encapsulation Efficiency of PBB-EVPN

When PBB encapsulation (shown in the figure 1) is adopted in PBB-EVPN, the B-DA, I-Tag, etc. fields in the encapsulation are useless in PBB-EVPN which reduce the effective payload.

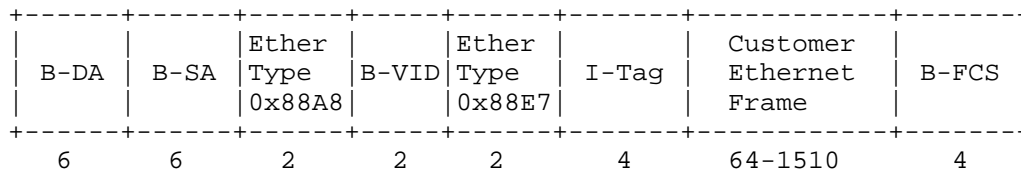


Figure 1: PBB Encapsulation

In the PBB encapsulation for PBB-EVPN, the source B-MAC is necessary since the egress PE need to learn the correspondence between C-MACs and B-MACs. The destination B-MAC is not necessary since the destination (egress PE) is reachable through the tunnel setup in advance instead of searching routes according to the destination B-MAC.

The I-SID is also not necessary any more. PBB divides the Ethernet network into two layers: I-Component and B-Component. In the egress PE, B-Component need identify I-Component through I-SID. For PBB-VPLS, MAC learning is through the data plane which is always to use broadcast or multicast for unknown unicast traffic. In order to indentify different forwarding instance, I-SID must be adopted. For PBB-EVPN, the forwarding instance is constructed through the control plane. That is, the forwarding instance is constructed through the RT matching of EVIs and identified by the label advertised. So I-SID information in PBB encapsulation for PBB-EVPN is not useful any more.

In addition B-VID in PBB encapsulation is almost never used. In a summary, in the PBB encapsulation for PBB-EVPN, only source B-MAC is indispensable. The encapsulation efficiency can be optimized.

### 3. Combination of PBB and E-VPN

The issues are dealt with by PBB-EVPN through the combination of two distinct technologies: PBB (layer 2 technology) and MPLS technology. In order to reduce the number of BGP MAC advertisement routes in E-VPN, PBB-EVPN can aggregate Customer/Client MAC (C- MAC) addresses via Provider Backbone MAC address (B-MAC). In fact, C-MAC addresses can be aggregated via MPLS label. Thus the issue solved by PBB-VPN can be solved in the method that is based on only MPLS technology. That is, the method is similar as E-VPN which is only based on MPLS technology. In other word, we can enhance E-VPN according to the similar way to gain PBB-EVPN benefits but not implement PBB on PE, which is a cleaner and simpler solution than PBB-EVPN.

#### 4. Architecture of S-EVPN

To implement C-MAC summarization scheme, Segment-based EVPN (S-EVPN) introduces a global label for each Ethernet Segment in an EVPN regardless single homed or multi-homed CE. BGP needs to advertise the global label and Ethernet Segment binding to all PEs. In data plane, the ingress PE inserts the source Ethernet Segment label into packets; the egress PE learns the C-MAC and source Ethernet Segment label binding upon receiving packets. S-EVPN purely relies on BGP IP/MPLS technology.

The encapsulation of S-EVPN is shown in figure 2 for the case that MPLS tunnel is used to transport EVPN traffic. The outmost label is the label for MPLS tunnel. The second label is the label which is allocated for Ethernet A-D route per EVI as E-VPN [I-D.ietf-l2vpn-evpn] and can identify a given <ESI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> (where the Ethernet Tag ID is set to 0). The third label is a global label which identifies an Ethernet Segment uniquely. The global label allocated for a specific Ethernet Segment will be described in section 4.2.

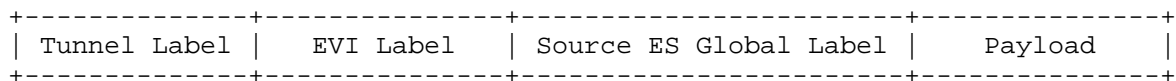


Figure 2: S-EVPN Encapsulation

##### 4.1. C-MAC Learning

In S-EVPN, C-MACs can be learned in the data plane to determine which source Ethernet Segment they are from and which EVI they belongs to. The forwarding entry to these learned C-MACs can be installed according to the source ES and EVI information.

In S-EVPN, the ingress PE needs to send unknown traffic with source C-MACs to all remote PEs according to the encapsulation as shown in figure 2. When a specific egress PE receives the packet:

1. it can learn the C-MAC and possible VLAN Tag in the payload;
2. it can learns the EVI which the C-MAC belongs to according to the EVI label which is allocated by the egress PE;
3. it can learns the Source Ethernet Segment which the CMAC belongs to according to the advertised the global label and Ethernet Segment binding in BGP.

Then the egress PE needs to install the forwarding entry to the learned C-MAC. The forwarding entry to the C-MAC need two types of information: the reachability information to the ingress PE which the C-MAC belongs to; the identification for the Ethernet Segment of the EVI on the ingress PE through which the packet can send to the C-MAC.

1. Tunnel to the ingress PE: the egress PE determines PE which the Source Ethernet Segment belongs to according to the advertised the global label and Ethernet Segment binding in BGP. Then egress PE can determine the tunnel to the ingress PE.

2. Label for the Ethernet Segment of the EVI on the ingress PE: The ingress PE needs to allocate label for the <ESI, EVI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> and advertise the corresponding Ethernet A-D Route per EVI to remote PEs. The egress PE can determines the Source Ethernet Segment, the EVI and the possible VLAN which the learned the C-MAC belongs to. Then it can determine the label binded to the <ESI, EVI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> which is advertised though the Ethernet A-D Route per EVI by the ingress PE.

Besides the two types of forwarding information, when the egress PE sends a specific packet to the learned C-MAC, it needs to determine the Ethernet Segment from which the packet come and encapsulate the global label for the Ethernet Segment firstly in the packet.

According to above procedures in S-EVPN, the egress PE can learn C-MACs and install forwarding entries to these C-MACs.

#### 4.2. ES Global Label Assignment

In S-EVPN, C-MAC summarization is done per an Ethernet Segment. The global ES label is introduced to identify the Ethernet Segment. The advantages of using global label are:

1. identify the ES globally;
2. leverage existing MPLS label stack implementation;
3. the label can be allocated dynamically to automate provision.

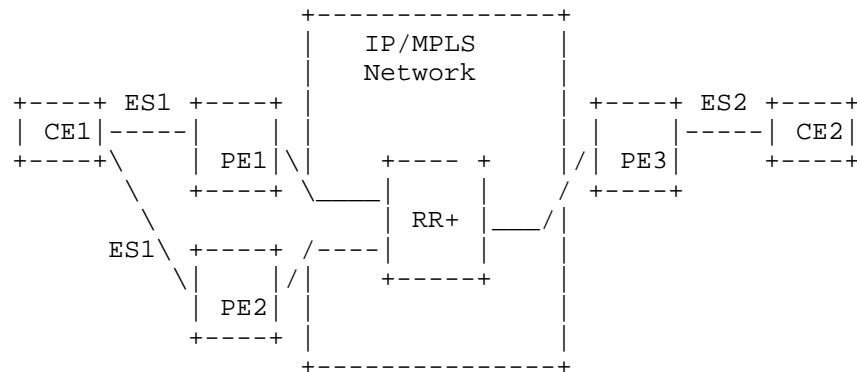


Figure 3: S-EVPN Network

In order to allocate a global label for an Ethernet Segment, there should be a centralized control point. Route Reflector (RR) of BGP may serve as this role and we call this type of RR as RR+. The S-EVPN network is shown in the figure 3. All PEs of S-EVPN connects with RR+. The procedure is as follows:

#### 1. Auto-Discovery of Ethernet Segment

RR+ can learn Ethernet Segment through the Ethernet A-D route per Ethernet Segment defined by [I-D.ietf-l2vpn-evpn]. Note that, in S-EVPN, every ES MUST has a unique identifier including the single-homed CEs. That is, ESI 0 cannot denote for a single-homed CE in S-EVPN. The ESI for the single-homed CE MUST be unique network wide and can be created automatically. The ESI is encoded as a ten octets integer. One way to generate ESI value for a single-homed CE is to use the MAC address of the Ethernet Segment with the low order 4 octets filled by value 0. The ESI value generation for multi-homed CE is specified in EVPN and can be reused in S-EVPN. Through Ethernet A-D route per Ethernet Segment, RR+ can learn all Ethernet Segments on all PEs.

#### 2. ES Global Label Allocation

[I-D.li-mpls-global-label-framework] specifies the framework for the global label allocation. In S-EVPN, RR+ allocates global labels for the Ethernet Segments discovered and advertises <Ethernet Segment, label> pair to all PEs. The PEs that are members of E-VPN keep track of the global label/Ethernet Segment mappings.

#### 4.3. Ethernet A-D Route Per EVI

The procedures defined for Ethernet A-D router per EVI in [I-D.ietf-l2vpn-evpn] will be reused by S-EVPN. In S-EVPN, both single home CE and multi-home CE have a unique ES identification. So for both single-homed CEs and multi-homed CEs, PEs need to allocate MPLS label for the <ESI, EVI, Ethernet Tag ID> tuple per EVI or per <ESI, EVI> and advertise corresponding Ethernet A-D routes per EVI. The MPLS label is used to identify a specific Ethernet Segment in an EVI.

#### 4.4. Ethernet A-D Route Per ES

In S-EVPN, support of Ethernet A-D Route per Ethernet Segment is still MANDATORY. PEs can learn Ethernet Segments through this type of route as E-VPN. In S-EVPN, RR+ which all PEs connect to can also learn Ethernet Segments. When constructing the Ethernet A-D Route per Ethernet Segment, there are following differences from E-VPN:

- The ESI for the single-homed CE in this route MUST be unique network wide instead of 0.

- The "ESI Label Extended Community" MUST be included in the route and the "Active-Standby" bit in the flags MUST be set accordingly. But the MPLS label in the extended community can be set as 0 (Invalid MPLS label value) since the ES global label is introduced in S-EVPN which can substitute ESI label.

### 5. Improvement on EVPN

When S-EVPN process is introduced, the E-VPN process defined by [I-D.ietf-l2vpn-evpn] can also be improved. The improvement includes split horizon, unifying unicast and multicast forwarding.

#### 5.1. Split Horizon

ES global label is introduced to identify the Ethernet Segment globally. Thus S-EVPN can fulfill requirements proposed by PBB-EVPN. Besides this, the ES global label can also be used for split horizon in EVPN. In order to achieve split horizon function, E-VPN adopts ESI label to encapsulate it in every BUM packet originating from a non-DF PE to identify the Ethernet Segment of origin. ES global label can use for the same purpose since it can identify the Ethernet Segment. Every BUM packet originating from a non-DF PE is encapsulated as the encapsulation which is shown in the figure 2. Since the original ESI label in E-VPN can be substituted by the ES global label, the ESI label in the ESI Label Extended Community can be an invalid label value. For the reason of compatibility, the ESI

Label Extended Community can carry a valid ESI label. Both ESI label and ES global label SHOULD be used for split horizon no matter which label is encapsulated in the packet.

[I-D.ietf-l2vpn-evpn-req] specifies the multicast optimization requirements to use MP2MP LSPs in EVPN. The ES global label can also solve the possible issue for split horizon when MP2MP LSP is used to transport BUM traffic. In E-VPN, when P2MP LSPs is used the upstream label assignment mechanism is introduced for split horizon. When PE received the packet, it decapsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label allocated by the ingress PE for a specific Ethernet Segment, the received PE will not forward the packet on the corresponding ES. In the MP2MP LSP scenarios, there are multiple roots and the upstream label allocated for Ethernet Segment maybe the same. So the received PE cannot determine a correct context label space according the top label for the MP2MP LSP. That is, the upstream label assignment mechanism for split horizon introduced in the P2MP LSP scenario can not be reused in the MP2MP LSP. But if the ES global label is used, in the MP2MP LSP scenario the received PE can also determine not to forward the packet on the specific ES which is identified by the ES global label. In one word, no matter ingress replication, P2MP LSP, or MP2MP, S-EVPN provides a unified solution for split horizon based on the ES global label. It can reduce the complexity of the split horizon mechanism in E-VPN.

## 5.2. Unifying MPLS Forwarding

S-EVPN adopts MPLS forwarding for C-MAC learning. In the control plane, it is just to add one new route type for E-VPN. It is a smooth upgrading of E-VPN and can switch easily between C-MAC learning through control plane and C-MAC learning through data plane.

When C-MACs is learned through the control plane, the unicast forwarding uses the label for the MAC route which is shown as follows:

```

+-----+-----+-----+
| Tunnel Label | MAC Label |   Payload   |
+-----+-----+-----+
```

Figure 4: E-VPN Unicast Forwarding Encapsulation

When C-MACs is learned through the data plane, the unicast forwarding uses the EVI label and the Segment global label which is shown in figure 2. In fact even if the C-MAC is learned through the data plane, the data plane can also use following encapsulation. In this case, the label in MAC advertisement route should not be used. From

the comparison, we can see that when E-VPN and S-EVPN are introduced, the forwarding encapsulation can be unified no matter which way C-MACs are learned by.

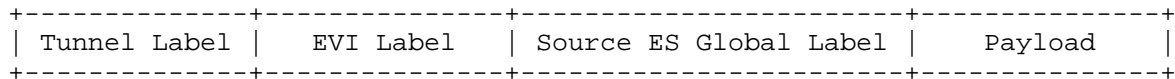
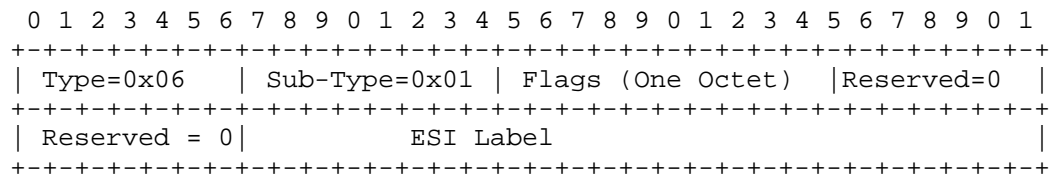


Figure 5: Unicast Forwarding Encapsulation without MAC Label

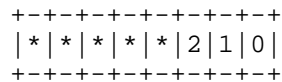
## 6. BGP E-VPN NLRI Extensions

### 6.1. ES Global Label Request Extended Community

ES Global Label Request Extended Community may be advertised along Ethernet A-D route per Ethernet Segment. ES Global Label Request Extended Community can reuse ESI Label Extended Community defined in [I-D.ietf-l2vpn-evpn] which is shown in the following figure:



There defines a new bit of the flag octet as the "Global Label Request" bit.



Bit0: "Active-Standby" bit  
 Bit1: "Root-Leaf" bit  
 Bit2: "Global Label Request" bit

The third low order bit of the flags octet is defined as the "Global Label Request". A value of 0 means there is no global label request for the Ethernet A-D route. A value of 1 means that global label request is associated with the Ethernet A-D route.

### 6.2. ES Global Label Mapping Route

A new route type is defined for E-VPN NLRI to allocate global label for Ethernet Segment:

+5 - ES Global Label Mapping Route



An ES Global Label Mapping route type specific E-VPN NLRI consists of the following:

```

+-----+
|      RD      (8 octets)      |
+-----+
|Ethernet Segment Identifier (10 octets)|
+-----+
|  Ethernet Tag ID (4 octets)  |
+-----+
|  MPLS Global Label (3 octets) |
+-----+
|          .....          |
+-----+
|  MPLS Global Label (3 octets) |
+-----+

```

## 7. Operations

### 7.1. ES Global Label Request

Global label request is only for the Ethernet A-D route per Ethernet Segment. The Ethernet A-D route per Ethernet Segment is constructed as defined by [I-D.ietf-l2vpn-evpn]. The Ethernet Segment Identifier MUST be a unique ten octet entity. Even if the CE is single-homed, the corresponding Ethernet Segment Identifier MUST NOT be the reserved value 0.

When request a global label for a specific Ethernet Segment, ES Global Label Request Extended Community MUST be used for the Ethernet A-D route. ES Global Label Request Extended Community of S-EVPN can reuse the ESI Label Extended Community. The "Global Label Request" bit of the flag octet MUST be set as 1 for Global Label Request. According to Section 5 "Improvement on E-VPN", if ES global label is introduced, the original ESI label MAY NOT be used. The "root-leaf" bit of the flag octet and the ESI Label value in the ESI Label Extended Community can always be 0 to simplify the process.

One or more Route Target(RT) MUST be carried with the Ethernet A-D route. These RTs are the set of RTs associated with all the EVIs to which the Ethernet Segment belongs. Since the Global label is allocated per Ethernet Segment, RTs carried by the Ethernet A-D route will be ignored by the RR+ when allocate global label for the Ethernet Segment specified in the Ethernet A-D routes. The global label per Ethernet Segment is advertised to all PEs. For multi-homed Ethernet Segment, if one EVI on one PE requests label allocation for the Ethernet Segment and the ES Global Label Mapping Route has been

advertised corresponding to the Ethernet Segment, other EVIs on other PEs SHOULD NOT send the global label request for the Ethernet Segment again, that is, the "Global Label Request" bit SHOULD set as 0 when advertise Ethernet A-D routes for the Ethernet Segment by these EVIs.

## 7.2. ES Global Label Allocation

When RR+ receives the Ethernet A-D route per Ethernet Segment and the "Global Label Request" bit of the ES Global Label Request Extended Community is set as 1, RR+ MUST allocate global label for the Ethernet Segment and advertise the ES Global Mapping route to all PEs.

The ES Global Label Mapping route is constructed as follows:

RD, Ethernet Segment Identifier and Ethernet Tag ID values can be directly derived from the corresponding Ethernet A-D route per Ethernet Segment.

The MPLS Global Label field carries one or more labels (that corresponds to the stack of labels [MPLS-ENCAPS]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [MPLS-ENCAPS]).

One or more Route Target(RT) MUST be carried with the ES Global Label Mapping route. These RTs can be directly derived from the RTs associated with the corresponding Ethernet A-D route.

For multi-homed Ethernet Segment, there maybe multiple global label request for the same Ethernet Segment advertised to RR+ by different PEs. When RR+ receives them, if RTs for these routes are same, owing to the Ethernet Segment Identifier is the same, it SHOULD advertise only one corresponding ES Global Label Mapping Route to all PEs. That is, the subsequent global label request for the same Ethernet Segment SHOULD be ignored. If RTs carried with the Ethernet A-D routes for the Ethernet Segment are different, RR+ SHOULD advertise multiple ES Global Label Mapping Routes with the same global label value and different RTs.

## 8. Solution Advantages

S-EVN has following advantages:

1. Remove the requirement of automating B-MAC address assignment to simplify provision of PBB-EVPN.
2. Improve the encapsulation efficiency of PBB-EVPN.

3. Seamless MPLS thoughts to solve the issue dealt with by PBB-EVPN instead of combination of two distinct technologies.

4. Be able to unify the split horizon mechanisms for ingress replication, P2MP LSP, and MP2MP LSP in E-VPN.

5. Be able to unify unicast traffic forwarding of E-VPN to implement seamless switch between C-MACs learning through control plane and C-MACs learning through data plane.

## 9. IANA Considerations

This document requires IANA to assign a new route type value for E-VPN NLRI.

## 10. Security Considerations

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

## 11. References

### 11.1. Normative References

[I-D.ietf-l2vpn-evpn-req]  
Sajassi, A., Aggarwal, R., Bitar, N., and A. Isaac,  
"Requirements for Ethernet VPN (EVPN)", draft-ietf-l2vpn-  
evpn-req-07 (work in progress), February 2014.

[I-D.ietf-l2vpn-evpn]  
Sajassi, A., Aggarwal, R., Henderickx, W., Isaac, A., and  
J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-  
l2vpn-evpn-05 (work in progress), February 2014.

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

### 11.2. References

[I-D.ietf-l2vpn-pbb-evpn]  
Sajassi, A., Salam, S., Boutros, S., Bitar, N., Isaac, A.,  
and L. Jin, "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-06 (work  
in progress), October 2013.

[I-D.li-mpls-global-label-framework]  
Li, Z., Zhao, Q., and T. Yang, "A Framework of MPLS Global  
Label", draft-li-mpls-global-label-framework-00 (work in  
progress), July 2013.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Lucy Yong  
Huawei Technologies  
1700 Alma Dr. Suite 500  
Plano, TX 75075  
USA

Email: lucyyong@huawei.com

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

Email: jackey.zhang@huawei.com

L2VPN Workgroup  
Internet Draft  
Intended status: Standards Track

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

F. Balus  
Nuage Networks

Expires: August 18, 2014

February 14, 2014

Data Center Interconnect Solution for EVPN Overlay networks  
draft-rabadan-l2vpn-dci-evpn-overlay-01.txt

#### Abstract

This document describes how Network Virtualization Overlay networks (NVO3) can be connected to a Wide Area Network (WAN) in order to extend the layer-2 connectivity required for some tenants. The solution will analyze the interaction between NVO3 networks running EVPN and other L2VPN technologies used in the WAN, such as VPLS/PBB-VPLS or EVPN/PBB-EVPN, and will propose a solution for the interworking between both.

#### 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 August 18, 2014.

#### 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 . . . . .	3
2. Decoupled DCI solution for EVPN overlay networks . . . . .	3
2.1. Interconnect requirements . . . . .	4
2.2. VLAN-based hand-off . . . . .	5
2.3. Pseudowire-based hand-off . . . . .	5
2.4. Multi-homing solution . . . . .	6
2.5. Data Center Gateway Optimizations . . . . .	7
2.5.1 Use of the Unknown MAC route to reduce unknown flooding . . . . .	7
2.5.2. MAC address advertisement control . . . . .	7
2.5.3. ARP flooding control . . . . .	8
3. Integrated DCI solution for EVPN overlay networks . . . . .	8
3.1. Interconnect requirements . . . . .	9
3.2. VPLS DCI for EVPN-Overlay networks . . . . .	10
3.2.1. Control/Data Plane setup procedures on the DC GWs . . . . .	10
3.2.2. Multi-homing procedures on the DC GWs . . . . .	11
3.3. PBB-VPLS DCI for EVPN-Overlay networks . . . . .	11
3.3.1. Control/Data Plane setup procedures on the DC GWs . . . . .	11
3.3.2. Multi-homing procedures on the DC GWs . . . . .	12
3.4. EVPN-MPLS DCI for EVPN-Overlay networks . . . . .	12
3.4.1. Control Plane setup procedures on the DC GWs . . . . .	12
3.4.2. Data Plane setup procedures on the DC GWs . . . . .	14
3.4.3. Multi-homing procedures on the DC GWs . . . . .	14
3.4.4. Impact on MAC Mobility procedures . . . . .	15
3.4.5. Data Center Gateway optimizations . . . . .	16
3.4.6. Benefits of the EVPN-MPLS DCI solution . . . . .	16
3.5. PBB-EVPN DCI for EVPN-Overlay networks . . . . .	17
3.5.1. Control/Data Plane setup procedures on the DC GWs . . . . .	17

3.5.2. Multi-homing procedures on the DC GWs . . . . .	18
3.5.3. Impact on MAC Mobility procedures . . . . .	18
3.5.4. Data Center Gateway optimizations . . . . .	18
5. Conventions and Terminology . . . . .	18
6. Security Considerations . . . . .	19
7. IANA Considerations . . . . .	19
8. References . . . . .	19
8.1. Normative References . . . . .	19
8.2. Informative References . . . . .	19
9. Acknowledgments . . . . .	20
10. Authors' Addresses . . . . .	20

## 1. Introduction

[EVPN-Overlays] discusses the use of EVPN as the control plane for Network Virtualization Overlay (NVO) networks, where VXLAN, NVGRE or MPLS over GRE can be used as possible data plane encapsulation options.

While this model provides a scalable and efficient multi-tenant solution within the Data Center, it might not be easily extended to the WAN in some cases due to the requirements and existing deployed technologies. For instance, a Service Provider might have an already deployed (PBB-)VPLS or (PBB-)EVPN network that must be used to interconnect Data Centers and WAN VPN users.

This document describes a Data Center Interconnect (DCI) solution for E-VPN overlay Data Center networks, assuming that the Data Center Gateway (DC GW) and the WAN Edge functions can be decoupled in two separate systems or integrated into the same system. The former option will be referred as "Decoupled DCI solution" throughout the document whereas the latter one will be referred as "Integrated DCI solution".

## 2. Decoupled DCI solution for EVPN overlay networks

This section describes the interconnect solution when the DC GW and WAN Edge functions implemented in different systems. Figure 1 depicts the reference model described in this section.

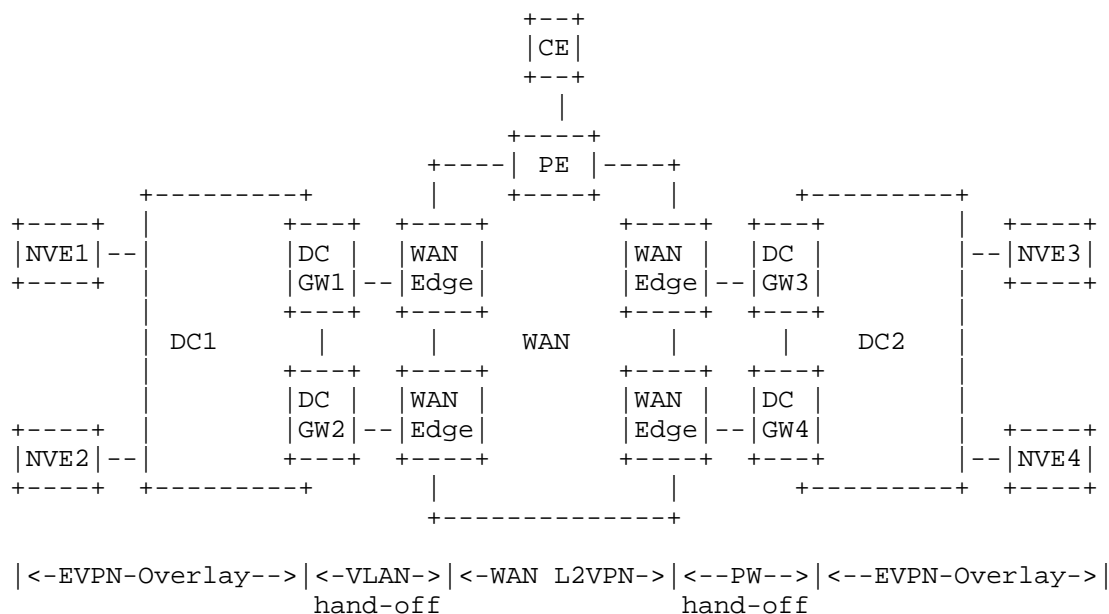


Figure 1 Decouple DCI model

The following section describes the interconnect requirements that make Service Providers select this model and the requirements of the solution itself.

## 2.1. Interconnect requirements

The proposed Interconnect architecture will be normally deployed in networks where the EVPN-Overlay provider and WAN providers are different entities and a clear demarcation is needed. The solution must observe the following requirements:

- o A simple connectivity hand-off must be provided between the EVPN-Overlay network provider and the WAN provider so that QoS and security enforcement are easily accomplished.
- o The solution must be independent of the L2VPN technology deployed in the WAN.
- o Multi-homing between DC GW and WAN Edge routers is required. Per-service load balancing MUST be supported. Per-flow load balancing MAY be supported but it is not a strong requirement since a deterministic path per service is usually required for an easy QoS and security enforcement.



- o Ethernet OAM and Connectivity Fault Management (CFM) functions must be supported between the EVPN-Overlay network and the WAN network.
- o The following optimizations MAY be supported at the DC GW:
  - + Unknown flooding reduction for the unicast traffic sourced from the DC Network Virtualization Edge devices (NVEs).
  - + Control of the WAN MAC addresses advertised to the DC.
  - + ARP flooding control for the requests coming from the WAN.

## 2.2. VLAN-based hand-off

In this option, the hand-off between the DC GWs and the WAN Edge routers is based on 802.1Q VLANs. This is illustrated in Figure 1, between the DC GWs in DC1 and the WAN Edge routers. Each EVPN Instance (EVI) in the DC GW is connected to a different VPLS/EVI instance in the WAN Edge router by using a different C-TAG VLAN ID or a different combination of S-TAG/C-TAG VLAN IDs that matches at both sides. In this use-case, the WAN Edge router becomes a VPLS/EVPN PE with regular Attachment Circuits.

This option provides the best possible demarcation between the DC and WAN providers and it does not require control plane interaction between both providers. The disadvantages of this model are the provisioning overhead and the reduced scalability (limited to the VLAN-ID space).

In this model, the DC GW acts as a regular Network Virtualization Edge (NVE) towards the DC. Its control plane, data plane procedures and interactions are described in [EVPN-Overlays].

The WAN Edge router acts as a (PBB-)VPLS or (PBB-)EVPN PE. Its functions are described in [RFC4761][RFC4762][RFC6074] or [EVPN][PBB-EVPN].

## 2.3. Pseudowire-based hand-off

If MPLS can be enabled between the DC GW and the WAN Edge router, a more scalable DCI solution can be deployed. In this option the hand-off between both routers is based on FEC128-based pseudowires or, alternatively, FEC129-based pseudowires for a greater level of network automation. Note that this model still provides a clear demarcation boundary between DC and WAN, and security/QoS policies may be applied on a per pseudowire basis. The PW-based hand-off interconnect is illustrated in Figure 1, between the DC2 DC GWs and the WAN Edge routers.

In this model, besides the usual MPLS procedures between DC GW and WAN Edge router, the DC GW MUST support an interworking function in

each EVI that requires extension to the WAN:

- o If a FEC128-based pseudowire is used between the EVI (DC GW) and the VSI (WAN Edge), the provisioning of the VCID for such pseudowire MUST be supported on the EVI and must match the VCID used in the peer VSI at the WAN Edge router.
- o If BGP Auto-discovery [RFC6074] and FEC129-based pseudowires are used between the DC GW EVI and the WAN Edge VSI, the provisioning of the VPLS-ID MUST be supported on the EVI and must match the VPLS-ID used in the WAN Edge VSI.

#### 2.4. Multi-homing solution

As already discussed, single-active multi-homing, i.e. per-service load-balancing multi-homing MUST be supported in this type of interconnect. All-active multi-homing may be considered in future revisions of this document.

The DC GWs will be provisioned with a unique ESI per WAN interconnect and the hand-off attachment circuits or pseudowires between the DC GW and the WAN Edge router will be assigned to such ESI. The ESI will be administratively configured on the DC GWs according to the procedures in [EVPN] and its use assumes that the DC GWs are connected to a single DC and to a single WAN domain. Multi-homing for cases where the DC GWs are connected to more than one DC and/or more than one WAN domain is for further study. This ESI will be referred as "DCI-ESI" hereafter.

The solution (on the DC GWs) MUST follow the single-active multi-homing procedures as described in [EVPN-Overlays] for the provisioned DCI-ESI, i.e. Ethernet A-D routes per ESI and per EVI will be advertised to the DC NVEs. The MAC addresses learnt (in the data plane) on the hand-off links will be advertised with the DCI-ESI encoded in the ESI field.

The use of OAM is recommended between the DC GWs and the WAN Edge routers:

- o If the DCI solution is based on a VLAN hand-off, 802.1ag/Y.1731 Ethernet-CFM can be used by the non-DF DC GW so that the peer WAN Edge router do not send any traffic to the DC GW for that particular EVI.
- o If the VPLS DCI solution is based on a pseudowire hand-off, the LDP PW Status bits TLV can be used by the non-DF to signal "Standby status" to the WAN Edge router for that particular EVI.

## 2.5. Data Center Gateway Optimizations

The following features MAY be supported on the DC GW in order to optimize the control plane and data plane in the DC.

### 2.5.1 Use of the Unknown MAC route to reduce unknown flooding

The use of EVPN, as the control plane of Network Virtualization Networks in the DC, brings a significant number of benefits as described in [EVPN-Overlays]. There are however some potential issues that SHOULD be addressed when the DC EVIs are connected to the WAN VPN instances.

The first issue is the additional unknown unicast flooding created in the DC due to the unknown MACs existing beyond the DC GW. In virtualized DCs where all the MAC addresses are learnt in the control/management plane, unknown unicast flooding is significantly reduced. This is no longer true if the DC GW is connected to a layer-2 domain with data plane learning.

The solution suggested in this document is based on the use of an "Unknown MAC route" that is advertised by the Designated Forwarder DC GW. The Unknown MAC route is a regular EVPN MAC/IP Advertisement route where the MAC Address Length is set to 48 and the MAC address to 00:00:00:00:00:00 (IP length is set to 0).

If this procedure is used, when an EVI is created in the DC GWs and the Designated Forwarder (DF) is elected, the DF will send the Unknown MAC route. The NVEs supporting this concept will prune their unknown unicast flooding list and will only send the unknown unicast packets to the owner of the Unknown MAC route. Note that the DCI-ESI will be encoded in the ESI field of the NLRI so that regular multi-homing procedures can be applied to this unknown MAC too (e.g. backup-path).

### 2.5.2. MAC address advertisement control

Another issue derived from the EVI interconnect to the WAN layer-2 domain is the potential massive MAC advertisement into the DC. All the MAC addresses learnt from the WAN on the hand-off attachment circuits or pseudowires must be advertised by BGP EVPN. Even if optimized BGP techniques like RT-constraint are used, the amount of MAC addresses to advertise or withdraw (in case of failure) from the DC GWs can be difficult to control and overwhelming for the DC network, especially when the NVEs reside in the hypervisors.

This document proposes the addition of administrative options so

that the user can enable/disable the advertisement of MAC addresses learnt from the WAN as well as the advertisement of the Unknown MAC route from the DF DC GW. In cases where all the DC MAC addresses are learnt in the control/management plane, the DC GW may disable the advertisement of WAN MAC addresses. Any frame with unknown destination MAC will be exclusively sent to the Unknown MAC route owner(s).

### 2.5.3. ARP flooding control

Another optimization mechanism, naturally provided by EVPN in the DC GWs, is the Proxy ARP function. The DC GWs SHOULD build a Proxy ARP cache table as per [EVPN]. When the active DC GW receives an ARP request coming from the WAN, the DC GW does a Proxy ARP table lookup and replies to the ARP request as long as the information is available in its table.

This mechanism is specially recommended on the DC GWs since it protects the DC network from external ARP-flooding.

## 3. Integrated DCI solution for EVPN overlay networks

When the DC and the WAN are operated by the same administrative entity, the Service Provider can decide to integrate the DC GW and WAN Edge PE functions in the same router for obvious CAPEX and OPEX saving reasons. This is illustrated in Figure 2. Note that this model does not provide an explicit demarcation link between DC and WAN anymore. ACLs or QoS policies between DC and WAN are not required.

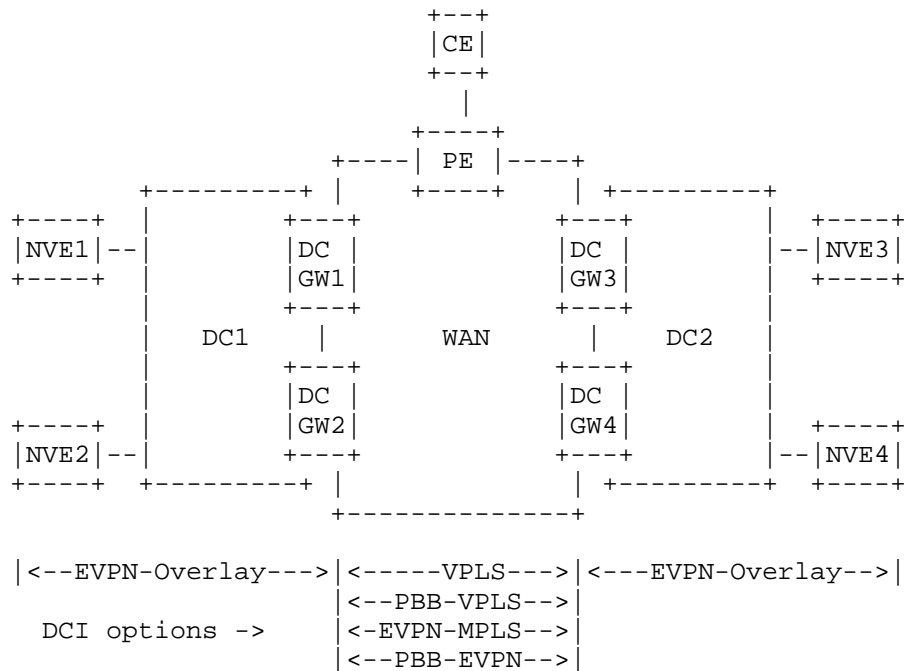


Figure 2 Integrated DCI model

### 3.1. Interconnect requirements

The solution must observe the following requirements:

- o The DC GW function must provide control plane and data plane interworking between the EVPN-overlay network and the L2VPN technology supported in the WAN, i.e. (PBB-)VPLS or (PBB-)EVPN, as depicted in Figure 2.
- o Multi-homing MUST be supported. Single-active multi-homing with per-service load balancing MUST be implemented. All-active multi-homing, i.e. per-flow load-balancing, MUST be implemented as long as the technology deployed in the WAN supports it.
- o If EVPN is deployed in the WAN, the MAC Mobility, Static MAC protection and other procedures (e.g. proxy-arp) described in [EVPN] must be supported end-to-end.
- o Any type of inclusive multicast tree MUST be independently supported in the WAN as per [EVPN], and in the DC as per [EVPN-Overlays].

### 3.2. VPLS DCI for EVPN-Overlay networks

#### 3.2.1. Control/Data Plane setup procedures on the DC GWs

Regular MPLS tunnels and TLDP/BGP sessions will be setup to the WAN PEs and RRs as per [RFC4761][RFC4762][RFC6074] and overlay tunnels and EVPN will be setup as per [EVPN-Overlays]. Note that different route-targets for the DC and for the WAN are normally required. A single type-1 RD per service can be used.

In order to support multi-homing, the DC GWs will be provisioned with a DCI-ESI (see section 2.4), that will be unique per interconnection. Note that Ethernet Segment is a system wide assigned value, as opposed to the Ethernet Segments defined in [EVPN]. All the [EVPN] procedures are still followed for the DCI-ESI, e.g. any MAC address learnt from the WAN will be advertised to the DC with the DCI-ESI in the ESI field.

A MAC-VRF per EVI will be created in each DC GW. The MAC-VRF will have two different types of tunnel bindings instantiated in two different split-horizon-groups:

- o VPLS pseudowires will be instantiated in the "WAN split-horizon-group".
- o Overlay tunnel bindings (e.g. VXLAN, NVGRE) will be instantiated in the "DC split-horizon-group".

Attachment circuits are also supported on the same MAC-VRF, but they will not be part of any of the above split-horizon-groups.

Traffic received in a given split-horizon-group will never be forwarded to a member of the same split-horizon-group.

As far as BUM flooding is concerned, a flooding list will be created with the sub-list created by the inclusive multicast routes and the sub-list created for VPLS in the WAN. BUM frames received from a local attachment circuit will be flooded to both sub-lists. BUM frames received from the DC or the WAN will be forwarded to the flooding list observing the split-horizon-group rule described above.

Note that the DC GWs are not allowed to have an EVPN binding and a pseudowire to the same far-end within the same MAC-VRF in order to avoid loops and packet duplication:

- o If an EVPN binding exists between two DC GWs and an attempt is made to setup a pseudowire between them, the pseudowire will be kept operationally down. The corresponding OAM signaling will be

triggered.

- o If a pseudowire exists between two DC GWs and an attempt is made to setup an EVPN binding, the pseudowire will be brought operationally down before establishing the EVPN binding.

The optimizations procedures described in section 2.5 can also be applied to this option.

### 3.2.2. Multi-homing procedures on the DC GWs

Single-active multi-homing MUST be supported on the DC GWs. All-active multi-homing is not supported by VPLS.

All the single-active multi-homing procedures as described by [EVPN-Overlays] will be followed for the DCI-ESI.

The non-DF DC GW for the DCI-ESI will block the transmission and reception of all the bindings in the "WAN split-horizon-group" for BUM and unicast traffic.

### 3.3. PBB-VPLS DCI for EVPN-Overlay networks

#### 3.3.1. Control/Data Plane setup procedures on the DC GWs

In this case, there is no impact on the procedures described in [RFC7041] for the B-component. However the I-component instances become EVI instances with EVPN-Overlay bindings and potentially local attachment circuits. M EVI instances can be multiplexed into the same B-component instance. This option provides significant savings in terms of pseudowires to be maintained in the WAN.

The DCI-ESI concept described in section 3.2.1 will also be used for the PBB-VPLS-based DCI.

B-component pseudowires and I-component EVPN-overlay bindings established to the same far-end will be compared. The following rules will be observed:

- o Attempts to setup a pseudowire between the two DC GWs within the B-component context will never be blocked.
- o If a pseudowire exists between two DC GWs for the B-component and an attempt is made to setup an EVPN binding on an I-component linked to that B-component, the EVPN binding will be kept operationally down. Note that the BGP EVPN routes will still be valid but not used.

- o The EVPN binding will only be up and used as long as there is no pseudowire to the same far-end in the corresponding B-component. The EVPN bindings in the I-components will be brought down before the pseudowire in the B-component is brought up.

The optimizations procedures described in section 2.5 can also be applied to this DCI option.

### 3.3.2. Multi-homing procedures on the DC GWs

Single-active multi-homing MUST be supported on the DC GWs. All-active multi-homing MAY be supported. Procedures for the support of all-active multi-homing are for further study.

All the single-active multi-homing procedures as described by [EVPN-Overlays] will be followed for the DCI-ESI for each EVI instance connected to B-component.

The non-DF DC GW for the DCI-ESI will block the transmission and reception of all the EVPN bindings in the corresponding I-components for BUM and unicast traffic.

### 3.4. EVPN-MPLS DCI for EVPN-Overlay networks

If EVPN for MPLS tunnels, EVPN-MPLS hereafter, is supported in the WAN, an end-to-end EVPN solution can be deployed. The following sections describe the proposed solution as well as the impact required on the [EVPN] procedures.

#### 3.4.1. Control Plane setup procedures on the DC GWs

The DC GWs MUST establish separate BGP sessions for sending/receiving EVPN routes to/from the DC and to/from the WAN. Normally each DC GW will setup one (two) BGP EVPN session(s) to the DC RR(s) and one(two) session(s) to the WAN RR(s). The route-distinguisher (RD) per EVI can be used for the EVPN routes sent to both, WAN and DC RRs. On the contrary, although reusing the same value is possible, different route-targets are expected to be handled for the same EVI in the WAN.

As in the other discussed options, a DCI-ESI will be configured on the DC GWs for multi-homing.

Received EVPN routes will never be reflected on the DC GWs but consumed and re-advertised (if needed):

- o Ethernet A-D routes, ES routes and inclusive multicast routes are consumed by the DC GWs and processed locally for the



corresponding [EVPN] procedures.

- o MAC/IP advertisement routes will be received, imported and if they become active in the MAC FIB, the information will be re-advertised as a new route:
  - + The RD will be the DC GW's RD for the service.
  - + The ESI will be set to the DCI-ESI.
  - + The Ethernet-tag will be 0 or a new value.
  - + The MAC length, MAC address, IP Length and IP address values will be kept from the previously received NLRI.
  - + The MPLS label will be 0 or a local label.
  - + The appropriate RTs and [RFC5512] BGP Encapsulation extended community will be used according to [EVPN-Overlays].

The DC GWs will also generate the following local EVPN routes that will be sent to the DC and WAN, with their corresponding RT and [RFC5512] BGP Encapsulation extended community values:

- o ES route for the DCI-ESI.
- o Ethernet A-D routes per ESI and EVI for the DCI-ESI.
- o Inclusive multicast routes with independent tunnel type value for the WAN and DC. E.g. a P2MP LSP may be used in the WAN whereas ingress replication is used in the DC.
- o MAC/IP advertisement routes for MAC addresses learnt in local attachment circuits. Note that these routes will not include the DCI-ESI, but ESI=0 or different from 0 for local Ethernet Segments (ES).

Note that each DC GW will receive two copies of each of the above routes generated by the peer DC GW (one copy for the DC encapsulation and one copy for the WAN encapsulation). This is the expected behavior on the DC GW:

- o ES and A-D (per ESI) routes: regular BGP selection will be applied.
- o Inclusive multicast routes: if the Ethernet Tag ID matches on both routes, regular BGP selection applies and only one route will be active. It is recommended to influence the BGP selection

so that the DC route is preferred. If the Ethernet Tag ID does not match, then BGP will consider them two separate routes. In that case, the EVI service will select the DC route.

- o MAC/IP advertisement routes for local attachment circuits: as above, the DC GW will select only one. The decision will be made at BGP or service level, depending on the Ethernet Tags.

The optimizations procedures described in section 2.5 can also be applied to this option.

#### 3.4.2. Data Plane setup procedures on the DC GWs

The procedure explained at the end of the previous section will make sure there are no loops or packet duplication between the DC GWs of the same DC since only one EVPN binding will be setup in the data plane between the two nodes.

As for the rest of the EVPN tunnel bindings, two flooding lists will be setup by each DC GW for the same MAC-VRF:

- o EVPN-overlay flooding list (composed of bindings to the remote NVEs or multicast tunnel to the NVEs).
- o EVPN-mpls flooding list (composed of MP2P and or LSM tunnel to the remote PEs)

Each flooding list will be part of a separate split-horizon group. Traffic generated from a local AC can be flooded to both split-horizon-groups. Traffic from a binding of a split-horizon-group can be flooded to the other split-horizon-group and local ACs, but never to a member of its own split-horizon-group.

#### 3.4.3. Multi-homing procedures on the DC GWs

Single-active as well as all-active multi-homing MUST be supported.

All the multi-homing procedures as described by [EVPN] will be followed for the DF election for DCI-ESI, as well as the backup-path (single-active) and aliasing (all-active) procedures on the remote PEs/NVEs. The following changes are required at the DC GW with respect to the DCI-ESI:

- o Single-active multi-homing; assuming a WAN split-horizon-group, a DC split-horizon-group and local ACs on the DC GWs:
  - + Forwarding behavior on the non-DF: the non-DF MUST NOT forward BUM or unicast traffic received from a given split-horizon-

group to a member of his own split-horizon group or to the other split-horizon-group. Only forwarding to local ACs is allowed (as long as they are not part of an ES for which the node is non-DF).

- + Forwarding behavior on the DF: the DF MUST NOT forward BUM or unicast traffic received from a given split-horizon-group to a member of his own split-horizon group or to the non-DF. Forwarding to the other split-horizon-group and local ACs is allowed (as long as they are not part of an ES for which the node is non-DF).
- o All-active multi-homing; assuming a WAN split-horizon-group, a DC split-horizon-group and local ACs on the DC GWs:
  - + Forwarding behavior on the non-DF: the non-DF follows the same behavior as the non-DF in the single-active case but only for BUM traffic. Unicast traffic received from a split-horizon-group MUST NOT be forwarded to a member of its own split-horizon-group but can be forwarded normally to the other split-horizon-group and local ACs. If a known unicast packet is identified as a "flooded" packet, the procedures for BUM traffic MUST be followed.
  - + Forwarding behavior on the DF: the DF follows the same behavior as the DF in the single-active case but only for BUM traffic. Unicast traffic received from a split-horizon-group MUST NOT be forwarded to a member of its own split-horizon-group but can be forwarded normally to the other split-horizon-group and local ACs. If a known unicast packet is identified as a "flooded" packet, the procedures for BUM traffic MUST be followed.
- o No ESI label is required to be signaled for DCI-ESI for its use by the non-DF in the data path. This is possible because the non-DF and the DF will never forward BUM traffic (coming from a split-horizon-group) to each other.

#### 3.4.4. Impact on MAC Mobility procedures

Since the MAC/IP Advertisement routes are not reflected in the DC GWs but rather consumed and re-advertised if active, the MAC Mobility procedures can be constrained to each domain (DC or WAN) and resolved within each domain. In other words, if a MAC moves within the DC, the DC GW MUST NOT re-advertise the route to the WAN with a change in the sequence number. Only when the MAC moves from the WAN domain to the DC domain, the DC GW will re-advertise the MAC with a higher sequence number in the MAC Mobility extended community. In respect to the MAC

Mobility procedures described in [EVPN] the MAC addresses learnt from the NVEs in the local DC or on the local ACs will be considered as local.

The sequence numbers MUST NOT be propagated between domains. The sticky bit indication in the MAC Mobility extended community MUST be propagated between domains.

#### 3.4.5. Data Center Gateway optimizations

All the Data Center Gateway optimizations described in section 2.5 MAY be applied to the DC GWs when the DCI is based on EVPN-MPLS.

In particular, the use of the Unknown MAC route, as described in section 2.5.1, reduces the unknown flooding in the DC but also solves some transient packet duplication issues in cases of all-active multi-homing. This is explained in the following paragraph.

Consider the diagram in Figure 2 for EVPN-MPLS DCI and all-active multi-homing, and the following sequence:

- a) MAC Address M1 is advertised from NVE3 in EVI-1.
- b) DC GW3 and DC GW4 learn M1 for EVI-1 and re-advertise M1 to the WAN with DCI2-ESI in the ESI field.
- c) DC GW1 and DC GW2 learn M1 and install DC GW3/GW4 as next-hops following the EVPN aliasing procedures.
- d) Before NVE1 learns M1, a packet arrives to NVE1 with destination M1. The packet is subsequently flooded.
- e) Since both DC GW1 and DC GW2 know M1, they both forward the packet to the WAN (hence creating packet duplication), unless there is an indication in the data plane that the packet has been flooded by NVE1. If the DC GWs signal the same VNI/VSID for MAC/IP advertisement and inclusive multicast routes for EVI-1, such data plane indication does not exist.

This undesired situation can be avoided by the use of the Unknown MAC route. If this route is used, the NVEs will prune their unknown unicast flooding list, and the non-DF DC GW will not receive unknown packets, only the DF will. This solves the MAC duplication issue described above.

#### 3.4.6. Benefits of the EVPN-MPLS DCI solution

Besides retaining the EVPN attributes between Data Centers and

throughout the WAN, the EVPN-MPLS DCI solution on the DC GWs has some benefits compared to pure BGP EVPN RR or Inter-AS model B solutions without a gateway:

- o The solution supports the connectivity of local attachment circuits on the DC GWs.
- o Different data plane encapsulations can be supported in the DC and the WAN.
- o Optimized multicast solution, with independent inclusive multicast trees in DC and WAN.
- o MPLS Label aggregation: for the case where MPLS labels are signaled from the NVEs for MAC/IP Advertisement routes, this solution provides label aggregation. A remote PE MAY receive a single label per DC GW MAC-VRF as opposed to a label per NVE.
- o The DC GW will not propagate MAC mobility for the MACs moving within a DC. Mobility intra-DC is solved by all the NVEs in the DC. The MAC Mobility procedures on the DC GWs are only required in case of mobility across DCs.
- o Proxy-ARP function on the DGWs can be leveraged to reduce ARP flooding in the DC or/and in the WAN.

### 3.5. PBB-EVPN DCI for EVPN-Overlay networks

[PBB-EVPN] is yet another DCI option. It requires the use of DC GWs where I-components and associated B-components are EVI instances.

#### 3.5.1. Control/Data Plane setup procedures on the DC GWs

EVPN will independently run in both components, the I-component EVI and B-component EVI. Compared to [PBB-EVPN], the DC C-MACs are no longer learnt in the data plane on the DC GW but in the control plane through EVPN running on the I-component. Remote C-MACs coming from remote PEs are still learnt in the data plane. B-MACs in the B-component will be assigned and advertised following the procedures described in [PBB-EVPN].

A DCI-ESI will be configured on the DC GWs for multi-homing, but it will only be used in the EVPN control plane for the I-component EVI. No ESI will be used in the control plane of the B-component EVI as per [PBB-EVPN].

The rest of the control plane procedures will follow [EVPN] for the I-component EVI and [PBB-EVPN] for the B-component EVI.

From the data plane perspective, the I-component and B-component EVPN bindings established to the same far-end will be compared and the I-component EVPN-overlay binding will be kept down following the rules described in section 3.3.1.

### 3.5.2. Multi-homing procedures on the DC GWs

Single-active as well as all-active multi-homing MUST be supported.

The forwarding behavior of the DF and non-DF will be changed based on the description outlined in section 3.4.3, only replacing the "WAN split-horizon-group" for the B-component.

### 3.5.3. Impact on MAC Mobility procedures

C-MACs learnt from the B-component will be advertised in EVPN within the I-component EVI scope. If the C-MAC was previously known in the I-component database, EVPN would advertise the C-MAC with a higher sequence number, as per [EVPN]. From a Mobility perspective and the related procedures described in [EVPN], the C-MACs learnt from the B-component are considered local.

### 3.5.4. Data Center Gateway optimizations

All the considerations explained in section 3.4.5 are applicable to the PBB-EVPN DCI option.

## 5. Conventions and Terminology

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

BUM: it refers to the Broadcast, Unknown unicast and Multicast traffic

DF: Designated Forwarder

DC GW: Data Center Gateway

DCI: Data Center Interconnect

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

DCI-ESI: ESI defined on the DC GWs for multi-homing to/from the WAN

EVI: EVPN Instance

MAC-VRF: it refers to an EVI instance in a particular node

NVE: Network Virtualization Edge

TOR: Top-Of-Rack switch

VNI/VSID: refers to VXLAN/NVGRE virtual identifiers

## 6. Security Considerations

This section will be completed in future versions.

## 7. IANA Considerations

## 8. References

### 8.1. Normative References

[RFC4761]Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

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

[RFC6074]Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

### 8.2. Informative References

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-05.txt, work in progress, February, 2014

[PBB-EVPN] Sajassi et al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-06, work in progress, October, 2014

[EVPN-Overlays] 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

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

## 10. Authors' Addresses

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

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

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

Florin Balus  
Nuage Networks  
Email: florin@nuagenetworks.net

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



L2VPN Workgroup  
INTERNET-DRAFT  
Intended Status: Standards Track

Ali Sajassi  
Samer Salam  
Samir Thoria  
Cisco

Wim Henderickx  
Alcatel-Lucent

Yakov Rekhter  
John Drake  
Juniper

Florin Balus  
Nuage Networks

Lucy Yong  
Linda Dunbar  
Huawei

Expires: August 13, 2014

February 13, 2014

IP Inter-Subnet Forwarding in EVPN  
draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-03

Abstract

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

Status of this Memo

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

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

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

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

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

#### Copyright and License Notice

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

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

#### Table of Contents

1	Introduction . . . . .	4
1.1	Traditional Inter-Subnet Forwarding . . . . .	4
1.2	Scenarios of EVPN NVEs as L3GW . . . . .	4
2	Inter-Subnet Forwarding Scenarios . . . . .	5
2.1	Switching among EVIs within a DC . . . . .	6
2.2	Switching among EVIs in different DCs without route aggregation . . . . .	7
2.3	Switching among EVIs in different DCs with route aggregation . . . . .	7
2.4	Switching among IP-VPN sites and EVIs with route aggregation . . . . .	8
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 . . . . .	11
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	Among EVPN NVEs within a DC . . . . .	14
6	VM Mobility . . . . .	16
6.1	VM Mobility & Optimum Forwarding for VM's Outbound Traffic . . . . .	16
6.2	VM Mobility & Optimum Forwarding for VM's Inbound Traffic . . . . .	16

6.2.1	Mobility without Route Aggregation . . . . .	16
6.2.2	Mobility with Route Aggregation . . . . .	17
7	Acknowledgements . . . . .	17
8	Security Considerations . . . . .	17
9	IANA Considerations . . . . .	17
10	References . . . . .	17
10.1	Normative References . . . . .	17
10.2	Informative References . . . . .	17
	Authors' Addresses . . . . .	17

## Terminology

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

IRB: Integrated Routing and Bridging

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

NVE: Network Virtualization Endpoint

## 1 Introduction

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

### 1.1 Traditional Inter-Subnet Forwarding

The inter-subnet communication is traditionally achieved at the L3 Gateway nodes where all the inter-subnet communication policies are enforced. Even for different subnets belonging to one IP-VPN or tenant, traffic may need to go through FW or IPS between the trusted and un-trusted zones.

Some operators may prefer centralized approach, i.e. only have a set of default L3 gateways (whose redundancy is typically achieved by VRRP) for all inter-subnet traffic to go through. Usually there are FW, IPS, or other network appliances directly attached to the centralized L3 Gateway nodes. The centralized approach makes it easier for maintaining consistent policies and less prone to configuration errors. However, such centralized approach suffers from a major drawback of requiring all traffic to be hair-pinned to the L3GW nodes.

Some operators may prefer fully distributed L3 gateway design, e.g. allowing all NVEs to have the policies to route traffic across subnets. Under this design, all traffic between hosts attached to one NVE can be routed locally, thus avoiding traffic hair-pinning issue at the centralized L3GW. The perceived drawback of this fully distributed approach may be the extra effort required in maintaining policy consistence across all the NVEs.

Some operators may prefer somewhere in the middle, i.e. allowing NVEs to route traffic across only selected subnets. For example, allow NVEs to route traffic among subnets belonging to one tenant or one security zone.

### 1.2. Scenarios of EVPN NVEs as L3GW

When an EVPN NVE node is not the L3GW for the subnets attached, the EVPN NVE performs only L2 switching function for the traffic initiated from or destined to the hosts attached to the NVE.

Some EVPN NVEs can be the default L3GWs for some subnets. In this situation, the EVPN NVEs can route traffic across the subnets for which they are default L3GWs.

When there are multiple subnets attached to an EVPN NVE, some of the subnets could have the EVPN NVE as their L3GW, some other subnets don't have the NVE as their L3GW. For example: "Subnet-X" can communicate with "Subnet-Y" via NVE "A", but "Subnet-X" can't communicate with "Subnet-Z" via NVE "A". So when the "Subnet-X" needs to communicate with "Subnet-Z", the traffic might need to be routed through another device (e.g. FW, IPS, or another L3GW node).

1. When the EVPN NVE is the L3GW for "Subnet-X", hosts within "Subnet-X" will have the NVE's IRB MAC address (or NVE's MAC address) as their default GW MAC address when they send data frames towards targets in different subnets.

2. When the EVPN NVE is not the L3GW for "Subnet-Y", hosts within "Subnet-Y", (even though still attached to the NVE), will use their own designated L3GW MAC address (that is different from the NVE's IRB address) in data frames destined towards targets in different subnets.

## 2 Inter-Subnet Forwarding Scenarios

The inter-subnet forwarding scenarios performed by an EVPN NVE can be divided into the following five categories. The last scenario, along with their corresponding solutions, are described in [EVPN-IPVPN-INTEROP]. The solutions for the first four scenarios are the focus of this document.

1. Switching among EVPN instances (subnets) within a DC
2. Switching among EVPN instances in different DCs without route aggregation
3. Switching among EVPN instances in different DCs with route aggregation
4. Switching among IP-VPN sites and EVPN instances with route aggregation
5. Switching among IP-VPN sites and EVPN instances without route aggregation

In the above scenario, the term "route aggregation" refers to the

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

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

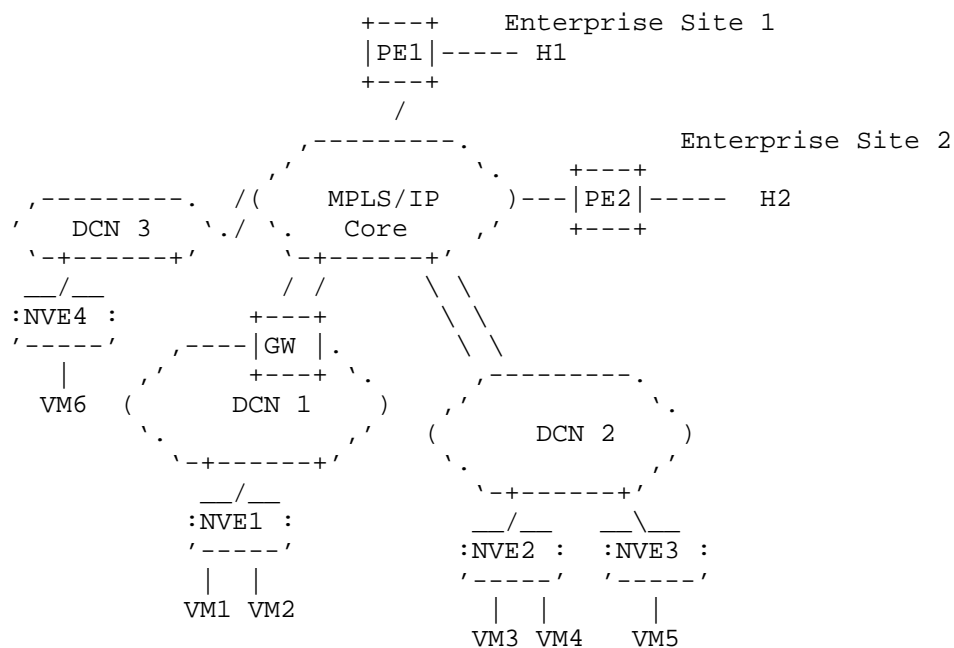


Figure 2: Interoperability Use-Cases

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

## 2.1 Switching among EVIs within a DC

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

As an example, consider VM3 and VM5 of Figure 2 above. Assume that connectivity is required between these two VMs where VM3 belongs to the IP-subnet 3 (SN3) whereas VM5 belongs to the IP-subnet 5 (SN5). Both SN3 and SN5 subnets 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 an 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, as discussed in [DC-MOBILITY]:

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



address correspond to the IRB interface of the EVI associated with that EVPN instance.

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

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

### 3.1 Heterogeneous Environment

For large data centers with thousands of servers and ToR (or Access) switches, some of them may not have the capability of maintaining or enforcing policies for inter-subnet switching. Even though policies among multiple subnets belonging to same tenant can be simpler, hosts belonging to one tenant can also send traffic to peers belonging to different tenants or security zones. A L3GW not only needs to enforce policies for communication among subnets belonging to a single tenant, but also it needs to know how to handle traffic destined towards peers in different tenants. Therefore, there can be a mixed environment where an NVE performs inter-subnet switching for some EVPN instances but not others.

## 4 Operational Models for 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.

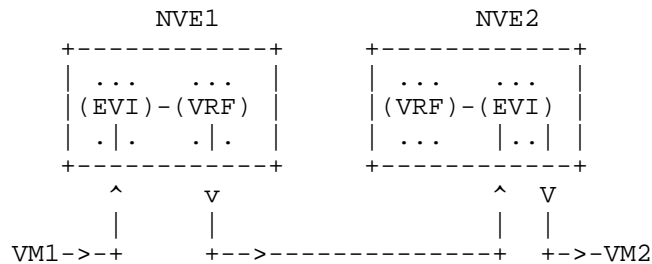


Figure 2: Inter-Subnet Forwarding Among EVPN NVEs within a DC

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.

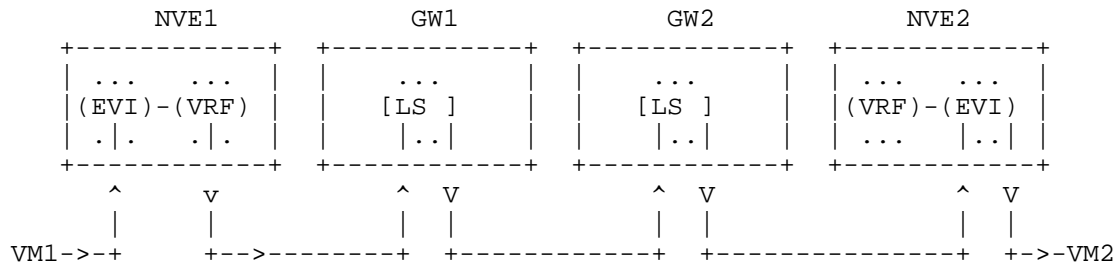


Figure 3: Inter-Subnet Forwarding Among EVPN NVEs in Different DCs without Route Aggregation

#### 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.

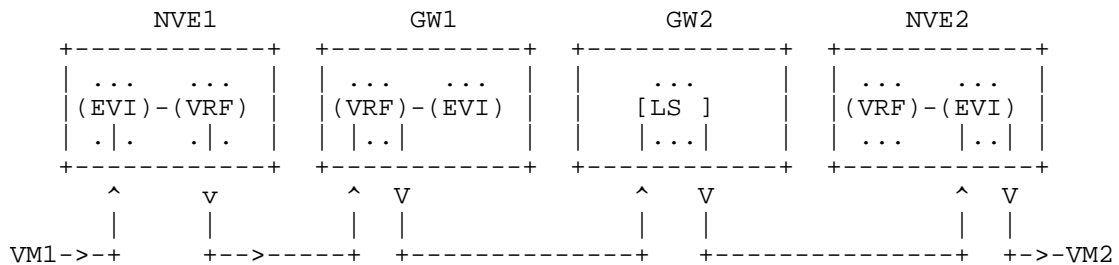


Figure 4: Inter-Subnet Forwarding Among EVPN NVEs in Different DCs with Route Aggregation

#### 4.4 Among IP-VPN Sites and EVPN NVEs with Route Aggregation

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

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 local WAN gateway after encapsulating it with the MPLS label stack. Note that this label

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

Figure 5 below depicts the forwarding model.

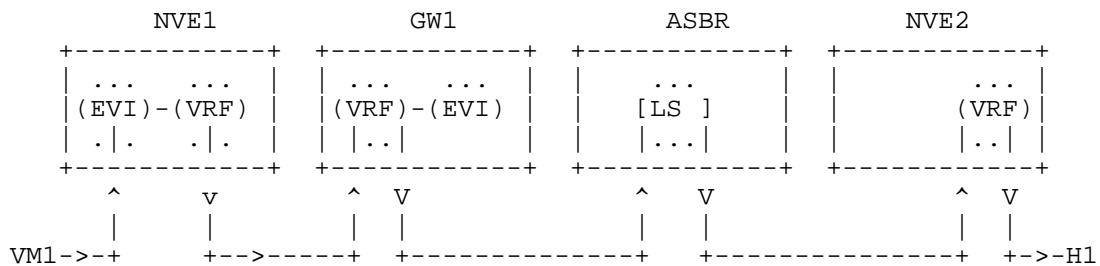


Figure 5: Inter-Subnet Forwarding Among IP-VPN Sites and EVPN NVEs with Route Aggregation

#### 4.5 Use of Centralized Gateway

In this scenario, the NVEs within a given data center need to forward traffic in L2 to a centralized L3GW for a number of reasons: a) they don't have IRB capabilities or b) they don't have required policy for switching traffic between different tenants or security zones. The centralized L3GW performs both the IRB function for switching traffic among different EVPN instances as well as it performs interworking function when the traffic needs to be switched between IP-VPN sites and EVPN instances.

## 5 Operational Models for Symmetric Inter-Subnet Forwarding

### 5.1 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. However, the received MAC address is not used to populate the adjacency associated with the IP route in the IP-VRF table, instead, the remote NVE's MAC address is used for this purpose.

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 MAC-VRF table. 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 next-hop NVE (egress 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 IP-VPN label (labeled in the MAC route) that was advertised by the egress NVE. When the MPLS encapsulated packet is received by the egress NVE, it uses the IP-VPN label to identify the IP-VRF table. It then performs an IP lookup in that table, which yields the outbound IRB interface to which the Ethernet frame must be forwarded. Next, a MAC lookup is performed on the destination MAC address of the frame in the MAC-VRF 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.

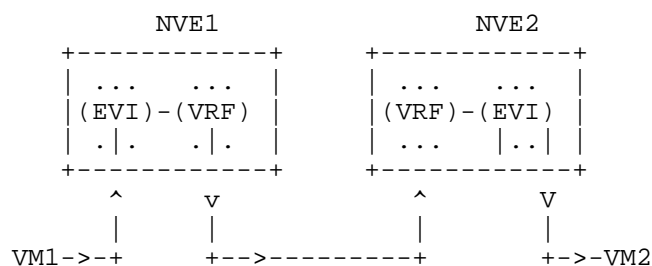


Figure 2: Inter-Subnet Forwarding Among EVPN NVEs within a DC

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.

## 6 VM Mobility

### 6.1 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].

### 6.2 VM Mobility & Optimum Forwarding for VM's Inbound Traffic

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

In what follows, we use the following terminology:

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

#### 6.2.1 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].

#### 6.2.2 Mobility with Route Aggregation

This section will be completed in the next revision.

### 7 Acknowledgements

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

### 8 Security Considerations

### 9 IANA Considerations

### 10 References

#### 10.1 Normative References

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

#### 10.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt, work in progress, July, 2014.

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

[DC-MOBILITY] Aggarwal et al., "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-05.txt, work in progress, June, 2013.

#### Authors' Addresses

Ali Sajassi  
Cisco

Email: sajassi@cisco.com

Samer Salam  
Cisco  
Email: ssalam@cisco.com

Yakov Rekhter  
Juniper Networks  
Email: yakov@juniper.net

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

Lucy Yong  
Huawei Technologies  
Email: lucy.yong@huawei.com

Linda Dunbar  
Huawei Technologies  
Email: linda.dunbar@huawei.com

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

Florin Balus  
Alcatel-Lucent  
Email: Florin.Balus@alcatel-lucent.com

Samir Thoria  
Cisco  
Email: sthoria@cisco.com

INTERNET-DRAFT  
Intended Status: Informational

Mingui Zhang  
Bin Wang  
Liang Xia  
Huawei  
Jie Hu  
China Telecom  
February 14, 2014

Expires: August 18, 2014

Tagging Customer Bridge Domains in VPLS  
draft-zhang-l2vpn-vpls-bd-tagging-01.txt

## Abstract

This document proposes to use Customer VLAN ID as a identifier for traffic isolation in Virtual Private LAN Service (VPLS). In this way, multiple bridge domains of customers can share a single VPLS instance while their traffic are separated. With this proposal, Service Providers can be relieved from the heavy provisioning overhead of large number of pseudowires in the environment where a mass of bridge domains need be connected.

## 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
2. Acronyms and Terminology . . . . .	3
2.1. Acronyms . . . . .	3
2.2. Terminology . . . . .	3
3. PE Model . . . . .	3
4. Use Cases of U-tag Awareness in VPLS . . . . .	4
4.1. No Duplicated MAC Address . . . . .	4
4.2. Scalable Interconnection of L2 Sites . . . . .	5
4.3. BUM Traffic Scoped per BD . . . . .	5
4.3.1. Advertising Interested VLANs in LDP . . . . .	5
4.3.2. Dynamic VLAN Registration with MVRP . . . . .	5
4.4. Per C-VLAN MAC Withdraw . . . . .	6
5. Backward Compatibility . . . . .	6
6. Contributors . . . . .	6
7. Security Considerations . . . . .	6
8. IANA Considerations . . . . .	6
9. References . . . . .	7
9.1. Normative References . . . . .	7
9.2. Informative References . . . . .	7
Author's Addresses . . . . .	8

## 1. Introduction

VPLS has been widely used to connect customers' bridge domains. Traffic segregation for customers is performed on a per VPLS instance basis. In the environment (e.g., Data Center Network) where a mass of customers multiplied with a plenty of bridge domains are to be connected, a large number of PWs need be maintained. Service Providers are therefore suffering from scalability issue.

This proposal suggests the Customer VLAN ID (U-tag) is used as an additional de-multiplexor for traffic segregation in VPLS. By doing this, multiple BDs can share the same VPLS instance while their traffic are isolated. This method can greatly reduce the number of PWs therefore reduce the provisioning overhead for operators. Use cases of this method are given in the document.

Two options arising from the industry are covered in the discussion. The first one is proposed in [V-aware]. It extends the LDP control plane for PEs to advertise supported VLANs. The second option makes use of VLAN registration protocol, such as [MVRP], to exchange supported C-VLANs between PEs.

## 2. Acronyms and Terminology

### 2.1. Acronyms

MVRP: Multiple VLAN Registration Protocol  
BD: Bridge Domain/Broadcast Domain  
PW: Pseudowire  
VSI: Virtual Switch Instance  
U-tag: Customer VLAN ID  
C-VLAN: Customer VLAN  
BUM: Broadcast, Unknown unicast and Multicast  
VLL: Virtual Leased Line

### 2.2. Terminology

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

## 3. PE Model

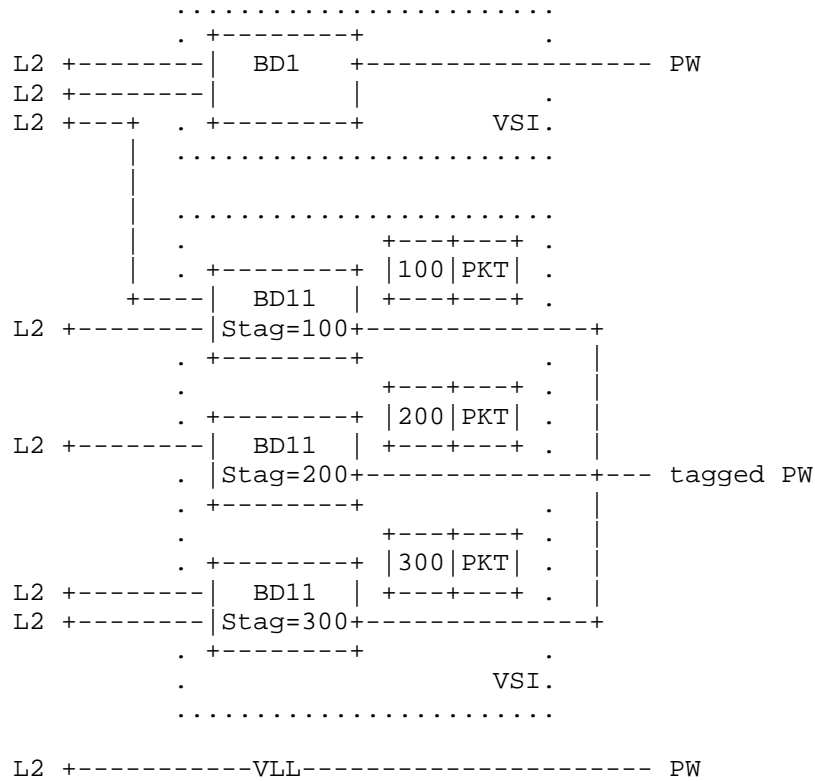


Figure 3.1: U-tag is used as the service de-multiplexor in tagged PW

In Figure 3.1, an example is used to show that service tag is used as a finer grained de-multiplexor for traffic segregation. Therefore, multiple BDs can be integrated into one VSI.

#### 4. Use Cases of U-tag Awareness in VPLS

##### 4.1. No Duplicated MAC Address

One MAC address might be used by multiple hosts in different customer VLANs (C-VLAN). This is illegal but it is the headache reality for providers. In the virtualization environment, Virtual Machines (VM) are more likely to have duplicated MAC addresses. When these hosts/VMs join in the same VSI of a PE, the PE will see MAC address duplication. In order to overcome this issue, the PE has to adopt qualified learning [RFC4762], i.e., the PE has to set up one VSI per C-VLAN. This brings the scalability issue as discussed in Section 4.2.

If the PE uses U-tag as the de-multiplexor to isolate traffic of customers' BDs, above MAC address duplication issue can be avoided.

#### 4.2. Scalable Interconnection of L2 Sites

For the qualified learning, providers need set up one PW per C-VLAN. When there is a large number of customers multiplied by C-VLANs interconnected using VPLS, a mass of PWs need be maintained. It brings heavy operating overhead to providers.

In this document, U-tag is used to distinguish BDs in VPLS. In this way, traffic from multiple C-VLANs can be handled by a single VPLS. As shown in Figure 3.1, one PW is set up for each VSI and this VSI may be an integration of multiple BDs. Operating overhead of operators can be greatly reduced.

#### 4.3. BUM Traffic Scoped per BD

Traditional VPLS limits a broadcast domain scope per PW. Suppose a customer has four sites in New York, Chicago, Atlanta and Dallas. BD1 = {New York, Chicago and Atlanta} while BD2 = {New York, Chicago and Dallas}. If one VSI per PE is set up to interconnect these four sites. BUM traffic of Atlanta site will be poured to Dallas site, and vice versa.

When PEs are aware of the U-tag, the BUM traffic can be confined per BD with multicast pruning. For above example, the operator need use two U-tags to distinguish the two BDs. In this way, BUM traffic of Atlanta site will be confined in BD1 and BUM traffic for Dallas site will be confined in BD2. This increases the efficiency of the bandwidth utilization of BUM traffic.

Two C-VLAN based multicast pruning techniques are listed below. (One is give in [V-aware] the other has been implemented by vendors.)

##### 4.3.1. Advertising Interested VLANs in LDP

With the PW VLAN Vector TLV defined in [V-aware], PEs can advertise in LDP the interested C-VLANs for its interfaces. In this way, PEs can prune the flooding on a per C-VLAN basis.

##### 4.3.2. Dynamic VLAN Registration with MVRP

It requires Multiple VLAN Registration Protocol (MVRP) to be supported by PEs for U-tag registration on the interfaces providing VPLS. With the help of MVRP, operators need not manually configure C-VLANs on PEs.

Only when a C-VLAN is registered in both directions of a PW, this PW will not be eliminated for this C-VLAN. Otherwise, this PW will be pruned for this C-VLAN. Multicast frames for a C-VLAN SHOULD only be forwarded on PWs that are not pruned for this C-VLAN.

#### 4.4. Per C-VLAN MAC Withdraw

With the awareness of U-tag, PEs can achieve a finer grained C-VLAN scoped MAC withdraw. For example, with the VLAN Vector TLV defined in [V-aware], a PE can specify VLANs that it wants their MAC address to be flushed.

#### 5. Backward Compatibility

Two PEs need negotiate their capability on supporting the awareness of U-tag. Unless both PEs are aware of U-tag, the tagged PW cannot be established. When a PE realizes the peering PE's interface is unaware of U-tag, it MUST fall back to establish a raw PW with this interface.

There are two ways to achieve the capability negotiation.

- a) As defined in Section 4 of [V-aware], PEs can negotiate this capability through LDP using the VLAN Aware Capability TLV.
- b) A tagged PW is established between two interfaces if they both enable MVRP.

For the tagged PW, PEs can achieve customer VLAN scoped MAC address flushing [V-aware]. However, PEs may as well send out the old type MAC withdraw message per Section 6.2 of [RFC4762]. The receiver PE parses this kind of message as that the peering PE is flushing MAC addresses across all customer VLANs supported by this PW.

#### 6. Contributors

Xingjian He, Huawei

#### 7. Security Considerations

This document raises no new security issues. For general security considerations, refer to [RFC4761] and [RFC4762].

#### 8. IANA Considerations

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



## 9. References

### 9.1. Normative References

- [V-aware] D. Cai, S. Boutros, and et al, "VLAN Aware VPLS services", draft-cai-l2vpn-vpls-vlan-aware-bundling-00.txt, working in progress.
- [MVRP] IEEE P802.1ak/D8.0, "IEEE Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks -- Amendment 07: Multiple Registration Protocol", November 29, 2006.
- [RFC4762] Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

### 9.2. Informative 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.

Author's Addresses

Mingui Zhang  
Huawei Technologies

Email: zhangmingui@huawei.com

Bin Wang  
Huawei Technologies

Email: wb.wangbin@huawei.com

Liang Xia  
Huawei Technologies

Email: frank.xialiang@huawei.com

Jie Hu  
China Telecom

Email: hujie@ctbri.com.cn