

BESS Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: May 7, 2020

P. Brissette, Ed.  
A. Sajassi  
L. Burdet  
Cisco Systems  
D. Voyer  
Bell Canada  
November 4, 2019

EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols  
draft-brissette-bess-evpn-l2gw-proto-05

Abstract

The existing EVPN multi-homing load-balancing modes defined are Single-Active and All-Active. Neither of these multi-homing mechanisms are appropriate to support access networks with Layer-2 Gateway protocols such as G.8032, MPLS-TP, STP, etc. These Layer-2 Gateway protocols require a new multi-homing mechanism defined in this draft.

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 <https://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 May 7, 2020.

Copyright Notice

Copyright (c) 2019 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

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

## Table of Contents

1. Introduction . . . . .	2
1.1. Requirements Language . . . . .	3
1.2. Terms and Abbreviations . . . . .	3
2. Solution . . . . .	3
2.1. Single-Flow-Active redundancy mode . . . . .	4
2.2. Backwards compatibility . . . . .	5
2.2.1. The two-ESI solution . . . . .	5
2.2.2. RFC7432 Remote PE . . . . .	6
3. Requirements . . . . .	6
4. Handling of Topology Change Notification (TCN) . . . . .	7
5. ESI-label Extended Community Extension . . . . .	9
6. EVPN MAC-Flush Extended Community . . . . .	9
7. EVPN Inter-subnet Forwarding . . . . .	10
8. Conclusion . . . . .	10
9. Security Considerations . . . . .	10
10. Acknowledgements . . . . .	11
11. IANA Considerations . . . . .	11
12. References . . . . .	11
12.1. Normative References . . . . .	11
12.2. Informative References . . . . .	11
Authors' Addresses . . . . .	11

## 1. Introduction

Existing EVPN multi-homing mechanisms of Single-Active and All-Active are not sufficient to support access Layer-2 Gateway protocols such as G.8032, MPLS-TP, STP, etc.

These Layer-2 Gateway protocols require that a given flow of a VLAN (represented by {MAC-SA, MAC-DA}) to be only active on one of the PEs in the multi-homing group. This is in contrast with Single-Active redundancy mode where all flows of a VLAN are active on one of the multi-homing PEs and it is also in contrast with All-Active redundancy mode where all L2 flows of a VLAN are active on all PEs in the redundancy group.

This draft defines a new multi-homing mechanism "Single-Flow-Active" which defines that a VLAN can be active on all PEs in the redundancy group but a single given flow of that VLAN can be active on only one of the PEs in the redundancy group. In fact, the carving scheme, performed by the DF (Designated Forwarder) election algorithm for

these L2 Gateway protocols, is not per VLAN but rather for a given VLAN. A selected PE in the redundancy group can be the only Designated Forwarder for a specific L2 flow but the decision is not taken by the PE. The loop-prevention blocking scheme occurs in the access network.

EVPN multi-homing procedures need to be enhanced to support Designated Forwarder election for all traffic (both known unicast and BUM) on a per L2 flow basis. This new multi-homing mechanism also requires new EVPN considerations for aliasing, mass-withdraw, fast-switchover and [EVPN-IRB] as described in the solution section.

### 1.1. Requirements Language

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

### 1.2. Terms and Abbreviations

AC:	Attachment Circuit
BUM:	Broadcast, Unknown unicast, Multicast
DF:	Designated Forwarder
GW:	Gateway
L2 Flow:	A given flow of a VLAN, represented by (MAC-SA, MAC-DA)
L2GW:	Layer-2 Gateway
G.8032:	Ethernet Ring Protection
MST-AG:	Multi-Spanning Tree Access Gateway
REP-AG:	Resilient Ethernet Protocol Access Gateway
TCN:	Topology Change Notification

## 2. Solution

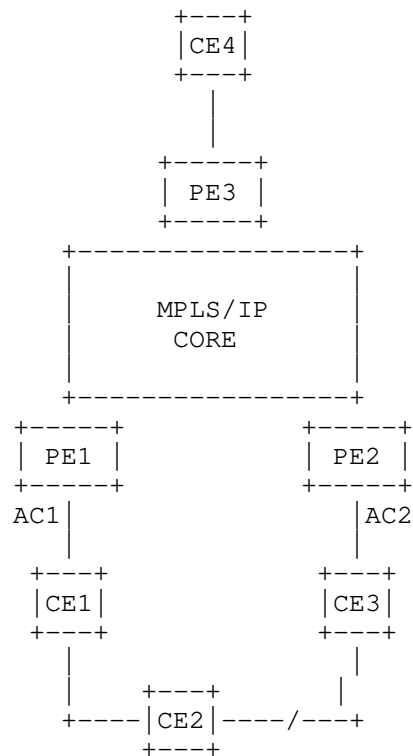


Figure 1: EVPN network with L2 access GW protocols

Figure 1 shows a typical EVPN network with an access network running a L2GW protocol, typically one of the following: G.8032, STP, MPLS-TP, etc. The L2GW protocol usually starts from AC1 (on PE1) up to AC2 (on PE2) in an open "ring" manner. AC1 and AC2 interfaces of PE1 and PE2 are participants in the access protocol.

The L2GW protocol is used for loop avoidance. In above example, the loop is broken on the right side of CE2.

### 2.1. Single-Flow-Active redundancy mode

PE1 and PE2 are peering PEs in a redundancy group, and sharing a same ESI. In the proposed Single-Flow-Active mode, PE1 and PE2 'Access Gateway' load-balancing mode shares similarities with both Single-Active and All-Active. DF election must not result in blocked ports or portions of the access may become isolated. Additionally, the reachability between CE1/CE2 and CE3 is achieved with the forwarding path through the EVPN MPLS/IP core side. Thus, the ESI-Label

filtering of [RFC7432] is disabled for Single-Flow-Active Ethernet segments.

Finally, PE3 behaves according to EVPN rules for traffic to/from PE1/PE2. Peering PE, selected per L2 flow, is chosen by the L2GW protocol in the access, and is out of EVPN control.

From PE3 point of view, some of the L2 flows coming from PE3 may reach CE3 via PE2 and some of the L2 flows may reach CE1/CE2 via PE1. A specific L2 flow never goes to both peering PEs. Therefore, aliasing cannot be performed by PE3. That node operates in a single-active fashion for each of these L2 flows.

The backup path which is also setup for rapid convergence, is not applicable here. For example, in Figure 1, if a failure happens between CE1 and CE2, L2 flows coming from CE4 behind PE3 destined to CE1 still goes through PE1 and shall not switch to PE2 as a backup path. On PE3, there is no way to know which L2 flow specifically is affected. During the transition time, PE3 may flood until unicast traffic recovers properly.

## 2.2. Backwards compatibility

### 2.2.1. The two-ESI solution

As background, an alternative solution which achieves some, but not all, of the requirements exists and is backwards compatible with [RFC7432]:

On the PE1 and PE2,

- a. A single-homed (different) non-zero ESI, or zero-ESI, is used for each PE;
- b. With no remote Ethernet-Segment routes received matching local ESI, each PE will be designated forwarder for all the local VLANs;
- c. Each L2GW PE will send Ethernet AD per-ES and per-EVI routes for its ESI if non-zero; and
- d. When the L2GW PEs receive a MAC-Flush notification (STP TCN, G.8032 mac-flush, LDP MAC withdrawal etc.), they send an update of the Ethernet AD per-EVI route with the MAC Mobility extended community defined in Section 6 and a higher sequence number.

While this solution is feasible, it is considered to fall short of the requirements listed in Section 3, namely for all aspects meant to achieve fast-convergence.

#### 2.2.2. RFC7432 Remote PE

A PE which receives an Ethernet AD per ES route with the Single-Flow-Active bit set in the ESI-flags, and which does not support/understand this bit, SHALL discard the bit and continue operating per [RFC7432] (All-Active). The operator should understand the usage of single-flow-active load-balancing mode else it is highly recommended to use the two-ESI approach as described in section 2.2.1.

The remote PE3 which does not support Single-Flow-Active redundancy mode as described, will ECMP traffic to peering PEs PE1 and PE2 in the example topology above (Figure 1), per [RFC7432], Section 8.4 aliasing and load-balancing rules. PE1 and PE2, which support the Single-Flow-Active redundancy mode MUST setup sub-optimal Layer-2 forwarding and sub-optimal Layer-3 routing towards the PE at which the flow is currently active.

Thus, while PE3 is ECMP (on average) 50% of the traffic to the incorrect PE in [RFC7432] operation, PE1 and PE2 will handle this gracefully in Single-Flow-Active mode and redirect across peering pair of PEs appropriately.

No extra route or information is required for this. The [RFC7432] and [EVPN-IRB] route advertisements are sufficient.

### 3. Requirements

The EVPN L2GW framework for L2GW protocols in Access-Gateway mode, consists of the following rules:

- o Peering PEs MUST share the same ESI.
- o The Ethernet-Segment DF election MUST NOT be performed and forwarding state MUST be dictated by the L2GW protocol. In Access Gateway mode, both PEs are usually in forwarding state. In fact, access protocol is responsible for operationally setting the forwarding state for each VLAN.
- o Split-horizon filtering is NOT needed because L2GW protocol ensures there will never be loop in the access network. The forwarding between peering PEs MUST also be preserved. In figure 1, CE1/CE2 device may need reachability with CE3 device. ESI-filtering capability MUST be disabled. PE MUST NOT advertise

corresponding ESI-label to other PEs in the redundancy group, or apply it if it is received.

- o ESI-label BGP-extcomm MUST support a new multi-homing mode named "Single-Flow-Active" corresponding to the single-active behaviour of [RFC7432], applied per flow.
- o Upon receiving ESI-label BGP-Extcomm with the single-flow-active load-balancing mode, remote PE MUST:

- \* Disable ESI-Label processing

- \* Disable aliasing (at Layer-2 and Layer-3 [EVPN-IRB])

- o The Ethernet-Segment procedures in the EVPN core such as Ethernet AD per-ES and per Ethernet AD per-EVI routes advertisement/withdraw, as well as MAC and MAC+IP advertisement, remains as explained in [RFC7432] and [EVPN-IRB].
- o For fast-convergence, remote PE3 MAY set up two distinct backup paths on a per-flow basis:
  - \* { PE1 active, PE2 backup }
  - \* { PE2 active, PE1 backup }

The backup paths so created, operate as in [RFC7432] section 8.4 where the backup PE of the redundancy group MAY immediately be selected for forwarding upon detection of a specific subset of failures: Ethernet AD per-ES route withdraw, Active PE loss of reachability (via IGP detection). An Ethernet AD per-EVI withdraw MUST NOT result in automatic switching to the backup PE as only a subset of the hosts may be changing reachability to the Backup PE, and the remote cannot determine which.

- o MAC mobility procedures SHALL have precedence in Single-Flow-Active for tracking host reachability over backup path procedure.

#### 4. Handling of Topology Change Notification (TCN)

In order to address rapid Layer-2 convergence requirement, topology change notification received from the L2GW protocols must be sent across the EVPN network to perform the equivalent of legacy L2VPN remote MAC flush.

The generation of TCN is done differently based on the access protocol. In the case of STP (REP-AG) and G.8032, TCN gets generated in both directions and thus both of the dual-homing PEs receive it.

However, with STP (MST-AG), TCN gets generated only in one direction and thus only a single PE can receive it. That TCN is propagated to the other peering PE for local MAC flushing, and relaying back into the access.

In fact, PEs have no direct visibility on failures happening in the access network neither on the impact of those failures over the connectivity between CE devices. Hence, both peering PEs require to perform a local MAC flush on corresponding interfaces.

There are two options to relay the access protocol's TCN to the peering PE: in-band or out-of-band messaging. The first method is better for rapid convergence, and requires a dedicated channel between peering PEs. An EVPN-VPWS connection MAY be dedicated for that purpose, connecting the Untagged ACs of both PEs. The latter choice relies on a new MAC flush extended community in the Ethernet Auto-discovery per EVI route, defined below. It is a slower method but has the advantage of avoid the usage of a dedicated channel between peering PEs.

Peering PE, upon receiving TCN from access, MUST:

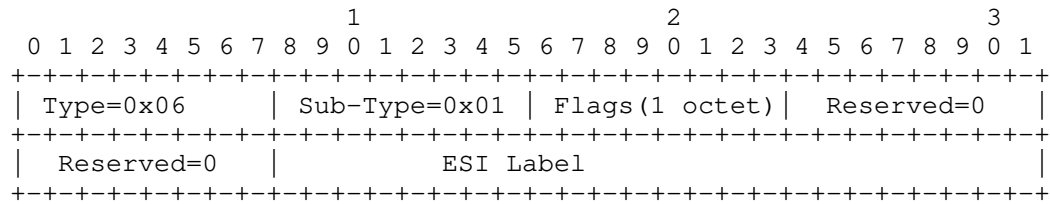
- o As per legacy VPLS, perform a local MAC flush on the access-facing interfaces. An ARP probe is also sent for all hosts previously locally-attached.
- o Advertise per EVI/EAD route along with a new MAC-flush BGP Extended Community in order to perform a remote MAC flush and steer L2 traffic to proper peering PE. The sequence number is incremented by one as a flushing indication to remote PEs.
- o Ensure MAC and MAC/IP route re-advertisement, with incremented sequence number when host reachability is NOT moving to peering PE. This is to ensure a re-advertisement of current MAC and MAC/IP which may have been flushed remotely upon MAC Flush extcomm reception. In theory, it should happen automatically since peering PE, receiving TCN from the access, performs local MAC flush on corresponding interface and will re-learn that local MAC or MAC/IP at ARP probe reply.
- o Where an access protocol relies on TCN BPDUs propagation to all participant nodes, a dedicated EVPN-VPWS connection MAY be used as an in-band channel to relay TCN between peering PEs. That connection may be auto-generated or can simply be directly configured by user.



## 5. ESI-label Extended Community Extension

In order to support the new EVPN load-balancing mode (single-flow-active), the ESI-label extended community is updated.

The 1 octet flag field, part of the ESI Label extended community, is modified as follows:



Low-order bit: [7:0]

[2:0]- 000 = all-active,  
           001 = single-active,  
           010 = single-flow-active,  
           others = unassigned  
 [7:3]- Reserved

Figure 2: ESI Label extended community

## 6. EVPN MAC-Flush Extended Community

The MAC mobility BGP Extended community, is required for the TCN procedures and MAC-Flushing. The well-known MAC-Flush procedure from [RFC7623] is borrowed, only for Ethernet AD per-EVI routes.

In this Single-Flow-Active mode, the MAC-Flush Extended Community is advertised along with Ethernet AD per EVI routes upon reception of TCN from the access. When this extended community is used, it indicates, to all remote PEs that all MAC addresses associated with that EVI/ESI are "flushed" i.e. unresolved. They remain unresolved until remote PE receives a route update / withdraw for those MAC addresses; the MAC may be re-advertised by the same PE, or by another, in the same ESI.

The sequence number used is of local significance from the originating PE, and is not used for comparison between peering PEs. Rather, it is used to signal via BGP successive MAC Flush requests from a given PE.

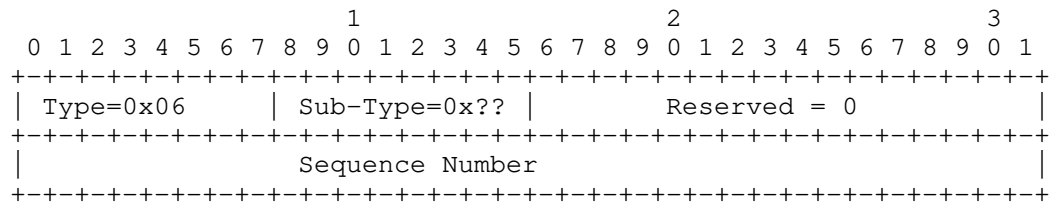


Figure 3: MAC-Flush Extended Community

## 7. EVPN Inter-subnet Forwarding

EVPN Inter-subnet forwarding procedures in [EVPN-IRB] works with the current proposal and does not require any extension. Host routes continue to be installed at PE3 with a single remote nexthop, no aliasing.

However, leveraging the same-ESI on both L2GW PEs enables ARP/ND synchronization procedures which are defined for All-Active redundancy in [EVPN-IRB]. In steady-state, on PE2 where a host is not locally-reachable the routing table will reflect PE1 as the destination. However, with ARP/ND synchronization based on a common ESI, the ARP/ND cache may be pre-populated with the local AC as destination for the host, should an AC failure occur on PE1. This achieves fast-convergence.

When a hosts moves to PE2 from the PE1 L2GW peer, the MAC mobility sequence number is incremented to signal to remote peers that a 'move' has occurred and the routing tables must be updated to PE2. This is required when an Access Protocol is running where the loop is broken between two CEs in the access and the L2GWs, and the host is no longer reachable from the PE1-side but now from the PE2-side of the access network.

## 8. Conclusion

EVPN style="symbols"Multi-Homing Mechanism for Layer-2 gateway Protocols solves a true problem due to the wide legacy deployment of these access L2GW protocols in Service Provider networks. The current draft has the main advantage to be fully compliant with [RFC7432] and [EVPN-IRB].

## 9. Security Considerations

The same Security Considerations described in [RFC7432] and [EVPN-IRB] remain valid for this document.

## 10. Acknowledgements

Authors would like to thank Thierry Couture for valuable review and inputs with respect to access protocol deployments related to procedures proposed in this document.

## 11. IANA Considerations

A new allocation of Extended Community Sub-Type for EVPN is required to support the new EVPN MAC flush mechanism..

## 12. References

### 12.1. Normative References

- [EVPN-IRB] Sajassi, A., "Integrated Routing and Bridging in EVPN", 2019.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <<https://www.rfc-editor.org/info/rfc7623>>.

### 12.2. Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

## Authors' Addresses

Patrice Brissette (editor)  
Cisco Systems  
Ottawa, ON  
Canada

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

Ali Sajassi  
Cisco Systems  
USA

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

Luc Andre Burdet  
Cisco Systems  
Ottawa, ON  
Canada

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

Daniel Voyer  
Bell Canada  
Montreal, QC  
Canada

Email: [daniel.voyer@bell.ca](mailto:daniel.voyer@bell.ca)

BESS Working Group  
INTERNET-DRAFT  
Intended Status: Proposed Standard

Patrice Brissette  
Ali Sajassi  
Cisco Systems

Bin Wen  
Comcast

Edward Leyton  
Verizon Wireless

Jorge Rabadan  
Nokia

Expires: May 3, 2020

October 31, 2019

EVPN multi-homing port-active load-balancing  
draft-brissette-bess-evpn-mh-pa-04

#### Abstract

The Multi-Chassis Link Aggregation Group (MC-LAG) technology enables the establishment of a logical link-aggregation connection with a redundant group of independent nodes. The purpose of multi-chassis LAG is to provide a solution to achieve higher network availability, while providing different modes of sharing/balancing of traffic. EVPN standard defines EVPN based MC-LAG with single-active and all-active multi-homing load-balancing mode. The current draft expands on existing redundancy mechanisms supported by EVPN and introduces support of port-active load-balancing mode. In the current document, port-active load-balancing mode is also referred to as per interface active/standby.

#### 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) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1	Introduction	3
1.1	Terminology	4
2.	Multi-Chassis Ethernet Bundles	4
3.	Port-active load-balancing procedure	4
4.	Algorithm to elect per port-active PE	5
4.1	Capability Flag	5
4.2	Modulo-based Designated Forwarder Algorithm	6
4.3	HRW Algorithm	6
4.4	Preferred-DF Algorithm	6
5.	Convergence considerations	6
6.	Applicability	7
7.	Overall Advantages	7
8	Security Considerations	8
9	IANA Considerations	8
10	References	8
10.1	Normative References	8
10.2	Informative References	8
	Authors' Addresses	9

## 1 Introduction

EVPN, as per [RFC7432], provides all-active per flow load balancing for multi-homing. It also defines single-active with service carving mode, where one of the PEs, in redundancy relationship, is active per service.

While these two multi-homing scenarios are most widely utilized in data center and service provider access networks, there are scenarios where active-standby per interface multi-homing redundancy is useful and required. The main consideration for this mode of redundancy is the determinism of traffic forwarding through a specific interface rather than statistical per flow load balancing across multiple PEs providing multi-homing. The determinism provided by active-standby per interface is also required for certain QOS features to work. While using this mode, customers also expect minimized convergence during failures. A new term of load-balancing mode "port-active load-balancing" is then defined.

This draft describes how that new redundancy mode can be supported via EVPN.

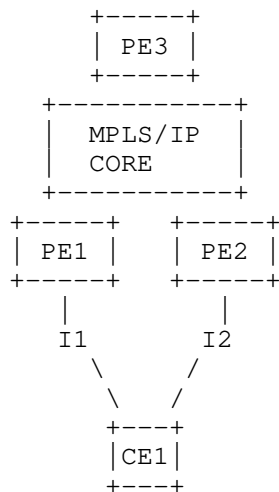


Figure 1. MC-LAG topology

Figure 1 shows a MC-LAG multi-homing topology where PE1 and PE2 are part of the same redundancy group providing multi-homing to CE1 via interfaces I1 and I2. Interfaces I1 and I2 are Bundle-Ethernet interfaces running LACP protocol. The core, shown as IP or MPLS enabled, provides wide range of L2 and L3 services. MC-LAG multi-homing functionality is decoupled from those services in the core and

it focuses on providing multi-homing to CE. With per-port active/standby redundancy, only one of the two interface I1 or I2 would be in forwarding, the other interface will be in standby. This also implies that all services on the active interface are in active mode and all services on the standby interface operate in standby mode.

### 1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Multi-Chassis Ethernet Bundles

When a CE is multi-homed to a set of PE nodes using the [802.1AX] Link Aggregation Control Protocol (LACP), the PEs must act as if they were a single LACP speaker for the Ethernet links to form a bundle, and operate as a Link Aggregation Group (LAG). To achieve this, the PEs connected to the same multi-homed CE must synchronize LACP configuration and operational data among them. InterChassis Communicated-based Protocol (ICCP) has been used for that purpose. EVPN LAG simplifies greatly that solution. Along with the simplification comes few assumptions:

- CE device connected to Multi-homing PEs may has a single LAG with all its active links i.e. Links in the Ethernet Bundle operate in all-active load-balancing mode.
- Same LACP parameters MUST be configured on peering PEs such as system id, port priority and port key.

Any discrepancies from this list is left for future study. Furthermore, mis-configuration and mis-wiring detection across peering PEs are also left for further study.

## 3. Port-active load-balancing procedure

Following steps describe the proposed procedure with EVPN LAG to support port-active load-balancing mode:

- 1- The Ethernet-Segment Identifier (ESI) MUST be assigned per access interface as described in [RFC7432], which may be auto derived or manually assigned. Access interface MAY be a Layer-2 or Layer3 interface. The usage of ESI over L3 interfce is newly described in this document.



2- Ethernet-Segment MUST be configured in port-active load-balancing mode on peering PEs for specific access interface

3- Peering PEs MAY exchange only Ethernet-Segment route (Route Type-4) when ESI is configured on a Layer3 interface.

4- PEs in the redundancy group leverage the DF election defined in [RFC8584] to determine which PE keeps the port in active mode and which one(s) keep it in standby mode. While the DF election defined in [RFC8584] is per <ES, Ethernet Tag> granularity, for port-active mode of multi-homing, the DF election is done per <ES>. The details of this algorithm are described in Section 4.

5- DF router MUST keep corresponding access interface in up and forwarding active state for that Ethernet-Segment

6- Non-DF routers MAY bring and keep peering access interface attached to it in operational down state. If the interface is running LACP protocol, then the non-DF PE MAY also set the LACP state to OOS (Out of Sync) as opposed to interface state down. This allows for better convergence on standby to active transition.

7- For EVPN-VPWS service, the usage of primary/backup bits of EVPN Layer2 attributes extended community [RFC8214] is highly recommended to achieve better convergence.

#### 4. Algorithm to elect per port-active PE

The ES routes, running in port-active load-balancing mode, are advertised with a new capability in the DF Election Extended Community as defined in [RFC8584]. Moreover, the ES associated to the port leverages existing procedure of single-active, and signals single-active bit along with Ethernet-AD per-ES route. Finally, as in RFC7432, the ESI-label based split-horizon procedures should be used to avoid transient echo'ed packets when L2 circuits are involved.

##### 4.1 Capability Flag

[RFC8584] defines a DF Election extended community, and a Bitmap field to encode "capabilities" to use with the DF election algorithm in the DF algorithm field. Bitmap (2 octets) is extended by the following value:

```

          1 1 1 1 1 1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+---+
|D|A|       |P|                         |
+---+---+---+---+---+---+---+---+---+

```

Figure 2 - Amended Bitmap field in the DF Election Extended Community

- Bit 0: 'Don't Preempt' bit, as explained in [PREF-DF].
- Bit 1: AC-Influenced DF Election, as explained in [RFC8584].
- Bit 5: (corresponds to Bit 25 of the DF Election Extended Community and it is defined by this document):  
P bit or 'Port Mode' bit (P hereafter), determines that the DF-Algorithm should be modified to consider the port only and not the Ethernet Tags.

#### 4.2 Modulo-based Designated Forwarder Algorithm

The default DF Election algorithm, or modulus-based algorithm as in [RFC7432] and updated by [RFC8584], is used here, at the granularity of <ES> only. Given the fact, ES-Import RT community inherits from ESI only byte 1-7, many deployments differentiate ESI within these bytes only. For Modulo calculation, bytes [3-7] are used to determine the designated forwarder using Modulo-based DF assignment.

#### 4.3 HRW Algorithm

Highest Random Weight (HRW) algorithm defined in [RFC8584] MAY also be used and signaled, and modified to operate at the granularity of <ES> rather than per <ES, VLAN>.

[RFC8584] describes computing a 32 bit CRC over the concatenation of Ethernet Tag and ESI. For port-active load-balancing mode, the Ethernet Tag is simply removed from the CRC computation.

#### 4.4 Preferred-DF Algorithm

When the new capability 'Port-Mode' is signaled, the algorithm is modified to consider the port only and not any associated Ethernet Tags. Furthermore, the "port-based" capability MUST be compatible with the 'DP' capability (for non-revertive). The AC-DF bit MUST be set to zero. When an AC (sub-interface) goes down, it does not influence the DF election.

#### 5. Convergence considerations

To improve the convergence, upon failure and recovery, when port-active load-balancing mode is used, some advanced synchronization between peering PEs may be required. Port-active is challenging in a sense that the "standby" port is in down state. It takes some time to bring a "standby" port in up-state and settle the network. For IRB and L3 services, ARP / ND cache may be synchronized. Moreover,

associated VRF tables may also be synchronized. For L2 services, MAC table synchronization may be considered.

Finally, for Bundle-Ethernet interface where LACP is running the ability to set the "standby" port in "out-of-sync" state aka "warm-standby" can be leveraged.

## 6. Applicability

A common deployment is to provide L2 or L3 service on the PEs providing multi-homing. The services could be any L2 EVPN such as EVPN VPWS, EVPN [RFC7432], etc. L3 service could be in VPN context [RFC4364] or in global routing context. When a PE provides first hop routing, EVPN IRB could also be deployed on the PEs. The mechanism defined in this draft is used between the PEs providing the L2 and/or L3 service, when the requirement is to use per port active.

A possible alternate solution is the one described in this draft is MC-LAG with ICCP [RFC7275] active-standby redundancy. However, ICCP requires LDP to be enabled as a transport of ICCP messages. There are many scenarios where LDP is not required e.g. deployments with VXLAN or SRv6. The solution defined in this draft with EVPN does not mandate the need to use LDP or ICCP and is independent of the underlay encapsulation.

## 7. Overall Advantages

The use of port-active multi-homing brings the following benefits to EVPN networks:

- Open standards based per interface single-active redundancy mechanism that eliminates the need to run ICCP and LDP.
- Agnostic of underlay technology (MPLS, VXLAN, SRv6) and associated services (L2, L3, Bridging, E-LINE, etc).
- Provides a way to enable deterministic QOS over MC-LAG attachment circuits
- Fully compliant with [RFC7432], does not require any new protocol enhancement to existing EVPN RFCs.
- Can leverage various DF election algorithms e.g. modulo, HRW, etc.
- Replaces legacy MC-LAG ICCP-based solution, and offers following additional benefits:
  - Efficiently supports 1+N redundancy mode (with EVPN using BGP

RR) where as ICCP requires full mesh of LDP sessions among PEs in redundancy group

- Fast convergence with mass-withdraw is possible with EVPN, no equivalent in ICCP

- Customers want per interface single-active redundancy, but don't want to enable LDP (e.g. they may be running VXLAN or SRv6 in the network). Currently there is no alternative to this.

## 8 Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

## 9 IANA Considerations

This document solicits the allocation of the following values:

- o Bit 5 in the [RFC8584] DF Election Capabilities registry, with name "P"(port mode load-balancing) Capability" for port-active ES.

## 10 References

### 10.1 Normative References

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <<https://www.rfc-editor.org/info/rfc8214>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

### 10.2 Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.
- [RFC7275] Martini, L., Salam, S., Sajassi, A., Bocci, M., Matsushima, S., and T. Nadeau, "Inter-Chassis Communication Protocol for Layer 2 Virtual Private Network (L2VPN) Provider Edge (PE) Redundancy", RFC 7275, DOI 10.17487/RFC7275, June 2014, <<https://www.rfc-editor.org/info/rfc7275>>.
- [PREF-DF] Rabadan et al. "Preference-based EVPN DF Election", draft-ietf-bess-evpn-pref-df, work-in-progress, June, 2019.

#### Authors' Addresses

Patrice Brisette  
Cisco Systems  
EMail: [pbrisset@cisco.com](mailto:pbrisset@cisco.com)

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

Luc Andre Burdet  
Cisco Systems  
EMail: [lburdet@cisco.com](mailto:lburdet@cisco.com)

Samir Thoria  
Cisco Systems  
EMail: [sthoria@cisco.com](mailto:sthoria@cisco.com)

Jorge Rabadan  
Nokia  
Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

Bin Wen

INTERNET DRAFT

draft-brisette-bess-evpn-mh-pa

October 31, 2019

Comcast

Email: Bin\_Wen@comcast.com

Edward Leyton

Verizon

Email: edward.leyton@verizonwireless.com

BESS Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: May 3, 2020

P. Brissette, Ed.  
A. Sajassi  
L. Burdet  
Cisco Systems  
J. Uttaro  
ATT  
October 31, 2019

EVPN-VPWS Seamless Integration with Legacy VPWS  
draft-brissette-bess-evpn-vpws-seamless-00

Abstract

This document specifies mechanisms for backward compatibility of Ethernet VPN Virtual Private Wire Service (EVPN-VPWS) solutions with legacy Virtual Private Wire Service (VPWS). It provides mechanisms for seamless integration in the same MPLS/IP network on a per-pseudowire or per-flexible-crossconnect basis. Implementation of this document enables service providers to introduce EVPN-VPWS PEs in their brown-field deployments of legacy VPWS networks. This document specifies control-plane and forwarding behavior needed for auto-discovery of a pseudowire in order to enable seamless integration between EVPN-VPWS and VPWS PEs.

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 <https://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 May 3, 2020.

Copyright Notice

Copyright (c) 2019 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

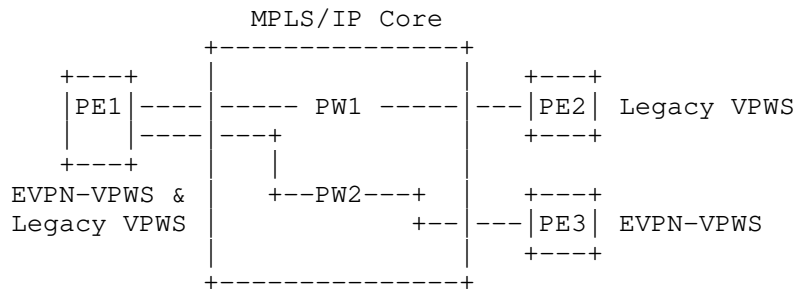
## Table of Contents

1. Introduction . . . . .	2
1.1. Requirements Language . . . . .	3
2. Terms and Abbreviations . . . . .	4
3. Solution Requirements . . . . .	5
4. Seamless Integration . . . . .	6
5. Capability Discovery . . . . .	6
6. Forwarding and Control Plane Operations . . . . .	6
6.1. Multi-homed Operations . . . . .	8
6.1.1. Operations with Port-Active MH PEs . . . . .	8
6.1.2. Operation with Single-Active MH PEs . . . . .	9
6.1.3. Operation with All-Active MH PEs . . . . .	9
6.1.3.1. Falling back to port-active . . . . .	9
6.1.3.2. All-active procedures . . . . .	10
7. IANA Considerations . . . . .	11
8. Security Considerations . . . . .	11
9. References . . . . .	11
9.1. Normative References . . . . .	11
9.2. Informative References . . . . .	11
Authors' Addresses . . . . .	12

## 1. Introduction

Virtual Private Wire Service (VPWS) is a widely-deployed Layer-2 VPN (L2VPN) technology. Many service providers, who are looking at adopting Ethernet VPN Virtual Private Wire Service (EVPN-VPWS), want to preserve their investment in the VPWS networks. Hence, they require mechanisms by which EVPN-VPWS can be introduced into their brown-field legacy VPWS networks without requiring any upgrades (software or hardware) to these networks. This document specifies procedures for the seamless integration of the two technologies (EVPN-VPWS and legacy VPWS) in the same MPLS/IP network. This document specifies control-plane and forwarding behaviour needed for auto-discovery of a pseudowire in order to enable seamless integration between EVPN-VPWS Provider Edge (PE) devices and PEs running legacy VPWS services.





Seamless Integration of EVPN-VPWS.

Figure 1

Figure 1 shows a simple network where PE1 runs in hybrid mode (EVPN-VPWS and legacy VPWS). It provides a pseudowire (PW1) with PE2 running legacy VPWS. It also provides a pseudowire (PW2) with PE2 running EVPN-VPWS. PE2 may be upgraded to EVPN-VPWS seamlessly. Legacy PEs may be setting up PWs per [RFC8077] or may be setting up a VPWS service by first auto-discovering VPN members using [RFC6074] and then setting up the PWs using [RFC8077] or [RFC6624].\

The seamless integration solution described in this document has the following attributes:

- It is backward compatible with [RFC8214] and EVPN Flexible crossconnect Service [evpn\_fxc] document.
- New PEs can leverage the multi-homing mechanisms and provisioning simplifications of EVPN Ethernet-Segment:
  - a. Auto-sensing of MHN / MHD
  - b. Auto-discovery of redundancy group
  - c. Auto-election of Designated Forwarder and VLAN carving
  - d. Support of various load-balancing mode such as port-active, single-active and all-active

#### 1.1. Requirements Language

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

## 2. Terms and Abbreviations

- o CE: A Customer Edge device, e.g., a host, router, or switch.
- o DF: EVPN Ethernet Segment Designated Forwarder.
- o NDF: EVPN Ethernet Segment Non-Designated Forwarder.
- o Ethernet Segment (ES): Refers to the set of Ethernet links that connects a customer site (device or network) to one or more PEs.
- o Ethernet Tag: An Ethernet Tag identifies a particular pseudowire, e.g. a PW-ID.
- o FEC: Forwarding Equivalence Class
- o LDP-LM: LDP Label Mapping Message
- o LDP-LW: LDP Label Withdraw Message
- o LSP: Label Switched Path
- o MHD: Multi-Homed Device
- o MHN: Multi-Homed Network
- o P2P: Point to Point - a P2P LSP typically refers to a LSP for Layer2 pseudowire
- o PE: Provider Edge device
- o VPWS: Virtual Private Wire Service. It refers to legacy VPWS circuit where pseudowires are signalled using LDP or BGP-AD protocol. The latter is referred as VPWS A-D.
- o EVPN-VPWS: Ethernet-VPN Virtual Private Wire Service. It refers to EVPN-VPWS circuit where pseudowires are signalled via BGP-EVPN. It also includes EVPN-FXC service.
- o EVPN-FXC: Ethernet-VPN Flexible Cross-connect Service [evpn\_fxc].
- o Port-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given interface, then the Ethernet Segment is defined to be operating in Port-Active redundancy mode.

- o Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet Segment is defined to be operating in Single-Active redundancy mode.
- o All-Active Redundancy Mode: When all PEs attached to an Ethernet Segment are allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.
- o VPWS A-D: refers to Virtual Private Wire Services with BGP-based Auto Discovery as in [RFC6074].
- o PW: Pseudowire

### 3. Solution Requirements

- o Following are the key requirements for backward compatibility between EVPN-VPWS and VPWS:
- o The solution MUST allow for staged migration towards EVPN-VPWS on a site-by-site basis - e.g., new EVPN-VPWS sites to be provisioned on EVPN-VPWS Provider Edge devices (PEs). Migration SHOULD be possible on a per-pseudowire basis.
- o The solution MUST NOT require any changes to existing VPWS or PEs, not even a software upgrade.
- o The solution MUST allow for the co-existence of PE devices running EVPN-VPWS and VPWS for the same pseudowire and single-homed segments.
- o The solution MUST support port-active redundancy of multi-homed networks and multi-homed devices for EVPN-VPWS PEs.
- o The solution MUST support single-active redundancy of multi-homed networks and multi-homed devices for EVPN-VPWS PEs.
- o The solution SHOULD support all-active redundancy of multi-homed Ethernet Segments for EVPN-VPWS PEs.

These requirements collectively allow for the seamless insertion of the EVPN-VPWS technology into brown-field VPWS deployments.

#### 4. Seamless Integration

In order to support seamless integration with Legacy PEs, this document may require Legacy PEs to setup PWs per [RFC8077] or may require Legacy PEs to setup VPWS service by auto-discovering VPN members using [RFC6074] and then setting up the PWs using [RFC8077] or [RFC6624]. Furthermore, EVPN-VPWS PEs must support BGP EVPN routes per [RFC8214] and one of method of legacy VPWS technologies. All the logic for seamless integration SHALL reside on the EVPN-VPWS PEs.

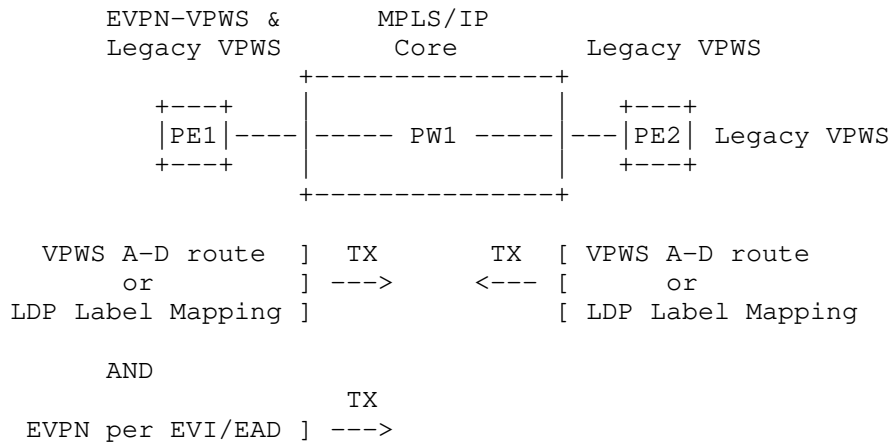
#### 5. Capability Discovery

The EVPN-VPWS PEs MUST advertise both the BGP VPWS Auto-Discovery (VPWS A-D) route or LDP-LM message as well as the BGP EVPN Ethernet-AD per EVI route for a given pseudowire. The VPWS PEs only advertise the BGP VPWS A-D route, per the procedures specified in [RFC4664] and [RFC6074]. The operator may decide to use the same BGP Route Target (RT) to identify a pseudowire on both EVPN-VPWS and VPWS networks. In this case, when a VPWS PE receives the EVPN Ethernet-AD per EVI route, it MUST ignore it on the basis that it belongs to an unknown SAFI. However, the operator may choose to use two RTs - one to identify the pseudowire on VPWS network and another for EVPN-VPWS network and employ RT-constrained [RFC4684] in order to prevent BGP EVPN routes from reaching the VPWS PEs.

When an EVPN-VPWS PE receives both a VPWS A-D route or a LDP-LM message as well as an EVPN-VPWS Ethernet-AD per EVI route from a given remote PE for the same pseudowire, it MUST give preference to the EVPN-VPWS route for the purpose of discovery. This ensures that, at the end of the route exchanges, all EVPN-VPWS capable PEs discover other EVPN-VPWS capable PEs. Furthermore, all the VPWS-only PEs will discover the EVPN-VPWS PEs as if they were standard VPWS PEs. In other words, when the discovery phase is complete, the EVPN-VPWS PEs will have discovered the remote PE per pseudowire along with their associated capability (EVPN-VPWS or VPWS-only), whereas the VPWS PE will have discovered the remote PE per pseudowire as if it was VPWS-only PEs.

#### 6. Forwarding and Control Plane Operations

The procedures for forwarding state setup on the VPWS PE are per [RFC8077], [RFC4761] and [RFC4762].



## EVPN-VPWS Single-Homed

Figure 2

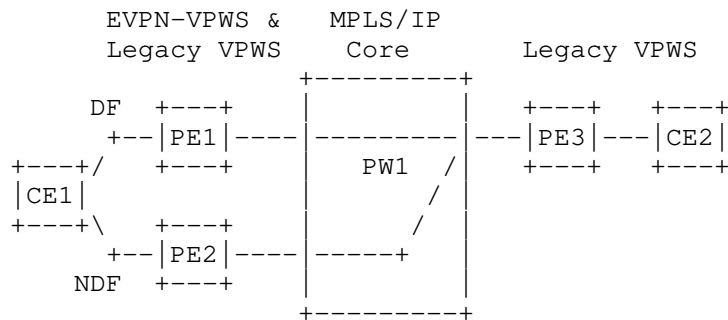
The procedures for forwarding state setup on the EVPN-VPWS PE are as follows:

- o The EVPN-VPWS PE MUST establish a PW to each remote PE from which it has received only a VPWS A-D route or a LDP-LM message for the corresponding pseudowire, and MUST set up the label stack corresponding to the PW FEC.
- o If an EVPN-VPWS PE receives a VPWS A-D route or a LDP-LM message from a given PE, it sets up a Legacy VPWS PW to that PE. If it then receives an EVPN Ethernet-AD per EVI route for that PW from the same PE, then the EVPN-VPWS PE may bring the Legacy PW operationally down and MUST forward traffic using the label information from the EVPN Ethernet-AD per EVI route.
- o If an EVPN-VPWS PE receives an EVPN Ethernet-AD per EVI route followed by a VPWS A-D route or a LDP-LM message from the same PE, then the EVPN-VPWS PE will setup the EVPN-VPWS PW. It may keep the Legacy VPWS PW operationally down and MUST forward traffic using the label information from that EVPN Ethernet-AD per EVI route.
- o For VPWS PE not using VPWS A-D or LDP signalling, the EVPN-VPWS PEs need to be provisioned manually with PWs to those remote VPWS PEs for each pseudowire. In that case, if an EVPN-VPWS PE receives an EVPN Ethernet-AD per EVI route from a PE to which a PW exists, it may keep VPWS PW operationally down and MUST forward

traffic using the label information from that EVPN Ethernet-AD per EVI route.

### 6.1. Multi-homed Operations

Figure 3 below demonstrates multi-homing scenarios. CE1 is connected to PE1 and PE2 where PE1 is the designated forwarder while PE2 is the non designated forwarder.



EVPN-VPWS Port-Active Redundancy

Figure 3

#### 6.1.1. Operations with Port-Active MH PEs

In Figure 3, PE1 and PE2 are configured in port-active load-balancing mode. Both PEs are advertising EVPN Ethernet-AD per ES route with the single-active bit set as described in EVPN port-active document [evpn\_pa]. In this example PE1 is DF elected for the shared Ethernet Segment identifier.

- o Only PE1, as DF, advertises the VPWS A-D route or LDP-LM message towards remote PE3.
- o PE1 advertises the EVPN Ethernet AD per EVI route for PW1 towards remote PE3. The P-bit in L2 Attributes Extended Community is set for PE1 as per [RFC8214]. The purpose is to have all required EVPN-VPWS routes on remote PE so during an upgrade from Legacy VPWS to EVPN-VPWS, those remote nodes are immediately upgraded.
- o PE2, as NDF, only advertises its EVPN Ethernet AD per EVI route corresponding to that same PW1. The B-bit in L2 Attributes Extended Community is set for PE1 as per [RFC8214]

Upon link failure between CE1 and PE1, PE1 and PE2 follows EVPN Ethernet Segment DF Election and/or procedures described in [RFC8214] for the EVPN-VPWS. Furthermore, PE1 withdraws its VPWS A-D route or sends LDP-LW message to remote PE3 to teardown the Legacy PW. Finally, PE2 advertises corresponding VPWS A-D route or LDP-LM message for that PW1 and re-establish Legacy PW with new PE2 destination.

If PE3 is running 2-way pseudowire redundancy and PW-status is enabled, PE2 may leverage the existence of standby/backup PW with PE3. In this particular scenario where PW-status is enabled, PE2 may advertise VPWS A-D route or LDP-LM message along with PW-status message.

Once PE3 is upgraded and supports EVPN-VPWS, EVPN-VPWS routes are exchanged by this PE. Higher precedence of EVPN-VPWS over VPWS allow all PEs to avoid the usage of legacy circuit. At that point in time, unpreferred legacy VPWS protocols and configuration may be removed from all PEs.

#### 6.1.2. Operation with Single-Active MH PEs

Single-active operation is similar to Port-active load-balancing mode described above but at the VLAN level instead being of at the port/interface level. Moreover, the procedures described in [RFC8214] are applied.

The main difference resides on the support of Legacy PW VC-type 4 vs PW VC-Type 5 mode on the EVPN-VPWS PE as per [RFC4448]. While services running in port-active load-balancing mode require raw mode, services running single-active load-balancing mode use tagged mode.

#### 6.1.3. Operation with All-Active MH PEs

In EVPN-VPWS all-active load-balancing mode, all PEs participating in a redundancy group forward traffic bidirectionally, reducing the importance of DF and NDF PE. However PEs running Legacy VPWS do NOT support all-active peering PEs as remote endpoint.

##### 6.1.3.1. Falling back to port-active

PE discovering remote PE running VPWS PW MAY fallback into port-active load-balancing mode. In that case, following rules are applied

- o Peering PEs advertises EVPN Ethernet-AD per ES route with the single-active bit set.

- o DF PE advertises VPWS AD routes or LDP-LM message and EVPN Ethernet AD per EVI route per PW.
- o NDF PE advertises only EVPN Ethernet AD per EVI route per PW.
- o If PE3 is running 2-ways pseudowire redundancy, PE2 may leverage the existence of standby/backup PW with PE3. PE2 may advertises VPWS AD route or LDP-LM message with proper PW-status message.

#### 6.1.3.2. All-active procedures

To support the case where CE device is forwarding traffic to peering PE, extensions to EVPN-VPWS MH procedure are required. In the example of Figure 3, traffic from CE1 going to PE1 gets forwarded to PE3 using the VPN label learned from VPWS AD route or LDP-LM message received from PE3. Traffic from CE1 going to PE2 should get forwarded to PE3 using that same VPN label. Traffic coming from CE3 to PE3 gets forwarded only over the primary PW towards PE1. It is an asymmetric forwarding.

Following rules are applied to achieve expected behaviour:

- o Peering PEs advertises EVPN Ethernet-AD per ES route with the single-active bit unset. Again, to this to get ready on remote PE in case of that legacy PE gets upgraded to EVPN-VPWS.
- o DF PE advertises VPWS AD routes or LDP-LM message and EVPN Ethernet AD per EVI route per PW.
- o NDF PE advertises only EVPN Ethernet AD per EVI route per PW.
- o If PE3 is running 2-ways pseudowire redundancy, PE2 may leverage the existence of standby/backup PW with PE3. PE2 may advertises VPWS AD route or LDP-LM message with proper PW-status message.

To support all-active load-balancing mode on EVPN-VPWS peering PEs, the tunnel encapsulation attribute [tun\_encap] is used to synchronize alias PW label between peering PEs. The tunnel encapsulation attribute, specifying the alias PW label and tunnel endpoint (nexthop) of the remote PE (PE3), is transmitted along with EVPN Ethernet-AD per EVI route. The NDF PEs uses the same VPN label per Legacy PW as DF PE when transmitting traffic coming from CE (CE1) towards remote PE(PE3).



## 7. IANA Considerations

This document has no actions for IANA.

## 8. Security Considerations

The same Security Considerations described in RFC 8214 [RFC8214] are valid for this document.

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <<https://www.rfc-editor.org/info/rfc6074>>.
- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, DOI 10.17487/RFC6624, May 2012, <<https://www.rfc-editor.org/info/rfc6624>>.
- [RFC8077] Martini, L., Ed. and G. Heron, Ed., "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", STD 84, RFC 8077, DOI 10.17487/RFC8077, February 2017, <<https://www.rfc-editor.org/info/rfc8077>>.
- [RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <<https://www.rfc-editor.org/info/rfc8214>>.

### 9.2. Informative References

- [evpn\_fxc] Sajassi, A. and P. Brissette, "draft-ietf-bess-evpn-vpws-fxc", 2019.
- [evpn\_pa] Brissette, P. and A. Sajassi, "draft-brissette-bess-evpn-mh-pa", 2019.

- [RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, DOI 10.17487/RFC4448, April 2006, <<https://www.rfc-editor.org/info/rfc4448>>.
- [RFC4664] Andersson, L., Ed. and E. Rosen, Ed., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, DOI 10.17487/RFC4664, September 2006, <<https://www.rfc-editor.org/info/rfc4664>>.
- [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, DOI 10.17487/RFC4684, November 2006, <<https://www.rfc-editor.org/info/rfc4684>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <<https://www.rfc-editor.org/info/rfc4762>>.
- [tun\_encap] Patel, K., Van de Velde, G., and S. Sangli, "draft-ietf-idr-tunnel-encaps", 2019.
- [vpls\_AA] Sajassi, A., Salam, S., Brissette, P., and L. Jalil, "draft-sajassi-bess-evpn-vpls-all-active", 2017.

## Authors' Addresses

Patrice Brissette (editor)  
Cisco Systems  
Ottawa, ON  
Canada

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

Ali Sajassi  
Cisco Systems  
USA

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

Luc Andre Burdet  
Cisco Systems  
Ottawa  
Canada

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

James Uttaro  
ATT  
USA

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

BESS Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: May 7, 2020

J. Drake  
Juniper Networks  
A. Farrel  
Old Dog Consulting  
L. Jalil  
Verizon  
A. Lingala  
AT&T  
November 4, 2019

BGP-LS Filters : A Framework for Network Slicing and Enhanced VPNs  
draft-drake-bess-enhanced-vpn-02

Abstract

Future networks that support advanced services, such as those enabled by 5G mobile networks, envision a set of overlay networks each with different performance and scaling properties. These overlays are known as network slices and are realized over a common underlay network.

In order to support network slicing, as well as to offer enhanced VPN services in general, it is necessary to define a mechanism by which specific resources (links and/or nodes) of an underlay network can be used by a specific network slice, VPN, or set of VPNs. This document sets out such a mechanism for use in Segment Routing networks.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://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 May 7, 2020.

## Copyright Notice

Copyright (c) 2019 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 (<https://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. Requirements Language . . . . .	3
3. Overview of Approach . . . . .	3
4. Detailed Protocol Operation . . . . .	5
4.1. The BGP-LS Filter Attribute . . . . .	7
4.1.1. The Filter TLV . . . . .	8
4.1.2. The DSCP List TLV . . . . .	9
4.1.3. The Color List TLV . . . . .	10
4.1.4. The Root TLV . . . . .	10
4.2. Error Handling . . . . .	11
5. Comparison With ACTN . . . . .	12
6. Examples . . . . .	12
6.1. MP2MP Connectivity . . . . .	13
6.2. P2MP Unidirectional Connectivity . . . . .	14
6.3. P2P Unidirectional Connectivity . . . . .	15
6.4. P2P Bidirectional Connectivity . . . . .	16
7. Security Considerations . . . . .	17
8. Manageability Considerations . . . . .	17
9. IANA Considerations . . . . .	17
9.1. New BGP Path Attribute . . . . .	17
9.2. New BGP-LS Filter attribute TLVs Type Registry . . . . .	18
10. Acknowledgements . . . . .	18
11. References . . . . .	18
11.1. Normative References . . . . .	18
11.2. Informative References . . . . .	20
Authors' Addresses . . . . .	21

## 1. Introduction

Network slicing is an approach to network operations that builds on the concept of network abstraction to provide programmability, flexibility, and modularity. Driven largely by needs surfacing from 5G, the concept of network slicing has gained traction, for example in [TS23501] and [TS28530]. Network slicing requires the underlying network to support partitioning the network resources to provide the client with dedicated (private) networking, computing, and storage resources drawn from a shared pool. The slices may be seen as (and operated as) virtual networks.

Advanced services drive a need to create virtual networks with enhanced characteristics. The tenant of such a virtual network can require a degree of isolation and performance that previously could only be satisfied by dedicated networks. Additionally, the tenant may ask for some level of control to their virtual networks, e.g., to customize the service forwarding paths in the underlying network.

The concepts of "enhanced VPNs" and "networkslicing" are introduced in [I-D.ietf-teas-enhanced-vpn].

In order to support network slicing, as well as to offer enhanced VPN services in general, it is necessary to define a mechanism by which specific resources (links and/or nodes) of an underlay network can be used by a specific network slice, VPN, or set of VPNs. This document sets out such a mechanism for use in Segment Routing networks [RFC8402] and builds on the ideas introduced in [I-D.ietf-idr-segment-routing-te-policy]. I.e., it generalizes that work to support multipoint-to-multipoint (MP2MP), point-to-multipoint (P2MP), and bidirectional point-to-point (P2P) topologies; it integrates BGP-based VPN support ([RFC4364], [RFC7432]); it supports DSCP as well as a Color-based forwarding, and it uses BGP Link-State (BGP-LS) [RFC7752] to distribute topology information.

## 2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 3. Overview of Approach

The approach is based on a network controller that uses the {source, destination} traffic matrix and the performance and scaling properties of each network slice, VPN, or set of VPNs in conjunction

with the topology of the underlay network to assign each network slice, VPN, or set of VPNs a set of underlay links and nodes that it can use. That is, each network slice, VPN, or set of VPNs gets a subset, either dedicated or shared, of the resources in the underlay network.

It should be noted that resources can be assigned at any of the following granularities:

- o All PEs in a given VPN
- o A set of PEs in a given VPN
- o An individual PE in a given VPN.

Once the network controller has determined the resource assignments, it distributes this information to the PEs that participate in each VPN using the usual VPN information dissemination tools, e.g., route targets (RT) [RFC4360], route reflectors (RR) [RFC4456], and RT constraints [RFC4684].

This information is distributed to the PEs by giving them a customized and limited view of the underlay network on the basis of a network slice, a VPN, or a set of VPNs. Each PE will have a complete view of the underlay network and this customized and limited view acts as filter on the underlay network telling the PE which underlay network resources it can use to direct the traffic of a given network slice, VPN, or set of VPNs to best deliver end-to-end services.

The resource allocation information is encoded using BGP-LS. This approach is chosen for the following reasons:

- o It is BGP-based so it integrates easily with the existing BGP-based VPN infrastructure ([RFC4364], [RFC4684])
- o It supports Segment Routing which is necessary to enforce the PEs' usage of the resources allocated to the VPN or set of VPNs
- o It supports Segment Routing which is necessary to enforce the PEs' usage of the resources allocated to the network slice, VPN, or set of VPNs. The use of RSVP-TE ([RFC3209]) rather than Segment Routing is at the discretion of the network operator as BGP-LS supports both and either confines a packet flow to a specific path.
- o It supports inter-AS connectivity which is a prerequisite for supporting the existing BGP-based VPN infrastructure

- o It is canonical, in that it can be used to advertise the resources of underlay networks that use either IS-IS or OSPF

It should be noted that this mechanism also follows the scalability model of the existing BGP-based VPN infrastructure, which is that the per-VPN information is restricted to only those PE routers that are supporting that VPN and that the P routers have no per-VPN state.

The PEs in non-enhanced VPNs do not receive this resource allocation information and would not confine their usage of the underlay network resources. In order to ensure that the underlay network resources allocated to enhanced VPNs are not inadvertently used by the PEs in non-enhanced VPNs, the network controller SHOULD ensure that the IGP and TE metrics for these resources is higher than the metrics for the underlay network resources allocated to non-enhanced VPNs. In certain situations, detailed in Section 4, PEs in enhanced VPNs will use the underlay networks resources allocated to non-enhanced VPNs.

Additional to the programming of the PEs and its computation and assignment of resources for use by network slices, VPNs, or sets of VPNs, the network controller also instructs the P routers to make the actual allocation of these resources by assigning link bandwidth to a specific DSCP or adjacency SID.

#### 4. Detailed Protocol Operation

We define a BGP-LS Filter to be a BGP-LS encoded description of a subset of the links and nodes in the underlay network. A BGP-LS Filter defines the topology for a network slice or a set of one or more VPNs. The topology defined by a BGP-LS Filter needs to provide connectivity between the PEs in a given network slice, VPN or set of VPNs. I.e., it connects the PEs in these VPNs and is used by them to send packets to each other. A given filter is tagged with the route targets of the VPNs whose PEs are to import the filter. A BGP-LS Filter is pushed southbound to those PEs by the network controller and SHOULD provide multiple paths between a given ingress/egress PE pair.

Note that there will be multiple BGP-LS Filters in a given network deployment and that a given underlay network link or node may appear in more than one of them. In order to provide disambiguation AFI 16388 (BGP-LS) and SAFI 72 (BGP-LS-VPN) are used in BGP-LS UPDATE messages and the network controller SHOULD allocate a different route distinguisher (RD) to each BGP-LS Filter.

Within a given VPN, when an ingress PE needs to send a packet to an egress PE it selects a path to that egress PE from the topology defined by the BGP-LS Filters it has imported for that VPN. It then



either adds a segment routing label stack specifying that path to the packet or places the packet in an RSVP-TE LSP which uses that path. The ingress PE may use any path computation it wishes if that path computation confines the path to the topology defined by the relevant set of BGP-LS Filters.

If Segment Routing is used and a nodal SID is placed in the segment routing label stack, then when that segment is active the P routers will forward the packet using the underlay network resources allocated to non-enhanced VPNs. Similarly, if the RSVP-TE LSP was established using a loose source route to the subject node, the path to that node was selected using the underlay network resources allocated to non-enhanced VPNs.

Because the BGP-LS UPDATE messages specifying a BGP-LS Filter may arrive in any order and the BGP-LS UPDATE messages of multiple BGP-LS Filters may be interleaved, there is a need for a new attribute that is attached to a BGP-LS UPDATE. This attribute contains a Filter ID, a Filter version number, a Filter type (MP2MP, P2MP, or P2P), the total number of fragments in the filter, and the specific fragment number of the piece in hand. I.e., it is assumed that a PE may import more than one BGP-LS Filter, that a given BGP-LS Filter may change over time, and that a given BGP-LS Filter may span multiple BGP-LS UPDATE messages. The Filter ID needs to be unique across the set of VPNs into which the BGP-LS Filter is to be imported.

A BGP-LS Filter that is created for a set of VPNs will contain a set of network resources sufficient to connect the PEs in each VPN in the set and each of the BGP-LS UPDATE messages for the filter MUST be tagged with the RT for each VPN in the set.

If a PE imports more than one BGP-LS Filter it may use the union of the links and nodes specified in each filter when selecting a path. A PE should give precedence to BGP-LS Filters of type P2MP and P2P when selecting a path. Routes targets specific to a given VPN/PE pair are needed for BGP-LS Filters of type P2MP and P2P.

A given BGP-LS Filter may change in response to updates to the PE membership in a VPN to which the BGP-LS Filter applies or to updates to the underlay network. When this occurs, the network controller should push a new version of the affected BGP-LS Filters. That is, it increments the version number of each BGP-LS Filter. Note that a network controller does not need to compute new BGP-LS Filters in response to an individual link or node failure in the underlay network if connectivity still exists among the PEs in the network slice, VPN or set or VPNs with the existing BGP-LS Filters.

A BGP-LS Filter cannot be used by a PE until it is completely assembled. If the BGP-LS Filter that is being assembled is a newer version of a BGP-LS Filter that the PE is currently using, the PE should continue to use its current version of the BGP-LS Filter until the newer version is completely assembled.

When selecting a path using one or more BGP-LS Filters, an ingress PE can use a link or node only if it is active in the underlay network. If this precludes connectivity to the egress PE it may use the underlay network resources allocated to non-enhanced VPNs to reach the egress PE.

Additionally, when there is a newly activated PE it will not be present in any of the BGP-LS Filters used by the other PEs. Until a new BGP-LS Filter or Filters that contain that PE has been distributed, other PEs will use the underlay network resources allocated to non-enhanced VPNs to reach the newly activated PE and it use these resources to reach other PEs.

#### 4.1. The BGP-LS Filter Attribute

[RFC4271] defines the BGP Path attribute. This document introduces a new Optional Transitive Path attribute called the BGP-LS Filter attribute with value TBD1 to be assigned by IANA.

The first BGP-LS Filter attribute MUST be processed and subsequent instances MUST be ignored.

The common fields of the BGP-LS Filter attribute are set as follows:

- o Optional bit is set to 1 to indicate that this is an optional attribute.
- o The Transitive bit is set to 1 to indicate that this is a transitive attribute.
- o The Extended Length bit is set according to the length of the BGP-LS Filter attribute as defined in [RFC4271].
- o The Attribute Type Code is set to TBD1.

The content of the BGP-LS Filter attribute is a series of Type-Length-Value (TLV) constructs. Each TLV may include sub-TLVs. All TLVs and sub-TLVs have a common format that is:

- o Type: A single octet indicating the type of the BGP-LS Filter attribute TLV. Values are taken from the registry described in Section 9.2.

- o Length: A two octet field indicating the length of the data following the Length field counted in octets.
- o Value: The contents of the TLV.

The formats of the TLVs defined in this document are shown in the following sections. The presence rules and meanings are as follows.

- o The BGP-LS Filter attribute MUST contain a Filter TLV.
- o The BGP-LS Filter attribute MAY contain a DSCP List TLV.
- o The BGP-LS Filter attribute MAY contain a Color List TLV.
- o The BGP-LS Filter attribute MAY contain a Root TLV.

#### 4.1.1. The Filter TLV

The BGP-LS Filter attribute MUST contain exactly one Filter TLV. Its format is shown in Figure 1. Note that a given BGP-LS Filter may span multiple UPDATE messages and the Topology, Version Number, and the Number of Fragments fields in the BGP-LS Filter attribute contained in each UPDATE message MUST be set to the same value or the BGP-LS Filter is unusable.

Type = 1 (1 octet)
Length (2 octets)
Topology (1 Octet)
ID (4 Octets)
Version Number (4 Octets)
Number of Fragments (4 Octets)
Fragment Number (4 Octets)

Figure 1: The Filter TLV Format

The fields are as follows:

- o Type is set to 1 to indicate a Filter TLV.

- o Length is set to 17 octets.
- o Topology indicates whether this BGP-LS Filter is MP2MP, P2MP, P2P unidirectional, or P2P bidirectional.
- o The ID of this BGP-LS Filter. This ID needs to be unique within the set of VPNs into which the BGP-LS Filter is to be imported.
- o The Version Number of this BGP-LS Filter. I.e., the contents of a BGP-LS Filter with a given ID may change over time and this field indicates the latest version of that BGP-LS Filter.
- o Number of Fragments indicates the number of BGP UPDATE messages defining this BGP-LS Filter.
- o Fragment Number indicates ordinal position of this UPDATE message within the set of UPDATE messages defining this BGP-LS Filter. A BGP-LS Filter is not complete, i.e., usable, until all UPDATE messages have been received with Fragment Numbers in the range 1 <= Fragment Number <= Number of Fragments. An UPDATE message with a Fragment Number outside this range is to be ignored.

#### 4.1.2. The DSCP List TLV

The DSCP List TLV MAY be included in the BGP-LS Filter attribute. If included, a packet whose DSCP matches a DSCP in the DSCP list is to be forwarded using the BGP-LS Filter defined by the containing BGP-LS Filter attribute. The first DSCP List TLV MUST be processed and subsequent instances MUST be ignored. The format of the DSCP List TLV is shown in Figure 2.

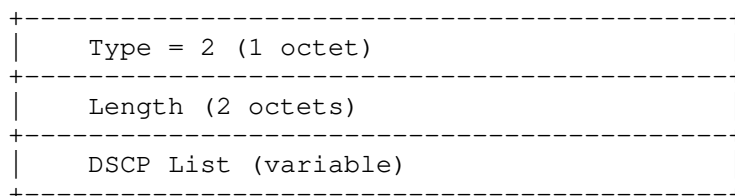


Figure 2: The DSCP List TLV Format

The fields are as follows:

- o Type is set to 2 to indicate a DSCP List TLV.
- o Length indicates the length in octets of the DSCP List.

- o DSCP List contains a list of DSCPs, each one octet in length and encoded in the standard format.

#### 4.1.3. The Color List TLV

The Color List TLV MAY be included in the BGP-LS Filter attribute. If a BGP UPDATE contains a Color extended community with a color (as defined by [RFC5512]) that matches an entry in the Color List, then a packet whose destination is covered by one of the routes in that UPDATE is to be forwarded using the BGP-LS Filter defined by the containing BGP-LS Filter attribute. The first Color List TLV MUST be processed and subsequent instances MUST be ignored. The format of the Color List TLV is shown in Figure 3.

Note that if both a DSCP List and a Color List TLV are included in a BGP-LS Filter attribute, packets matching an entry in either list are to be forwarded using the BGP-LS Filter defined by the containing BGP-LS Filter attribute. If neither list is included then all packets for that network slice, VPN, or set of VPNs can be forwarded using the BGP-LS Filter defined by the containing BGP-LS Filter attribute.

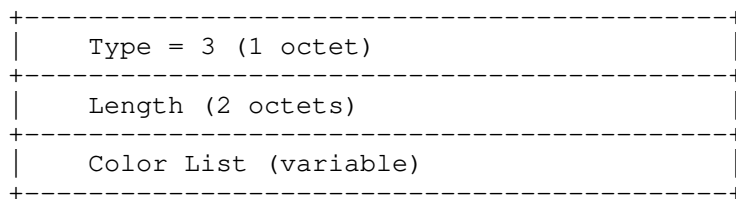


Figure 3: The Color List TLV Format

The fields are as follows:

- o Type is set to 3 to indicate a Color List TLV.
- o Length indicates the length in octets of the Color List.
- o Color List contains a list of Colors, each four octets in length.

#### 4.1.4. The Root TLV

The Root TLV MUST be included in the BGP-LS Filter attribute if its topology is of type P2MP or P2P unidirectional. It defines the root node for that topology and if it is not present the BGP-LS Filter is

unusable. The TLV, if present, MUST be ignored if the topology is of type MP2MP or P2P bidirectional.

The Root TLV is structured as shown in Figure 4 and MAY contain any of the sub-TLVs defined in section 3.2.1.4 of [RFC7752].

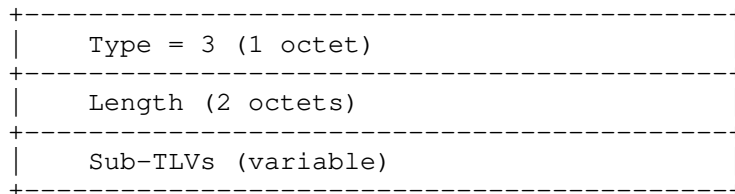


Figure 4: The Root TLV Format

The fields are as follows:

- o Type is set to 3 to indicate a Color List TLV.
- o Length indicates the length in octets of the Color List.
- o There follows a sequence of zero or more sub-TLVs as defined in section 3.2.1.4 of [RFC7752]. The presence of sub-TLVs can be deduced from the Length field of the Root TLV and from the Length fields of each of the sub-TLVs.

#### 4.2. Error Handling

Section 6 of [RFC4271] describes the handling of malformed BGP attributes, or those that are in error in some way. [RFC7606] revises BGP error handling specifically for the for UPDATE message, provides guidelines for the authors of documents defining new attributes, and revises the error handling procedures for a number of existing attributes. This document introduces the BGP-LS Filter attribute and so defines error handling as follows:

- o When parsing a message, an unknown Attribute Type code or a length that suggests that the attribute is longer than the remaining message is treated as a malformed message and the "treat-as-withdraw" approach used as per [RFC7606].
- o When parsing a message that contains an BGP-LS Filter attribute, the following cases constitute errors:
  1. Optional bit is set to 0 in BGP-LS Filter attribute.

2. Transitive bit is set to 0 in BGP-LS Filter attribute.
  3. The attribute does not contain a Filter TLV or it contains more than one Filter TLV.
  4. The TLV length indicates that the TLV extends beyond the end of the BGP-LS Filter attribute.
  5. There is an unknown TLV type field found in BGP-LS Filter attribute.
- o The errors listed above are treated as follows:
- 1., 2., 3., 4.: The attribute MUST be treated as malformed and the "treat-as-withdraw" approach used as per [RFC7606].
  - 5.: Unknown TLVs SHOULD be ignored, and message processing SHOULD continue.
5. Comparison With ACTN
- TBD
6. Examples
- Figure 5 shows a sample underlay topology. Six PEs (PE1 through PE6) are connected across a network of twelve P nodes (P1 through P12). Each PE is dual-homed, and the P nodes are variously connected so that there are multiple routes between PEs.

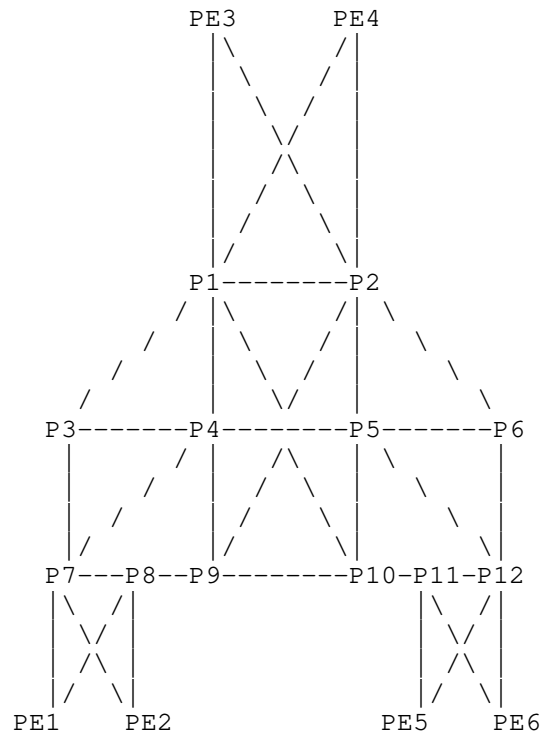


Figure 5: Underlay Network Topology

#### 6.1. MP2MP Connectivity

Figure 6 shows how a Multi-point-to-multipoint (MP2MP) service that connects PE1, PE3, and PE6 can be installed over the underlay network. Path have been computed so that, for example, PE1 is connected to both PE3 and PE6 via a pair of redundant paths. Similarly, PE3 is connected to PE1 and PE6, and PE6 is connected to PE1 and PE3.



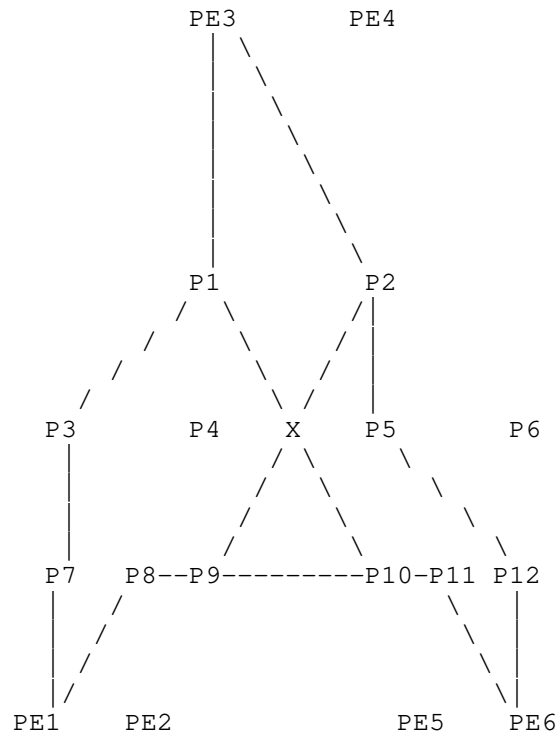


Figure 6: An MP2MP Service Installed at PE1, PE3, and PE6

## 6.2. P2MP Unidirectional Connectivity

Figure 7 shows the provision of a Point-to-Multipoint (P2MP) rooted at PE3 and connected to PE1 and PE6. As in the previous example, a redundant pair of paths is established between PE3 and each of PE1 and PE6. Thus, the two paths from PE3 to PE1 are PE3-P1-P4-P7-PE1 and PE3-P2-P9-P8-PE1.

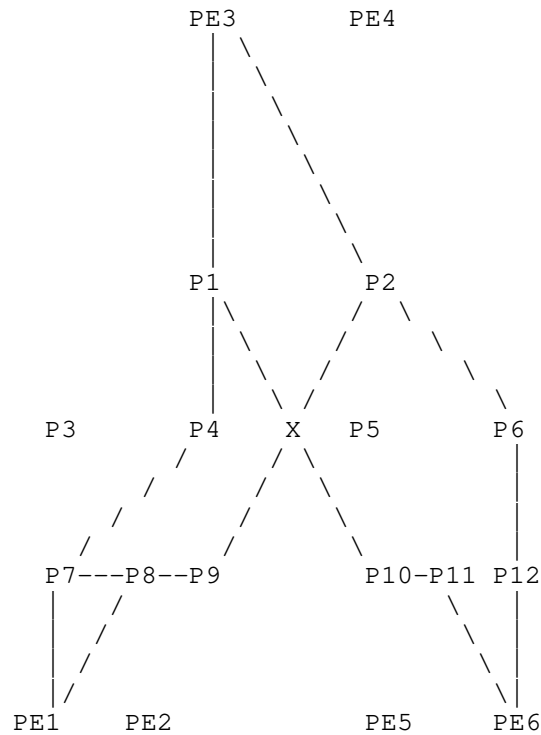


Figure 7: A P2MP Unidirectional Service Installed at PE3

### 6.3. P2P Unidirectional Connectivity

Figure 8 shows a Point-to-Point (P2P) service rooted at PE1 and connected to PE3. This is equivalent to a Segment Routing Traffic Engineering (SR TE) Policy [I-D.ietf-idr-segment-routing-te-policy] installed at PE1.

As in the previous examples, a pair of redundant paths are computed.

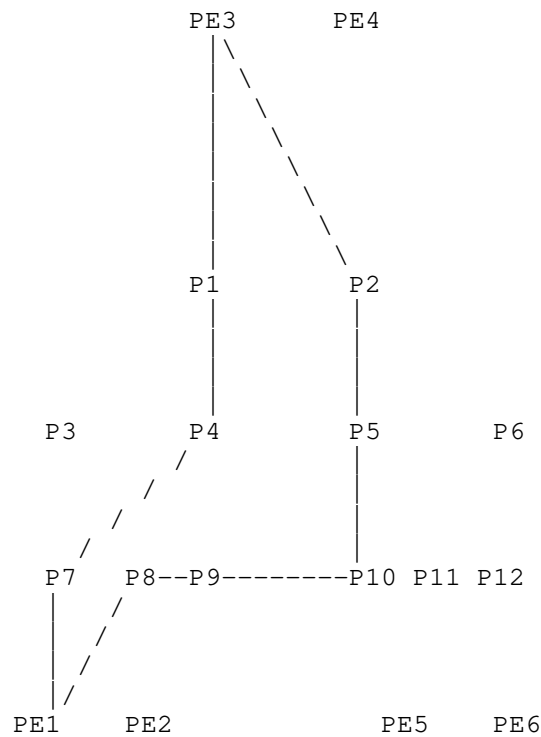


Figure 8: A P2P Unidirectional Service (SR TE Policy) Installed at PE1

#### 6.4. P2P Bidirectional Connectivity

Figure 9 show a bidirectional P2P service connecting PE1 and PE6. This is equivalent to a Segment Routing Traffic Engineering (SR TE) Policy [I-D.ietf-idr-segment-routing-te-policy] installed at PE1 and PE6.

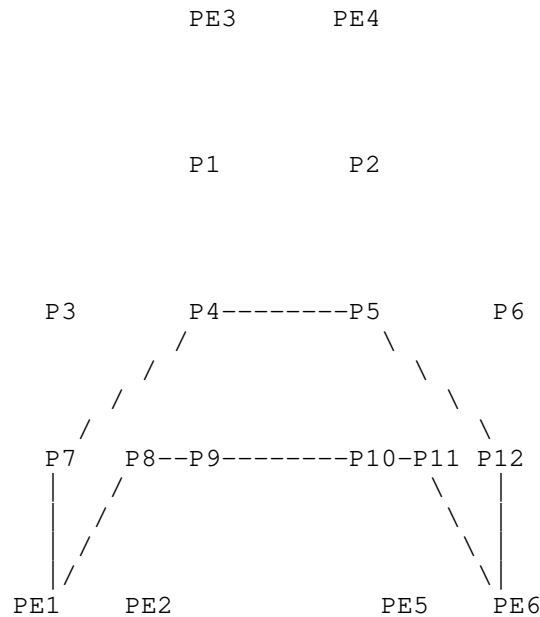


Figure 9: A P2P Bidirectional Service Installed at PE1 and PE6

## 7. Security Considerations

TBD

## 8. Manageability Considerations

Per VPN OAM and telemetry will be required in order to monitor and verify the performance of network slices. This is particularly important when the performance of a network slice has been committed to a customer through a Service Level Agreement.

TBD

## 9. IANA Considerations

### 9.1. New BGP Path Attribute

IANA maintains a registry of "Border Gateway Protocol (BGP) Parameters" with a subregistry of "BGP Path Attributes". IANA is requested to assign a new Path attribute called "BGP-LS Filter attribute" (TBD1 in this document) with this document as a reference.

## 9.2. New BGP-LS Filter attribute TLVs Type Registry

IANA maintains a registry of "Border Gateway Protocol (BGP) Parameters". IANA is request to create a new subregistry called the "BGP-LS Filter attribute TLVs" registry.

Valid values are in the range 0 to 255.

- o Values 0 and 255 are to be marked "Reserved, not to be allocated".
- o Values 1 through 254 are to be assigned according to the "First Come First Served" policy [RFC8126]

This document should be given as a reference for this registry. The new registry should track:

- o Type
- o Name
- o Reference Document or Contact
- o Registration Date

The registry should initially be populated as follows:

Type	Name	Reference	Date
1	Filter TLV	[This.I-D]	Date-to-be-set
2	DSCP List TLV	[This.I-D]	Date-to-be-set
3	Color List TLV	[This.I-D]	Date-to-be-set
4	Root TLV	[This.I-D]	Date-to-be-set

## 10. Acknowledgements

The authors are grateful to all those who contributed to the discussions that led to this work: Ron Bonica, Stewart Bryant, Jie Dong, Keyur Patel, and Colby Barth.

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <<https://www.rfc-editor.org/info/rfc5512>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <<https://www.rfc-editor.org/info/rfc7606>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

## 11.2. Informative References

- [I-D.ietf-idr-segment-routing-te-policy]  
Previdi, S., Filsfils, C., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-07 (work in progress), July 2019.
- [I-D.ietf-teas-enhanced-vpn]  
Dong, J., Bryant, S., Li, Z., Miyasaka, T., and Y. Lee, "A Framework for Enhanced Virtual Private Networks (VPN+) Service", draft-ietf-teas-enhanced-vpn-03 (work in progress), September 2019.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<https://www.rfc-editor.org/info/rfc4360>>.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <<https://www.rfc-editor.org/info/rfc4456>>.
- [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, DOI 10.17487/RFC4684, November 2006, <<https://www.rfc-editor.org/info/rfc4684>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.
- [TS23501] 3GPP, "System architecture for the 5G System (5GS) - 3GPP TS23.501", 2016, <<https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>>.

[TS28530] 3GPP, "Management and orchestration; Concepts, use cases and requirements - 3GPP TS28.530", 2016,  
<<https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>>.

Authors' Addresses

John Drake  
Juniper Networks

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

Adrian Farrel  
Old Dog Consulting

Email: [adrian@olddog.co.uk](mailto:adrian@olddog.co.uk)

Luay Jalil  
Verizon

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

Avinash Lingala  
AT&T

Email: [ar977m@att.com](mailto:ar977m@att.com)





Network Working Group  
Internet Draft  
Intended status: Informational  
Expires: January 13, 2021

L. Dunbar  
J. Guichard  
Futurewei  
Ali Sajassi  
Cisco  
J. Drake  
Juniper  
B. Najem  
Bell Canada  
Ayan Barnerjee  
D. Carrel  
Cisco

July 13, 2020

BGP Usage for SDWAN Overlay Networks  
draft-dunbar-bess-bgp-sdwan-usage-08

Abstract

The document describes three distinct SDWAN scenarios and discusses the applicability of BGP for each of those scenarios. The goal of the document is to demonstrate how BGP-based control plane is used for large scale SDWAN overlay networks with little manual intervention.

SDWAN edge nodes are commonly interconnected by multiple underlay networks which can be owned and managed by different network providers.

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

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

other groups may also distribute working documents as Internet-Drafts.

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

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

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

This Internet-Draft will expire on January 13, 2009.

#### Copyright Notice

Copyright (c) 2020 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. Conventions used in this document.....	4
3. Use Case Scenario Description and Requirements.....	6
3.1. Requirements.....	6
3.1.1. Supporting Multiple SDWAN Segmentations.....	6
3.1.2. Client Service Requirement.....	6
3.1.3. Application Flow Based Segmentation.....	7
3.1.4. Zero Touch Provisioning.....	8
3.1.5. Constrained Propagation of SDWAN Edge Properties.....	9

3.2. Scenarios #1: Homogeneous WAN.....	10
3.3. Scenario #2: CPE based SDWAN over Hybrid WAN Underlay....	11
3.4. Scenario #3: Private VPN PE based SDWAN.....	14
4. BGP Walk Through.....	15
4.1. BGP Walk Through for Homogeneous SDWAN.....	15
4.2. BGP Walk Through for Application Flow Based Segmentation.	18
4.3. Client Service Provisioning Model.....	19
4.4. WAN Ports Provisioning Model.....	20
4.5. Why BGP as Control Plane for SDWAN?.....	20
5. SDWAN Traffic Forwarding Walk Through.....	21
5.1. SDWAN Network Startup Procedures.....	21
5.2. Packet Walk-Through for Scenario #1.....	22
5.3. Packet Walk-Through for Scenario #2.....	22
5.4. Packet Walk-Through for Scenario #3.....	24
6. Manageability Considerations.....	24
7. Security Considerations.....	24
8. IANA Considerations.....	25
9. References.....	25
9.1. Normative References.....	25
9.2. Informative References.....	25
10. Acknowledgments.....	27

## 1. Introduction

Here are some key characteristics of "SDWAN" networks:

- Augment of transport, which refers to utilizing overlay paths over different underlay networks. Very often there are multiple parallel overlay paths between any two SDWAN edges, some of which are private networks over which traffic can traverse with or without encryption, others require encryption, e.g. over untrusted public networks.
- Enable direct Internet access from remote sites, instead hauling all traffic to Corporate HQ for centralized policy control.
- Some traffic are routed based on application IDs instead of based on destination IP addresses.
- The Application Routing can also be based on specific performance criteria (e.g. packets delay, packet loss, jitter) to provide better application performance by choosing the right underlay that meets or exceeds the specified criteria.

[Net2Cloud-Problem] describes the network related problems that enterprises face to connect enterprises' branch offices to dynamic workloads in different Cloud DCs, including using SDWAN to aggregate

multiple paths provided by different service providers to achieve better performance and to accomplish application ID based forwarding.

Even though SDWAN has been positioned as a flexible way to reach dynamic workloads in third party Cloud data centers over different underlay networks, scaling becomes a major issue when there are hundreds or thousands of nodes to be interconnected by an SDWAN overlay networks.

BGP is widely used by underlay networks. This document describes using BGP for edge nodes to exchange information across the SDWAN overlay networks.

## 2. Conventions used in this document

Cloud DC: Third party data centers that host applications and workloads owned by different organizations or tenants.

Controller: Used interchangeably with SDWAN controller to manage SDWAN overlay path creation/deletion and monitor the path conditions between sites.

CPE: Customer Premise Equipment

CPE-Based VPN: Virtual Private Secure network formed among CPEs. This is to differentiate from more commonly used PE-based VPNs [RFC 4364].

Homogeneous SDWAN: A type of SDWAN network in which all traffic to/from the SDWAN edge nodes has to be encrypted regardless of underlay networks. For lack of better terminology, we call this Homogeneous SDWAN throughout this document.

ISP: Internet Service Provider

NSP: Network Service Provider. NSP usually provides more advanced network services, such as MPLS VPN, private leased lines, or managed Secure WAN connections, many

times within a private trusted domain, whereas an ISP usually provides plain internet services over public untrusted domains.

PE: Provider Edge

SDWAN Edge Node: an edge node, which can be physical or virtual, maps the attached clients' traffic to the wide area network (WAN) overlay tunnels.

SDWAN: Software Defined Wide Area Network. In this document, "SDWAN" refers to the solutions of pooling WAN bandwidth from multiple underlay networks to get better WAN bandwidth management, visibility & control. When the underlay networks are private, traffic can traverse without additional encryption; when the underlay networks are public, such as the Internet, some traffic may need to be encrypted when traversing through (depending on user provided policies).

SDWAN IPsec SA: IPsec Security Association between two SDWAN ports or nodes.

SDWAN over Hybrid Networks: SDWAN over Hybrid Networks typically have edge nodes utilizing bandwidth resources from multiple service providers. In Hybrid SDWAN network, packets over private networks can go natively without encryption and are encrypted over the untrusted network, such as the public Internet.

WAN Port: A Port or Interface facing an ISP or Network Service Provider (NSP), with address (usually public routable address) allocated by the ISP or the NSP.

C-PE: SDWAN Edge node, which can be CPE for customer managed SDWAN, or PE that is for provider managed SDWAN services).

ZTP: Zero Touch Provisioning

### 3. Use Case Scenario Description and Requirements

SDWAN networks can have different topologies and have different traffic patterns. To make it easier for the focused discussion in subsequent drafts on SDWAN control plane and data plane, this section describes several SDWAN scenarios that may have different impact on their corresponding control planes & data planes.

#### 3.1. Requirements

##### 3.1.1. Supporting Multiple SDWAN Segmentations

The term "network segmentation", a.k.a. SDWAN instances, is referring to the process of dividing the network into logical sub-networks using isolation techniques on a forwarding device such as a switch, router, or firewall. For a homogeneous network, such as MPLS VPN or Layer 2 network, VRF or VLAN are used to achieve the network segmentation.

As SDWAN is an overlay network arching over multiple types of networks, MPLS L2VPN/L3VPN or pure L2 underlay can continue using the VRF, VN-ID or VLAN to differentiate SDWAN network segmentations. For public internet, the IPsec inner encapsulation header can carry the SDWAN Instance Identifier to differentiate the packets belonging to different SDWAN instances.

BGP already has the capability to differentiate control packets for different network instances. When using BGP for SDWAN, the SDWAN segmentations can be differentiated by the SDWAN Target ID in the BGP Extended Community in the same way as VPN instances being represented by the Route Target. Same as Route Target, need to use a different name to differentiate from VPN if a CPE supports traditional VPN with multiple VRFs and supports multiple SDWAN Segmentations (instances). The actual SDWAN Target ID encoding is proposed by [SDWAN-EDGE-Discovery].

##### 3.1.2. Client Service Requirement

Client interface of SDWAN nodes can be IP or Ethernet based.

For Ethernet based client interfaces, SDWAN edge should support VLAN-based service interfaces (EVI100), VLAN bundle service interfaces (EVI200), or VLAN-Aware bundling service interfaces. EVPN service requirements are applicable to the Client traffic, as described in the Section 3.1 of RFC8388.

For IP based client interfaces, L3VPN service requirements are applicable.

### 3.1.3. Application Flow Based Segmentation

Application Flow based Segmentation, also known as SDWAN Traffic Segmentation, enables the separation of the traffic based on the business and the security needs for different users' groups and/or application requirements. Each user group and/or applications may need different isolated topology and/or policies to fulfill the business requirements.

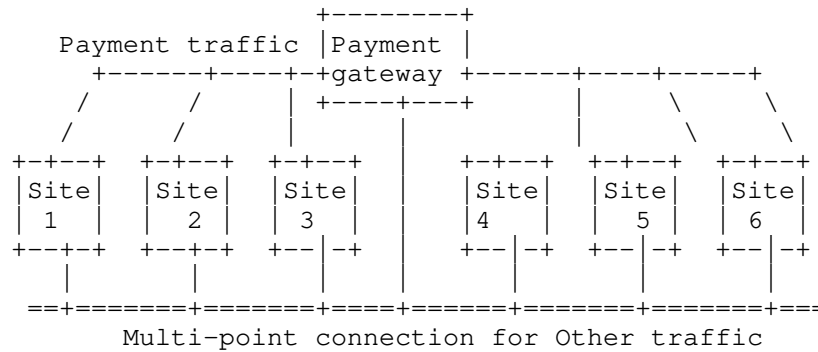
The Application Flow based Segmentation concept is analogous to VLAN (in L2 network) and VRF (in L3 network).

One can think about the Application Flow based Segmentation as a feature that can be provided or enabled on a single SDWAN service (or domain) to a single Subscriber. Each SDWAN Service can have one or more overlay Segments to support the business requirement; each Segment has its own policy, topology and application/user groups. Applications/users' group can belong to more than one Segment.

For example, a retail business requires the point-of-sales (PoS) application in all stores to be isolated from other applications AND routed only to the payment processing entity at a hub site (i.e. hub and spoke); however, the same retail business requires the other applications to be routed to all sites (i.e. multipoint-to-multipoint) AND isolated from the PoS application.

In the figure below, the traffic from the PoS application follows a Tree topology, whereas other traffic can be multipoint-to-multipoint topology.





Another example is an enterprise who wants to isolate the traffic for each department and have different topology and policy for different department; the HR department may need to access certain applications that are NOT accessible by the engineering department. In addition, the contractors may have a limited access to the enterprise resources.

#### 3.1.4. Zero Touch Provisioning

Unlike traditional EVPN or L3VPN whose PEs are deployed for long term, SDWAN edge nodes (virtual or physical) deployment at a specific location can be ephemeral. Therefore, Zero Touch Provisioning (ZTP), or Plug and Play, is a common requirement for SDWAN. When an SDWAN edge is physically installed at a location or instantiated on a VM in a Cloud DC, ZTP automates follow-up steps, including updates to the OS, software version, and configuration prior to connection. From network control perspective, ZTP includes the following:

- Upon power up, an SDWAN node can establish transport layer secure connection (such as TLS, SSL, etc.) to its controller whose address can be burned or preconfigured on the device.
- The SDWAN Controller can designate a Local Network Controller in the proximity of the SDWAN node; the Local Network Controller manages and monitor the communication policies of the edge node.

### 3.1.5. Constrained Propagation of SDWAN Edge Properties

One SDWAN edge node may only be authorized to communicate with a small number of other SDWAN edge nodes. Under this circumstance, the property of the SDWAN edge node cannot be propagated to any other nodes who are not authorized to communicate. But a remote SDWAN edge node upon powering up might not have the proper policies to know who the authorized peers are. Therefore, it is very essential for SDWAN deployment have a central point to distribute the properties of each SDWAN edge node to its authorized peers.

BGP is well suited for this purpose. RFC 4684 has specified the procedure to constrain the distribution of BGP UPDATE to only a subset of SDWAN edges. Basically, each edge node informs the Route Reflector (RR) [RFC4456] on its interested SDWAN instances. The RR only propagates the BGP UPDATE for the relevant SDWAN instances to the edge.

Usually the connection between a SDWAN edge node and its RR is over insecure network. Therefore, upon power up, a SDWAN node needs to establish a secure transport layer connection (TLS, SSL, etc.) to its designated RR. The BGP UPDATE messages need to be sent over the secure channel (TLS, SSL, etc.) to the RR.

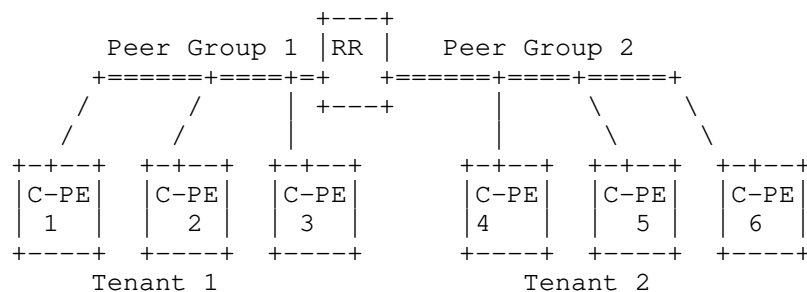


Figure 1: Peer Groups managed by RR

Tenant separation is achieved by the SDWAN instance identification represented in control plane and data plane, respectively.

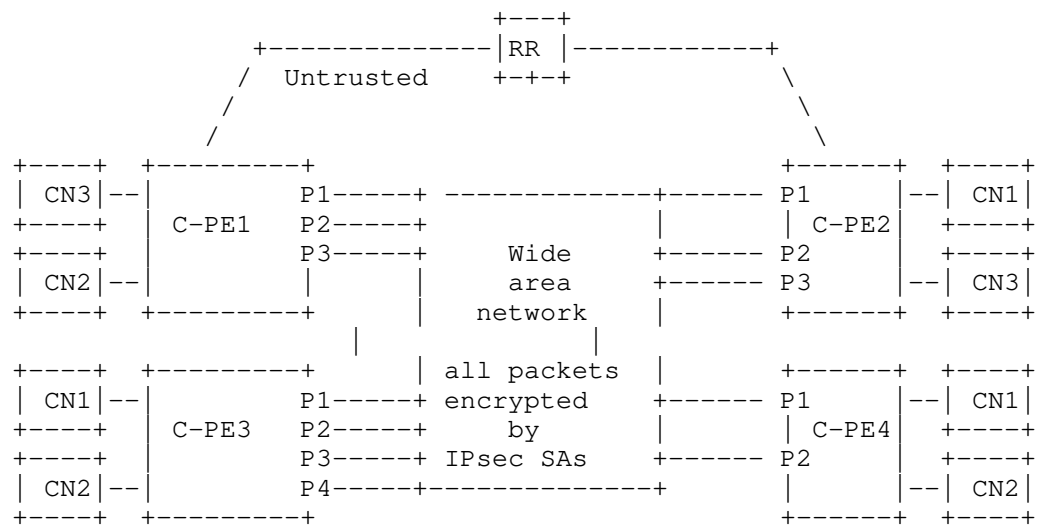
### 3.2. Scenarios #1: Homogeneous WAN

This is referring to a type of SDWAN network with edge nodes encrypting all traffic over WAN to other edge nodes, regardless of whether the underlay is private or public. For lack of better terminology, we call this Homogeneous SDWAN throughout this document.

Some typical scenarios for the use of a Homogeneous SDWAN network are as follows:

- A small branch office connecting to its HQ offices via the Internet. All sensitive traffic to/from this small branch office has to be encrypted, which is usually achieved using IPsec SAs.
- A store in a shopping mall may need to securely connect to its applications in one or more Cloud DCs via the Internet. A common way of achieving this is to establish IPsec SAs to the Cloud DC gateway to carry the sensitive data to/from the store.

As described in [SECURE-EVPN], the granularity of the IPsec SAs for Homogeneous SDWAN can be per site, per subnet, per tenant, or per address. Once the IPsec SA is established for a specific subnet/tenant/site, all traffic to/from the subnets/tenants/site are encrypted.



CN: Client Networks, which is same as Tenant Networks used by NVo3

Figure 2: Homogeneous SDWAN

One of the key properties of homogeneous SDWAN is that the SDWAN Local Network Controller (RR) is connected to C-PEs via untrusted public network, therefore, requiring secure connection between RR and C-PEs (TLS, DTLS, etc.).

Homogeneous SDWAN has some similarity to commonly deployed IPsec VPN, albeit the IPsec VPN is usually point-to-point among a small number of nodes and with heavy manual configuration for IPsec between nodes, whereas an SDWAN network can have a large number of edge nodes with an SDWAN controller to manage requiring zero touch provisioning upon powering up.

Existing Private VPNs (e.g. MPLS based) can use homogeneous SDWAN to extend over public network to remote sites to which the VPN operator does not own or lease infrastructural connectivity, as described in [SECURE-EVPN] and [SECURE-L3VPN]

### 3.3. Scenario #2: CPE based SDWAN over Hybrid WAN Underlay

In this scenario, SDWAN edge nodes (a.k.a. C-PEs) have some WAN ports connected to PEs of Private VPNs over which packets can be forwarded natively without encryption, and some WAN ports connected to the public Internet over which sensitive traffic have to be encrypted (usually by IPsec SA).

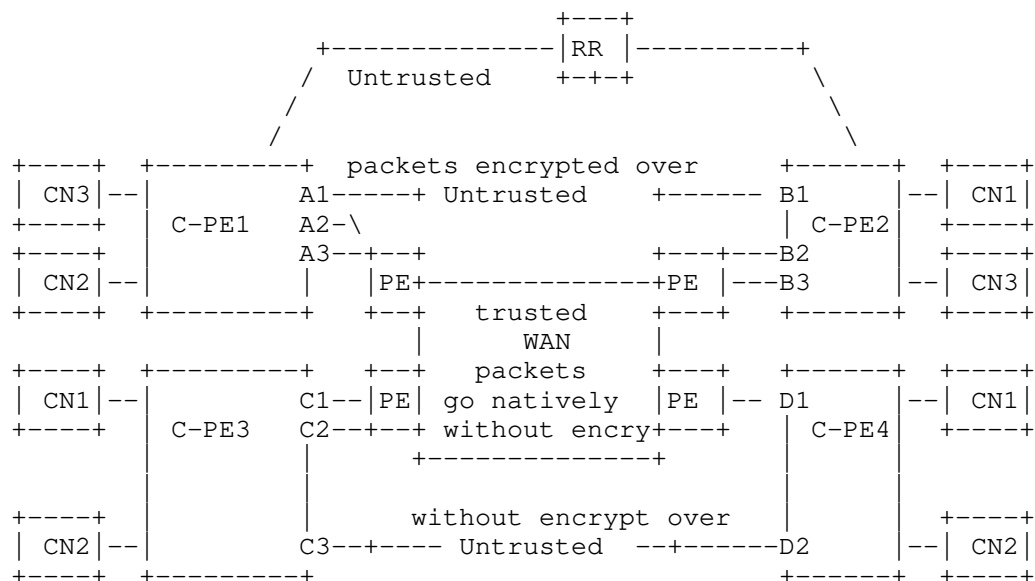
In this scenario, the SDWAN edge nodes' egress WAN ports are all IP/Ethernet based, either egress to PEs of the VPNs or egress to the public Internet. Even if the VPN is a MPLS network, the VPN's PEs have IP/Ethernet links to the SDWAN edge (C-PEs). Throughout this document, this scenario is also called CPE based SDWAN over Hybrid Networks.

Even though IPsec SA can secure the packets traversing the Internet, it does not offer the premium SLA commonly offered by Private VPNs, especially over long distance. Clients need to have policies to specify criteria for flows only traversing private VPNs or traversing either as long as encrypted when over the Internet. For example, client can have those policies for the flows:

1. A policy or criteria for sending the flows over a private network without encryption (for better performance),
2. A policy or criteria for sending the flows over any networks as long as the packets of the flows are encrypted when traversing untrusted networks, or
3. A policy of not needing encryption at all.

If a flow traversing multiple segments, such as A<->B<->C<->D, has either Policy 2 or 3 above, the flow can traverse different underlays in different network segments, such as over Private network underlay between A<->B without encryption, or over the public internet between B<->C in an IPsec SA.

As shown in the figure below, C-PE-1 has two different types of interfaces (A1 to Internet and A2 & A3 to VPN). The C-PEs' loopback addresses and addresses attached to C-PEs may or may not be visible to the ISPs/NSPs. The addresses for the WAN ports can have addresses allocated by service providers or dynamically assigned (e.g. by DHCP). One WAN port shown in the figure below (e.g. A1, A2, A3 etc.) is a logical representation of potential multiple physical ports on the C-PEs.



CN: Client Network

Figure 3: Hybrid SDWAN

Some key characteristics of a Hybrid SDWAN overlay network are as follows:

- one C-PE may be connected to different ISPs/NSPs, with some of its WAN ports addresses being assigned by different ISPs/NSPs.
- The WAN ports connected to PEs of trusted private networks (e.g. MPLS VPN) hand off IP/Ethernet packets, just like today's CPE that do not handle MPLS packets and do not participate in the underlay VPN networks' control plane. Traffic can flow natively without encryption when be forwarded out through those WAN ports for better performance.
- The WAN ports connected to untrusted networks, e.g. the Internet, requires sensitive traffic to be encrypted, i.e. encrypted by IPsec SA.
- An SDWAN local Network Controller (RR) is connected to C-PEs via the untrusted public network, therefore, requiring secure connection between RR and C-PEs via TLS, DTLS, etc.
- The SDWAN nodes' [loopback] addresses might not be routable nor visible in the underlay ISP/NSP networks. Routes & services attached to SDWAN edges at the SDWAN overlay layer are in different address spaces than the underlay networks.
- There could be multiple SDWAN devices sharing a common property, such as a geographic location. Some applications over SDWAN may need to traverse specific geographic locations for various reasons, such as to comply with regulatory rules, to utilize specific value added services, or others.
- The underlay path selection between sites can be a local decision. Some policies allow one service from C-PE1 -> C-PE2 -> C-PE3 using one ISP/NSP underlay in the first segment (C-PE1 -> C-PE2) and using a different ISP/NSP in the second segment (C-PE2-> CPE3).
- Services may not be congruent, i.e. the packets from A-> B may traverse one underlay network, and the packets from B -> A may traverse a different underlay.
- Different services, routes, or VLANs attached to SDWAN nodes can be aggregated over one underlay path; same service/routes/VLAN can spread over multiple SDWAN underlays at different times depending on the policies specified for the service. For example, one tenant's packets to HQ need to be encrypted when sent over the Internet or have to be sent over private networks, while the same

tenant's packets to Facebook can be sent over the Internet without encryption.

### 3.4. Scenario #3: Private VPN PE based SDWAN

This scenario refers to existing VPN (e.g. MPLS based VPN, such as EVPN or IPVPN) adding extra ports facing untrusted public networks allowing PEs to offload some low priority traffic to ports facing public networks when the VPN MPLS paths are congested. Throughout this document, this scenario is also called Internet Offload for Private VPN, or PE based SDWAN.

In this scenario, the packets offloaded to untrusted public network must be encrypted.

PE based SDWAN can be used by VPN service providers to temporarily increase bandwidth between sites when they are not sure if the demand will sustain for long period of time or as a temporary solution before the permanent infrastructure is built or leased.

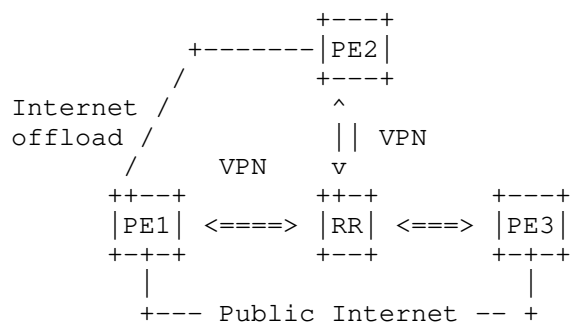


Figure 4: Additional Internet paths added to the VPN

Here are some key properties for PE based SDWAN:

- For MPLS based VPN, PEs continue having MPLS encapsulation handoff to existing paths.

- The BGP RR is connected to PEs in the same way as VPN, i.e. via the trusted network.
- For the added Internet ports, PEs have IP packets handoff, i.e. sending and receiving IP data frames. Internally, PEs can have the option to encapsulate the MPLS payload in IP, as specified by RFC4023.
- The ports facing public internet might get IP addresses assigned by ISPs, which may not be in the same address domain as PEs'.
- Ports facing public internet are not as secure as the ports facing private infrastructure. There could be spoofing, or DDOS attacks to the ports facing public internet. Extra consideration must be given when injecting the new routes learned from public network into VRFs.
- Even though packets are encrypted over public internet, the performance SLA is not guaranteed over public internet. Therefore, clients may have policies only allowing some flows to be offloaded to internet path.

#### 4. BGP Walk Through

##### 4.1. BGP Walk Through for Homogeneous SDWAN

In the figure below, packets destined towards multiple routes attached to the C-PE2 can be carried by one IPsec tunnel. Then one BGP UPDATE can be announced by C-PE2 to its RR.



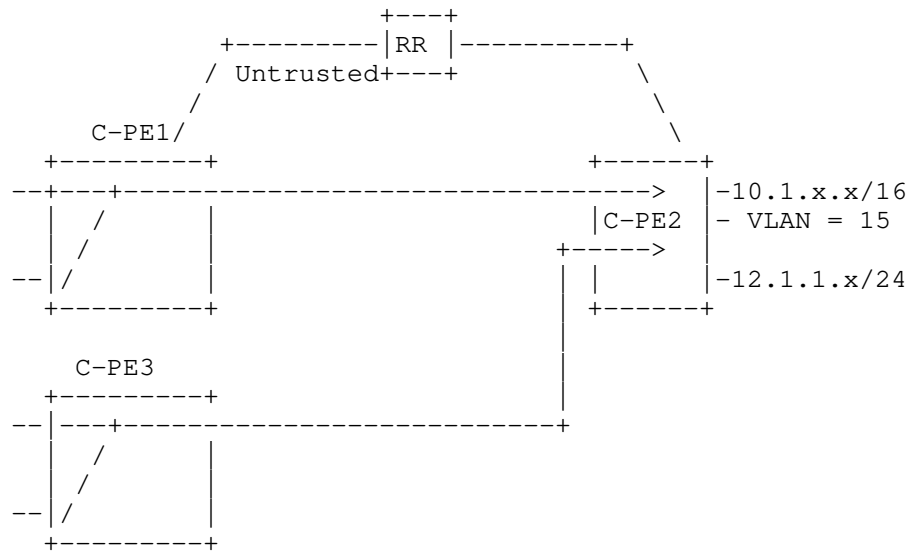


Figure 5: Homogeneous SDWAN

The BGP UPDATE Message from C-PE2 to RR should have the client routes encoded in the MP-NLRI Path Attribute and the IPsec Tunnel associated information encoded in the Tunnel-Encap Path Attributes as described in the [SECURE-EVPN]:

- MP-NLRI Path Attribute: to indicate multiple routes attached to the C-PE2:
  - 10.1.x.x/16
  - VLAN #15
  - 12.1.1.x/24
- Tunnel-Encap Path Attribute: to describe the IPsec attributes for routes encoded in the NLRI Path Attribute:
  - IPsec attributes for remote nodes to establish the IPsec tunnel to C-PE2.

If different client routes attached to C-PE2 needs to be reached by separate IPsec tunnels, then multiple BGP UPDATE messages need to be sent to the remote nodes via RR. If C-PE2 doesn't have the policy on authorized peers for the specific client routes, RR needs to check the client routes policies to propagate the BGP UPDATE messages to the remote authorized edge nodes.

There could be policies governing the topologies of a client's different routes attached to an edge node. For example, VLAN #25 and

route 22.1.1.x/24 could be the Payment Applications described in the Section 3.1.2 that can only communicate with Payment Gateway attached to C-PE3. If C-PEs don't have the policy to govern the communication peers, RR can take over the responsibility of only send BGP UPDATE to the authorized peers.

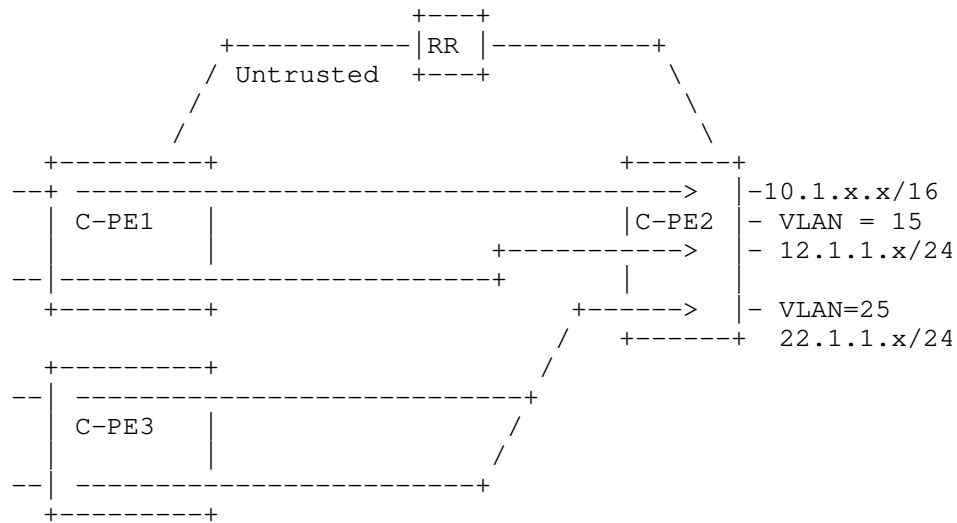


Figure 6: (see \*.pdf for more accurate figure)

#### UPDATE 1:

- MP-NLRI Path Attribute:  
10.1.x.x/16  
VLAN #15  
12.1.1.x/24
- Tunnel-Encap Path Attribute:  
IPsec SA attributes for IPsec tunnels to C-PE2 from any node for reaching 10.1.x.x/16, VLAN #15, and 12.1.1.x/24.

#### UPDATE 2 (only sent to C-PE3)

- MP-NLRI Path Attribute:  
VLAN #25  
22.1.1.x/24
- Tunnel-Encap:

IPsec SA attributes for IPsec tunnels to C-PE2 from C-PE3 for reaching VLAN #25 and subnet 22.1.1./24.

#### 4.2. BGP Walk Through for Application Flow Based Segmentation

If the applications are assigned with unique IP addresses, the Application Flow based Segmentation described in Section 3.1.2 can be achieved by advertising different BGP UPDATE messages to different nodes. In the Figure below, the following BGP Updates can be advertised to ensure that Payment Application only communicates with the Payment Gateway:

BGP UPDATE #1 from C-PE2 to RR for the P2P topology that is only propagated to Payment GW node:

- MP-NLRI Path Attribute:
  - 30.1.1.x/24
- Tunnel Encap Path Attribute
  - IPsec Attributes for PaymentGW ->C-PE2

BGP UPDATE #2 from C-PE2 to RR for the routes to be reached by C-PE1 and C-PE2:

- MP-NLRI Path Attribute:
  - 10.1.x.x
  - 12.4.x.x
- Tunnel-Encap Path Attribute:
  - Any node to C-PE2

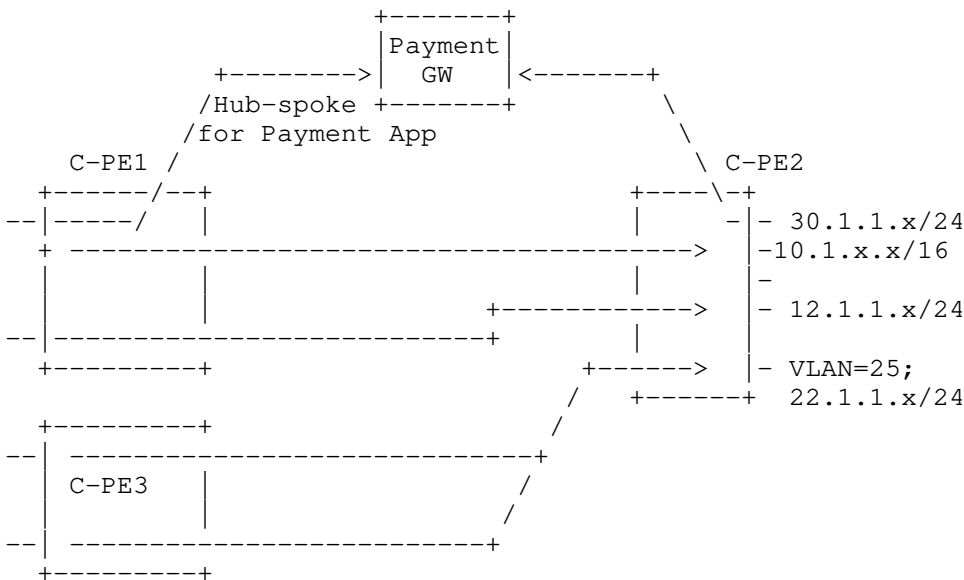


Figure 7: Application Based SDWAN Segmentation

4.3. Client Service Provisioning Model

The provisioning tasks described in Section 4 of RFC8388 are the same for the SDWAN client traffic. When client traffic is multi-homed to two (or more) C-PEs, the Non-Service-Specific parameters need to be provisioned per the Section 4.1.1 of RFC8388.

Since some SDWAN nodes are ephemeral and have small number of IP subnets or VLANs attached to their client ports, it is recommended to have default and simplified Service-specific parameters for each client port, remotely managed by the SDWAN Network Controller via the secure channel (TLS/DTLS) between the controller and the C-PEs.

#### 4.4. WAN Ports Provisioning Model

Since the deployment of PEs to MPLS VPN are for relatively long term, the common provisioning procedure for PE's WAN ports is via CLI.

A SDWAN node deployment can be ephemeral and its location can be in remote locations, manual provisioning for its WAN ports is not acceptable. In addition, a SDWAN WAN port's IP address can be dynamically assigned or using private addresses. Therefore, it is necessary to have a separate control protocol; something like NHRP did for ATM, for a SDWAN node to register its WAN property to its controller dynamically.

Unlike a PE to MPLS based VPN where its WAN ports are homogeneously facing MPLS private network and all traffic are egressed in MPLS data frames through its WAN ports, the WAN ports of a SDWAN node can be connected to a PE of VPN with Ethernet/IP, MPLS private network directly via MPLS headers, or the public Internet.

For Scenario #1 described in Section 3.2, the WAN ports can face public internet or VPN.

For Scenario #2 described in Section 3.3, WAN ports are either configured as connecting to PEs of VPN where traffic can be sent as IP/Ethernet without encryption, or configured as connecting to public Internet that requires encryption for packets egress out.

For Scenario #3 described in Section 3.4, the WAN ports are either configured as VPN egress ports (hand off MPLS data frames), or as connecting to the public internet that requires MPLS in IP in IPsec encapsulation.

#### 4.5. Why BGP as Control Plane for SDWAN?

For a small sized SDWAN network, traditional hub & spoke model using NHRP or DSVPN/DMVPN with a hub node (or controller) managing SDWAN node WAN ports mapping (e.g. local & public addresses and tunnel identifiers mapping) can work reasonably well. However, for a large SDWAN network, say more than 100 nodes with different types of topologies, the traditional approach becomes very messy, complex and error prone.

Here are some of the compelling reasons of using BGP instead of extending NHRP/DSVPN/DMVPN. (Same as the reasons quoted by LSVR on why using BGP):

- BGP has the built-in capability to constrain the propagation of SDWAN edge node properties to a small number of edge nodes [RFC4684].
- RR already has the capability to apply policies to communications among peers.
- BGP is widely deployed as sole protocol (see RFC 7938)
- Robust and simple implementation
- Wide acceptance - minimal learning
- Reliable transport
- Guaranteed in-order delivery
- Incremental updates
- Incremental updates upon session restart
- No flooding and selective filtering

## 5. SDWAN Traffic Forwarding Walk Through

BGP based EVPN control plane are still applicable to routes attached to the client ports of SDWAN nodes. Section 5 of RFC8388 describes the BGP EVPN NLRI Usage for various routes of client traffic. The procedures described in the Section 6 of RFC8388 are same for the SDWAN client traffic.

The only additional consideration for SDWAN is to control how traffic egress the SDWAN edge node to various WAN ports.

### 5.1. SDWAN Network Startup Procedures

A SDWAN network can add or delete SDWAN edge nodes on regular basis depending on user requests.

- For Scenario #1: a SDWAN edge node in a shopping mall or Cloud DC can be added or removed on demand. The Zero Touch Provisioning described in 3.1.2 are required for the node startup.
- For Scenario #2: this can be Data Centers or enterprises upgrading their CPEs to add extra bandwidth via public internet in addition to VPN services that they already purchased. Before the node powers up

or upgraded, there should be links connected to the PEs of a provider VPNs.

- For Scenario #3, the Internet facing WAN ports are added to (or removed from) existing VPN PEs.

## 5.2. Packet Walk-Through for Scenario #1

Upon power up, a SDWAN node can learn client routes from the Client facing ports, in the same way as EVPN described in RFC8388. Controller facilitates the IPsec SA establishment and rekey management as described in [SECURE-EVPN]. Controller manages how client's routes are associated with individual IPsec SA.

[SECURE-EVPN] describes a solution for SDWAN Scenario #1. It utilizes the BGP RR to facilitate the key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

When C-PEs do not support MPLS, the approaches described by RFC8365 can be used, with addition of IPsec encrypting the IP packets when sending packets over the Black Interfaces.

## 5.3. Packet Walk-Through for Scenario #2

In this scenario, C-PEs have some WAN ports connected to the public internet and some WAN ports with direct connect to PEs of trusted VPN. The C-PEs in Scenario #2 have the plain IP/Ethernet data frames egress to the PEs of the VPN, encrypted data frames egress the WAN ports facing the public Internet.

Users specify the policy or criteria on which flows can only egress WAN ports facing the trusted VPN without encryption, which can egress the WAN ports facing the public Internet with encryption, or which can egress WAN ports facing the public Internet without encryption.

The internet facing WAN ports can face potential DDoS attacks, additional anti-DDoS mechanism has to be enabled on those WAN ports and the Control Plane should not learn routes from the Public Network facing WAN ports.

For the Scenario #2, if a client route can be reached by MPLS VPN and IPsec Tunnel via public network, the BGP UPDATE for the client

route should indicate all available tunnels in the Tunnel Path Attribute of the BGP NLRI.

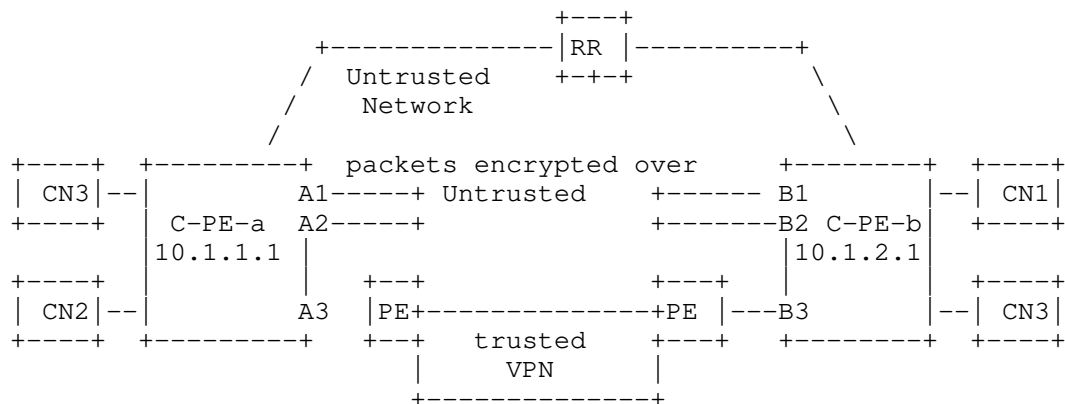


Figure 8: SDWAN Scenario #2

For example, if the CN1 route can be reached by both VPN and Public internet, the CN1's BGP route UPDATE should include the following:

- MP-NLRI Path Attribute:

CN1

- Tunnel-Encap Path Attribute:

Tunnel 1: MPLS-in-GRE encapsulation

With the MPLS-in-GRE Sub-TLV specified by Tunnel-Encap;

Tunnel 2: IPsec-GRE encapsulation

With the IPsec Sub-TLVs specified by the [SECURE-EVPN] and [BGP-EDGE-DISCOVERY]

There could be multiple IPsec SA tunnels terminated at the edge node loopback address or terminated at WAN ports. For the Scenario #2, there can be policies to determine which IPsec SA tunnels that the client route can be carried. When a client route can be carried by multiple IPsec SA tunnels terminated by two different WAN ports, multiple Tunnel Path Attributes with different Tunnel-end-point Sub-TLVs need to be included in the NLRI of the BGP UPDATE for the client route.



#### 5.4. Packet Walk-Through for Scenario #3

The behavior described in [SECURE-L3VPN] applies to this scenario.

[SECURE-L3VPN] describes how to extend the RFC4364 VPN to allow some PEs being connected to other PEs via public networks. In this scenario, the PEs is the SDWAN Edge nodes. [SECURE-L3VPN] introduces the concept of RED Interface & Black Interface on those PEs. RED interfaces face the VPN over which packets can be forwarded natively without encryption. Black Interfaces face public network over which only IPsec-protected packets are forwarded. [SECURE-L3VPN] assumes PEs terminate MPLS packets, and use MPLS over IPsec when sending over the Black Interfaces.

The C-PEs not only have RED interfaces facing clients but also have RED interface facing MPLS backbone, with additional BLACK interfaces facing the untrusted public networks for the WAN side. The C-PEs cannot mix the routes learned from the Black Interfaces with the Routes from RED Interfaces. The routes learned from core-facing RED interfaces are for underlay and cannot be mixed with the routes learned over access-facing RED interfaces that are for overlay. Furthermore, the routes learned over core-facing interfaces (both RED and BLACK) can be shared in the same GLOBAL route table.

There may be some added risks of the packets from the ports facing the Internet. Therefore, special consideration has to be given to the routes from WAN ports facing the Internet. RFC4364 describes using an RD to create different routes for reaching same system. A similar approach can be considered to force packets received from the Internet facing ports to go through special security functions before being sent over to the VPN backbone WAN ports.

#### 6. Manageability Considerations

SDWAN overlay networks utilize the SDWAN controller to facilitate route distribution, central configurations, and others. SDWAN Edge nodes need to advertise the attached routes to their controller (i.e. RR in BGP case).

#### 7. Security Considerations

Having WAN ports facing the public Internet introduces the following security risks:

- 1) Potential DDoS attack to the C-PEs with ports facing internet.
- 2) Potential risk of provider VPN network being injected with illegal traffic coming from the public Internet WAN ports on the C-PEs.

## 8. IANA Considerations

None

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4364] E. rosen, Y. Rekhter, "BGP/MPLS IP Virtual Private networks (VPNs)", Feb 2006.
- [RFC7296] C. Kaufman, et al, "Internet Key Exchange Protocol Version 2 (IKEv2)", Oct 2014.
- [RFC7432] A. Sajassi, et al, "BGP MPLS-Based Ethernet VPN", Feb 2015.
- [RFC8365] A. Sajassi, et al, "A network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", March 2018.

### 9.2. Informative References

- [RFC8192] S. Hares, et al, "Interface to Network Security Functions (I2NSF) Problem Statement and Use Cases", July 2017
- [RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.
- [BGP-SDWAN-Port] L. Dunbar, H. Wang, W. Hao, "BGP Extension for SDWAN Overlay Networks", draft-dunbar-idr-bgp-sdwan-overlay-ext-03, work-in-progress, Nov 2018.

- [Net2Cloud-Gap] L. Dunbar, A. Malis, C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work in progress, Oct. 2018.
- [SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-dunbar-idr-sdwan-edge-discovery-00, work-in-progress, July 2020.
- [VPN-over-Internet] E. Rosen, "Provide Secure Layer L3VPNs over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, work-in-progress, July 2018
- [DMVPN] Dynamic Multi-point VPN:  
<https://www.cisco.com/c/en/us/products/security/dynamic-multipoint-vpn-dmvpn/index.html>
- [DSVPN] Dynamic Smart VPN:  
<http://forum.huawei.com/enterprise/en/thread-390771-1-1.html>
- [SECURE-EVPN] A. Sajassi, et al, "Secure EVPN", draft-sajassi-bess-secure-evpn-01, Work-in-progress, March 2019.
- [SECURE-L3VPN] E. Rosen, R. Bonica, "Secure Layer L3VPN over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, Work-in-progress, June 2018.
- [ITU-T-X1036] ITU-T Recommendation X.1036, "Framework for creation, storage, distribution and enforcement of policies for network security", Nov 2007.
- [Net2Cloud-Problem] L. Dunbar and A. Malis, "Seamless Interconnect Underlay to Cloud Overlay Problem Statement", draft-dm-net2cloud-problem-statement-02, June 2018
- [Net2Cloud-gap] L. Dunbar, A. Malis, and C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work-in-progress, Aug 2018.
- [Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

## 10. Acknowledgments

Acknowledgements to Jim Guichard, John Scudder, Darren Dukes, Andy Malis and Donald Eastlake for their review and contributions.

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

Authors' Addresses

Linda Dunbar  
Futurewei  
Email: ldunbar@futurewei.com

James Guichard  
Futurewei  
Email: james.n.guichard@futurewei.com

Ali Sajassi  
Cisco  
Email: sajassi@cisco.com

John Drake  
Juniper  
Email: jdrake@juniper.net

Basil Najem  
Bell Canada  
Email: basil.najem@bell.ca

David Carrel  
Cisco  
Email: carrel@cisco.com

Ayan Banerjee  
Cisco  
Email: ayabaner@cisco.com



BESS Working Group  
Internet-Draft  
Intended Status: Standards Track

Ali Sajassi  
Samir Thoria  
Cisco  
Keyur Patel  
Arrcus  
John Drake  
Wen Lin  
Juniper

Expires: April 2, 2020

September 30, 2019

IGMP and MLD Proxy for EVPN  
draft-ietf-bess-evpn-igmp-mld-proxy-04

Abstract

Ethernet Virtual Private Network (EVPN) solution is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) and DC interconnect (DCI) services, and in service provider (SP) applications for next generation virtual private LAN services.

This draft describes how to support efficiently endpoints running IGMP for the above services over an EVPN network by incorporating IGMP proxy procedures on EVPN PEs.

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) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1	Introduction . . . . .	4
2	Specification of Requirements . . . . .	5
3	Terminology . . . . .	5
4	IGMP/MLD Proxy . . . . .	6
4.1	Proxy Reporting . . . . .	6
4.1.1	IGMP/MLD Membership Report Advertisement in BGP . . . . .	7
4.1.2	IGMP/MLD Leave Group Advertisement in BGP . . . . .	8
4.2	Proxy Querier . . . . .	9
5	Operation . . . . .	9
5.1	PE with only attached hosts/VMs for a given subnet . . . . .	10
5.2	PE with a mix of attached hosts/VMs and multicast source . . . . .	11
5.3	PE with a mix of attached hosts/VMs, a multicast source and a router . . . . .	11
6	All-Active Multi-Homing . . . . .	11
6.1	Local IGMP/MLD Join Synchronization . . . . .	12
6.2	Local IGMP/MLD Leave Group Synchronization . . . . .	12
6.2.1	Remote Leave Group Synchronization . . . . .	13
6.2.2	Common Leave Group Synchronization . . . . .	14
6.3	Mass Withdraw of Multicast join Sync route in case of failure . . . . .	14
7	Single-Active Multi-Homing . . . . .	14
8	Selective Multicast Procedures for IR tunnels . . . . .	14
9	BGP Encoding . . . . .	15
9.1	Selective Multicast Ethernet Tag Route . . . . .	15
9.1.1	Constructing the Selective Multicast Ethernet Tag route . . . . .	17



9.1.2 Default Selective Multicast Route . . . . .	18
9.2 Multicast Join Synch Route . . . . .	19
9.2.1 Constructing the Multicast Join Synch Route . . . . .	21
9.3 Multicast Leave Synch Route . . . . .	22
9.3.1 Constructing the Multicast Leave Synch Route . . . . .	24
9.4 Multicast Flags Extended Community . . . . .	25
9.5 EVI-RT Extended Community . . . . .	26
9.6 Rewriting of RT ECs and EVI-RT ECs by ASBRs . . . . .	29
10 IGMP/MLD Immediate Leave . . . . .	29
11 IGMP Version 1 Membership Request . . . . .	29
12 Security Considerations . . . . .	30
13 IANA Considerations . . . . .	30
14 References . . . . .	30
14.1 Normative References . . . . .	30
14.2 Informative References . . . . .	31
15 Acknowledgement . . . . .	31
16 Contributors . . . . .	32
Authors' Addresses . . . . .	32

## 1 Introduction

Ethernet Virtual Private Network (EVPN) solution [RFC7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) and DC interconnect (DCI) services, and in service provider (SP) applications for next generation virtual private LAN services.

In DC applications, a point of delivery (POD) can consist of a collection of servers supported by several top of rack (TOR) and Spine switches. This collection of servers and switches are self contained and may have their own control protocol for intra-POD communication and orchestration. However, EVPN is used as standard way of inter-POD communication for both intra-DC and inter-DC. A subnet can span across multiple PODs and DCs. EVPN provides robust multi-tenant solution with extensive multi-homing capabilities to stretch a subnet (VLAN) across multiple PODs and DCs. There can be many hosts/VMs (several hundreds) attached to a subnet that is stretched across several PODs and DCs.

These hosts/VMs express their interests in multicast groups on a given subnet/VLAN by sending IGMP membership reports (Joins) for their interested multicast group(s). Furthermore, an IGMP router periodically sends membership queries to find out if there are hosts on that subnet that are still interested in receiving multicast traffic for that group. The IGMP/MLD Proxy solution described in this draft accomplishes has three objectives:

1) Reduce flooding of IGMP messages: just like the ARP/ND suppression mechanism in EVPN to reduce the flooding of ARP messages over EVPN, it is also desired to have a mechanism to reduce the flooding of IGMP messages (both Queries and Reports) in EVPN.

2) Distributed anycast multicast proxy: it is desirable for the EVPN network to act as a distributed anycast multicast router with respect to IGMP/MLD proxy function for all the hosts attached to that subnet.

3) Selective Multicast: to forward multicast traffic over EVPN network such that it only gets forwarded to the PEs that have interest in the multicast group(s), multicast traffic will not be forwarded to the PEs that have no receivers attached to them for that multicast group. This draft shows how this objective may be achieved when Ingress Replication is used to distribute the multicast traffic among the PEs. Procedures for supporting selective multicast using P2MP tunnels can be found in [bum-procedure-updates]

The first two objectives are achieved by using IGMP/MLD proxy on the

PE and the third objective is achieved by setting up a multicast tunnel (e.g., ingress replication) only among the PEs that have interest in that multicast group(s) based on the trigger from IGMP/MLD proxy processes. The proposed solutions for each of these objectives are discussed in the following sections.

## 2 Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 3 Terminology

POD: Point of Delivery

ToR: Top of Rack

NV: Network Virtualization

NVO: Network Virtualization Overlay

EVPN: Ethernet Virtual Private Network

IGMP: Internet Group Management Protocol

MLD: Multicast Listener Discovery

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE

IR: Ingress Replication

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet Segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet Segment is called an 'Ethernet Segment Identifier'.

PE: Provider Edge.

BD: Broadcast Domain. As per [RFC7432], an EVI consists of a single or multiple BDs. In case of VLAN-bundle and VLAN-aware bundle service model, an EVI contains multiple BDs. Also, in this document, BD and subnet are equivalent terms.

Ethernet Tag: An Ethernet tag identifies a particular broadcast domain, e.g., a VLAN. An EVPN instance consists of one or more broadcast domains.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, 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 known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

This document also assumes familiarity with the terminology of [RFC7432]. Though most of the place this document uses term IGMP membership request (Joins), the text applies equally for MLD membership request too. Similarly, text for IGMPv2 applies to MLDv1 and text for IGMPv3 applies to MLDv2. IGMP / MLD version encoding in BGP update is stated in section 9

#### 4 IGMP/MLD Proxy

The IGMP Proxy mechanism is used to reduce the flooding of IGMP messages over an EVPN network similar to ARP proxy used in reducing the flooding of ARP messages over EVPN. It also provides a triggering mechanism for the PEs to setup their underlay multicast tunnels. The IGMP Proxy mechanism consists of two components: a) Proxy for IGMP Reports and b) Proxy for IGMP Queries.

##### 4.1 Proxy Reporting

When IGMP protocol is used between hosts/VMs and their first hop EVPN router (EVPN PE), Proxy-reporting is used by the EVPN PE to summarize (when possible) reports received from downstream hosts and propagate them in BGP to other PEs that are interested in the information. This is done by terminating the IGMP Reports in the first hop PE, and translating and exchanging the relevant information among EVPN BGP speakers. The information is again translated back to IGMP message at the recipient EVPN speaker. Thus it helps create an IGMP overlay subnet using BGP. In order to facilitate such an overlay, this document also defines a new EVPN route type NLRI, the EVPN Selective

Multicast Ethernet Tag route, along with its procedures to help exchange and register IGMP multicast groups [section 7].

#### 4.1.1 IGMP/MLD Membership Report Advertisement in BGP

When a PE wants to advertise an IGMP membership report (Join) using the BGP EVPN route, it follows the following rules (BGP encoding stated in section 9.1):

1) When the first hop PE receives several IGMP membership reports (Joins), belonging to the same IGMP version, from different attached hosts/VMs for the same (\*,G) or (S,G), it only SHOULD send a single BGP message corresponding to the very first IGMP Join (BGP update as soon as possible) for that (\*,G) or (S,G). This is because BGP is a stateful protocol and no further transmission of the same report is needed. If the IGMP Join is for (\*,G), then multicast group address MUST be sent along with the corresponding version flag (v2 or v3) set. In case of IGMPv3, the exclude flag MUST also need to be set to indicate that no source IP address to be excluded (include all sources"\*").

If the IGMP Join is for (S,G), then besides setting multicast group address along with the version flag v3, the source IP address and the include/exclude flag MUST be set. It should be noted that when advertising the EVPN route for (S,G), the only valid version flag is v3 (v1 and v2 flags MUST be set to zero).

2) When the first hop PE receives an IGMPv3 Join for (S,G) on a given BD, it SHOULD advertise the corresponding EVPN Selective Multicast Ethernet Tag (SMET) route regardless of whether the source (S) is attached to itself or not in order to facilitate the source move in the future.

3) When the first hop PE receives an IGMP version-X Join first for (\*,G) and then later it receives an IGMP version-Y Join for the same (\*,G), then it MUST re-advertise the same EVPN SMET route with flag for version-Y set in addition to any previously-set version flag(s). In other words, the first hop PE MUST not withdraw the EVPN route before sending the new route because the flag field is not part of BGP route key processing.

4) When the first hop PE receives an IGMP version-X Join first for (\*,G) and then later it receives an IGMPv3 Join for the same multicast group address but for a specific source address S, then the PE MUST advertise a new EVPN SMET route with v3 flag set (and v1 and v2 reset). The include/exclude flag also need to be set accordingly. Since source IP address is used as part of BGP route key processing,

it is considered as a new BGP route advertisement.

5) When a PE receives an EVPN SMET route with more than one version flag set, it will generate the corresponding IGMP report for (\*,G) for each version specified in the flags field. With multiple version flags set, there MUST not be source IP address in the receive EVPN route. If there is, then an error SHOULD be logged. If the v3 flag is set (in addition to v2), then the include/exclude flag MUST indicate "exclude". If not, then an error SHOULD be logged. The PE MUST generate an IGMP membership report (Join) for that (\*,G) and each IGMP version in the version flag.

6) When a PE receives a list of EVPN SMET NLRIs in its BGP update message, each with a different source IP address and the same multicast group address, and the version flag is set to v3, then the PE generates an IGMPv3 membership report with a record corresponding to the list of source IP addresses and the group address along with the proper indication of inclusion/exclusion.

7) Upon receiving EVPN SMET route(s) and before generating the corresponding IGMP Join(s), the PE checks to see whether it has any CE multicast router for that BD on any of its ES's. The PE provides such a check by listening for PIM Hello messages on that AC (i.e., <ES,BD>). If the PE does have the router's ACs, then the generated IGMP Join(s) are sent to those ACs. If it doesn't have any of the router's AC, then no IGMP Join(s) needs to be generated. This is because sending IGMP Joins to other hosts can result in unintentionally preventing a host from joining a specific multicast group using IGMPv2 - i.e., if the PE does not receive a join from the host it will not forward multicast data to it. Per [RFC4541], when an IGMPv2 host receives a membership report for a group address that it intends to join, the host will suppress its own membership report for the same group, and if the PE does not receive an IGMP Join from host it will not forward multicast data to it. In other words, an IGMPv2 Join MUST NOT be sent on an AC that does not lead to a CE multicast router. This message suppression is a requirement for IGMPv2 hosts. This is not a problem for hosts running IGMPv3 because there is no suppression of IGMP Membership reports.

#### 4.1.2 IGMP/MLD Leave Group Advertisement in BGP

When a PE wants to withdraw an EVPN SMET route corresponding to an IGMPv2 Leave Group (Leave) or IGMPv3 "Leave" equivalent message, it follows the following rules:

1) When a PE receives an IGMPv2 Leave Group or its "Leave" equivalent

message for IGMPv3 from its attached host, it checks to see if this host is the last host that is interested in this multicast group by sending a query for the multicast group. If the host was indeed the last one (i.e. no responses are received for the query), then the PE MUST re-advertises EVPN SMET Multicast route with the corresponding version flag reset. If this is the last version flag to be reset, then instead of re-advertising the EVPN route with all version flags reset, the PE MUST withdraws the EVPN route for that (\*,G).

2) When a PE receives an EVPN SMET route for a given (\*,G), it compares the received version flags from the route with its per-PE stored version flags. If the PE finds that a version flag associated with the (\*,G) for the remote PE is reset, then the PE MUST generate IGMP Leave for that (\*,G) toward its local interface (if any) attached to the multicast router for that multicast group. It should be noted that the received EVPN route SHOULD at least have one version flag set. If all version flags are reset, it is an error because the PE should have received an EVPN route withdraw for the last version flag. Error MUST be considered as BGP error and SHOULD be handled as per [RFC7606].

3) When a PE receives an EVPN SMET route withdraw, it removes the remote PE from its OIF list for that multicast group and if there are no more OIF entries for that multicast group (either locally or remotely), then the PE MUST stop responding to queries from the locally attached router (if any). If there is a source for that multicast group, the PE stops sending multicast traffic for that source.

#### 4.2 Proxy Querier

As mentioned in the previous sections, each PE MUST have proxy querier functionality for the following reasons:

- 1) To enable the collection of EVPN PEs providing L2VPN service to act as distributed multicast router with Anycast IP address for all attached hosts/VMs in that subnet.
- 2) To enable suppression of IGMP membership reports and queries over MPLS/IP core.

#### 5 Operation

Consider the EVPN network of Figure-1, where there is an EVPN instance configured across the PEs shown in this figure (namely PE1, PE2, and PE3). Let's consider that this EVPN instance consists of a single bridge domain (single subnet) with all the hosts, sources, and

the multicast router connected to this subnet. PE1 only has hosts connected to it. PE2 has a mix of hosts and a multicast source. PE3 has a mix of hosts, a multicast source, and a multicast router. Furthermore, let's consider that for (S1,G1), R1 is used as the multicast router. The following subsections describe the IGMP proxy operation in different PEs with regard to whether the locally attached devices for that subnet are:

- only hosts/VMs
- mix of hosts/VMs and multicast source
- mix of hosts/VMs, multicast source, and multicast router

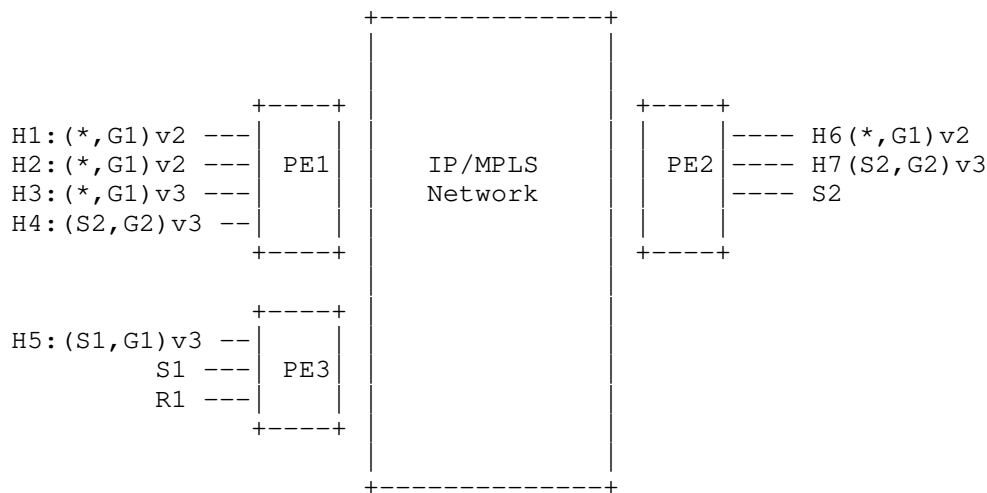


Figure 1: EVPN network

#### 5.1 PE with only attached hosts/VMs for a given subnet

When PE1 receives an IGMPv2 Join Report from H1, it does not forward this join to any of its other ports (for this subnet) because all these local ports are associated with the hosts/VMs. PE1 sends an EVPN Multicast Group route corresponding to this join for (\*,G1) and setting v2 flag. This EVPN route is received by PE2 and PE3 that are the members of the same BD (i.e., same EVI in case of VLAN-based service or <EVI,VLAN> in case of VLAN-aware bundle service). PE3 reconstructs the IGMPv2 Join Report from this EVPN BGP route and only sends it to the port(s) with multicast routers attached to it (for



that subnet). In this example, PE3 sends the reconstructed IGMPv2 Join Report for (\*,G1) only to R1. Furthermore, even though PE2 receives the EVPN BGP route, it does not send it to any of its ports for that subnet; viz, ports associated with H6 and H7.

When PE1 receives the second IGMPv2 Join from H2 for the same multicast group (\*,G1), it only adds that port to its OIF list but it doesn't send any EVPN BGP route because there is no change in information. However, when it receives the IGMPv3 Join from H3 for the same (\*,G1). Besides adding the corresponding port to its OIF list, it re-advertises the previously sent EVPN SMET route with the v3 & exclude flag set.

Finally when PE1 receives the IMGMPv3 Join from H4 for (S2,G2), it advertises a new EVPN SMET route corresponding to it.

## 5.2 PE with a mix of attached hosts/VMs and multicast source

The main difference in this case is that when PE2 receives the IGMPv3 Join from H7 for (S2,G2), it does advertise it in BGP to support source move even though PE2 knows that S2 is attached to its local AC. PE2 adds the port associated with H7 to its OIF list for (S2,G2). The processing for IGMPv2 received from H6 is the same as the IGMPv2 Join described in previous section.

## 5.3 PE with a mix of attached hosts/VMs, a multicast source and a router

The main difference in this case relative to the previous two sections is that IGMP v2/v3 Join messages received locally needs to be sent to the port associated with router R1. Furthermore, the Joins received via BGP (SMET) need to be passed to the R1 port but filtered for all other ports.

## 6 All-Active Multi-Homing

Because the LAG flow hashing algorithm used by the CE is unknown at the PE, in an All-Active redundancy mode it must be assumed that the CE can send a given IGMP message to any one of the multi-homed PEs, either DF or non-DF; i.e., different IGMP Join messages can arrive at different PEs in the redundancy group and furthermore their corresponding Leave messages can arrive at PEs that are different from the ones that received the Join messages. Therefore, all PEs attached to a given ES must coordinate IGMP Join and Leave Group (x,G) state, where x may be either '\*' or a particular source S, for each BD on that ES. This allows the DF for that [ES,BD] to correctly advertise or withdraw a Selective Multicast Ethernet Tag (SMET) route for that (x,G) group in that BD when needed.

All-Active multihoming PEs for a given ES MUST support IGMP synchronization procedures described in this section if they need to perform IGMP proxy for hosts connected to that ES.

### 6.1 Local IGMP/MLD Join Synchronization

When a PE, either DF or non-DF, receives on a given multihomed ES operating in All-Active redundancy mode, an IGMP Membership Report for (x,G), it determines the BD to which the IGMP Membership Report belongs. If the PE doesn't already have local IGMP Join (x,G) state for that BD on that ES, it MUST instantiate local IGMP Join (x,G) state and MUST advertise a BGP IGMP Join Synch route for that [ES, BD]. Local IGMP Join (x, G) state refers to IGMP Join (x,G) state that is created as a result of processing an IGMP Membership Report for (x,G).

The IGMP Join Synch route MUST carry the ES-Import RT for the ES on which the IGMP Membership Report was received. Thus it MUST only be sent to the PEs attached to that ES and not any other PEs.

When a PE, either DF or non-DF, receives an IGMP Join Synch route it installs that route and if it doesn't already have IGMP Join (x,G) state for that [ES,BD], it MUST instantiate that IGMP Join (x,G) state - i.e., IGMP Join (x,G) state is the union of the local IGMP Join (x,G) state and the installed IGMP Join Synch route. If the DF did not already advertise (originate) a SMET route for that (x,G) group in that BD, it MUST do so now.

When a PE, either DF or non-DF, deletes its local IGMP Join (x, G) state for that [ES,BD], it MUST withdraw its BGP IGMP Join Synch route for that [ES,BD].

When a PE, either DF or non-DF, receives the withdrawal of an IGMP Join Synch route from another PE it MUST remove that route. When a PE has no local IGMP Join (x,G) state and it has no installed IGMP Join Synch routes, it MUST remove IGMP Join (x,G) state for that [ES, BD]. If the DF no longer has IGMP Join (x,G) state for that BD on any ES for which it is DF, it MUST withdraw its SMET route for that (x,G) group in that BD.

In other words, a PE advertises an SMET route for that (x,G) group in that BD when it has IGMP Join (x,G) state in that BD on at least one ES for which it is DF and it withdraws that SMET route when it does not have IGMP Join (x,G) state in that BD on any ES for which it is DF.

### 6.2 Local IGMP/MLD Leave Group Synchronization

When a PE, either DF or non-DF, receives, on a given multihomed ES operating in All-Active redundancy mode, an IGMP Leave Group message for (x,G) from the attached CE, it determines the BD to which the IGMPv2 Leave Group belongs. Regardless of whether it has IGMP Join (x,G) state for that [ES,BD], it initiates the (x,G) leave group synchronization procedure, which consists of the following steps:

- 1) It computes the Maximum Response Time, which is the duration of (x,G) leave group synchronization procedure. This is the product of two locally configured values, Last Member Query Count and Last Member Query Interval (described in Section 3 of [RFC2236]), plus a delta corresponding to the time it takes for a BGP advertisement to propagate between the PEs attached to the multihomed ES (delta is a consistently configured value on all PEs attached to the multihomed ES).
- 2) It starts the Maximum Response Time timer. Note that the receipt of subsequent IGMP Leave Group messages or BGP Leave Synch routes for (x,G) do not change the value of a currently running Maximum Response Time timer and are ignored by the PE.
- 3) It initiates the Last Member Query procedure described in Section 3 of [RFC2236]; viz, it sends a number of Group-Specific Query (x,G) messages (Last Member Query Count) at a fixed interval (Last Member Query Interval) to the attached CE.
- 4) It advertises an IGMP Leave Synch route for that that [ES,BD]. This route notifies the other multihomed PEs attached to the given multihomed ES that it has initiated an (x,G) leave group synchronization procedure; i.e., it carries the ES-Import RT for the ES on which the IGMP Leave Group was received. It also contains the Maximum Response Time and the Leave Group Synchronization Procedure Sequence number. The latter identifies the specific (x,G) leave group synchronization procedure initiated by the advertising PE, which increments the value whenever it initiates a procedure.
- 5) When the Maximum Response Timer expires, the PE that has advertised the IGMP Leave Synch route withdraws it.

#### 6.2.1 Remote Leave Group Synchronization

When a PE, either DF or non-DF, receives an IGMP Leave Synch route it installs that route and it starts a timer for (x,G) on the specified [ES,BD] whose value is set to the Maximum Response Time in the received IGMP Leave Synch route. Note that the receipt of subsequent IGMPv2 Leave Group messages or BGP Leave Synch routes for (x,G) do not change the value of a currently running Maximum Response Time

timer and are ignored by the PE.

#### 6.2.2 Common Leave Group Synchronization

If a PE attached to the multihomed ES receives an IGMP Membership Report for (x,G) before the Maximum Response Time timer expires, it advertises a BGP IGMP Join Synch route for that [ES,BD]. If it doesn't already have local IGMP Join (x, G) state for that [ES, BD], it instantiates local IGMP Join (x,G) state. If the DF is not currently advertising (originating) a SMET route for that (x,G) group in that BD, it does so now.

If a PE attached to the multihomed ES receives an IGMP Join Synch route for (x,G) before the Maximum Response Time timer expires, it installs that route and if it doesn't already have IGMP Join (x,G) state for that BD on that ES, it instantiates that IGMP Join (x,G) state. If the DF has not already advertised (originated) a SMET route for that (x,G) group in that BD, it does so now.

When the Maximum Response Timer expires a PE that has advertised an IGMP Leave Synch route, withdraws it. Any PE attached to the multihomed ES, that started the Maximum Response Time and has no local IGMP Join (x,G) state and no installed IGMP Join Synch routes, it removes IGMP Join (x,G) state for that [ES,BD]. If the DF no longer has IGMP Join (x,G) state for that BD on any ES for which it is DF, it withdraws its SMET route for that (x,G) group in that BD.

#### 6.3 Mass Withdraw of Multicast join Sync route in case of failure

A PE which has received an IGMP Join, would have synced the IGMP Join by the procedure defined in section 6.1. If a PE with local join state goes down or the PE to CE link goes down, it would lead to a mass withdraw of multicast routes. Remote PEs (PEs where these routes were remote IGMP Joins) SHOULD not remove the state immediately; instead General Query SHOULD be generated to refresh the states. There are several ways to Some of the way to detect failure at a peer, e.g. using IGP next hop tracking or ES route withdraw.

#### 7 Single-Active Multi-Homing

Note that to facilitate state synchronization after failover, the PEs attached to a multihomed ES operating in Single-Active redundancy mode SHOULD also coordinate IGMP Join (x,G) state. In this case all IGMP Join messages are received by the DF and distributed to the non-DF PEs using the procedures described above.

#### 8 Selective Multicast Procedures for IR tunnels

If an ingress PE uses ingress replication, then for a given (x,G) group in a given BD:

- 1) It sends (x,G) traffic to the set of PEs not supporting IGMP Proxy. This set consists of any PE that has advertised an Inclusive Multicast Tag route for the BD without the "IGMP Proxy Support" flag.
- 2) It sends (x,G) traffic to the set of PEs supporting IGMP Proxy and having listeners for that (x,G) group in that BD. This set consists of any PE that has advertised an Inclusive Multicast Tag route for the BD with the "IGMP Proxy Support" flag and that has advertised a SMET route for that (x,G) group in that BD.

If an ingress PE's Selective P-Tunnel for a given BD uses P2MP and all of the PEs in the BD support that tunnel type and IGMP proxy, then for a given (x,G) group in a given BD it sends (x,G) traffic using the Selective P-Tunnel for that (x,G) group in that BD. This tunnel includes those PEs that have advertised a SMET route for that (x,G) group on that BD (for Selective P-tunnel) but it may include other PEs as well (for Aggregate Selective P-tunnel).

## 9 BGP Encoding

This document defines three new BGP EVPN routes to carry IGMP membership reports. The route type is known as:

- + 6 - Selective Multicast Ethernet Tag Route
- + 7 - Multicast Join Synch Route
- + 8 - Multicast Leave Synch Route

The detailed encoding and procedures for this route type are described in subsequent sections.

### 9.1 Selective Multicast Ethernet Tag Route

A Selective Multicast Ethernet Tag route type specific EVPN NLRI consists of the following:

RD (8 octets)
Ethernet Tag ID (4 octets)
Multicast Source Length (1 octet)
Multicast Source Address (variable)
Multicast Group Length (1 octet)
Multicast Group Address (Variable)
Originator Router Length (1 octet)
Originator Router Address (variable)
Flags (1 octet)

For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the one-octet flag field. The Flags fields are defined as follows:

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

The least significant bit, bit 7 indicates support for IGMP version 1.

The second least significant bit, bit 6 indicates support for IGMP version 2.

The third least significant bit, bit 5 indicates support for IGMP version 3.

The forth least significant bit, bit 4 indicates whether the (S,G) information carried within the route-type is of an Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is not set.

This EVPN route type is used to carry tenant IGMP multicast group information. The flag field assists in distributing IGMP membership interest of a given host/VM for a given multicast route. The version bits help associate IGMP version of receivers participating within

the EVPN domain.

The include/exclude bit helps in creating filters for a given multicast route.

If route is used for IPv6 (MLD) then bit 7 indicates support for MLD version 1. The second least significant bit, bit 6 indicates support for MLD version 2. Since there is no MLD version 3, in case of IPv6 route third least significant bit MUST be 0. In case of IPv6 routes, the fourth least significant bit MUST be ignored if bit 6 is not set.

#### 9.1.1 Constructing the Selective Multicast Ethernet Tag route

This section describes the procedures used to construct the Selective Multicast Ethernet Tag (SMET) route.

The Route Distinguisher (RD) SHOULD 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 Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0  
EVI is VLAN-Aware Bundle service without translation - set to the customer VID for that BD  
EVI is VLAN-Aware Bundle service with translation - set to the normalized Ethernet Tag ID - e.g., normalized VID

The Multicast Source Length MUST be set to length of the multicast Source address in bits. If the Multicast Source Address field contains an IPv4 address, then the value of the Multicast Source Length field is 32. If the Multicast Source Address field contains an IPv6 address, then the value of the Multicast Source Length field is 128. In case of a (\*, G) Join, the Multicast Source Length is set to 0.

The Multicast Source Address is the source IP address from the IGMP membership report. In case of a (\*, G), this field is not used.

The Multicast Group Length MUST be set to length of multicast group address in bits. If the Multicast Group Address field contains an IPv4 address, then the value of the Multicast Group Length field is 32. If the Multicast Group Address field contains an IPv6 address, then the value of the Multicast Group Length field is 128.

The Multicast Group Address is the Group address from the IGMP membership report.

The Originator Router Length is the length of the Originator Router Address in bits.

The Originator Router Address is the IP address of router originating the prefix. It should be noted that using the "Originating Router's IP address" field is needed for local-bias procedures and may be needed for building inter-AS multicast underlay tunnels where the BGP next-hop can get overwritten.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had the INCLUDE or EXCLUDE bit set.

IGMP is used to receive group membership information from hosts/VMs by TORs. Upon receiving the hosts/VMs expression of interest of a particular group membership, this information is then forwarded using Ethernet Multicast Source Group Route NLRI. The NLRI also keeps track of receiver's IGMP protocol version and any source filtering for a given group membership. All EVPN SMET routes are announced with per-EVI Route Target extended communities.

#### 9.1.2 Default Selective Multicast Route

If there is multicast router connected behind the EVPN domain, the PE MAY originate a default SMET (\*,\*) to get all multicast traffic in domain.



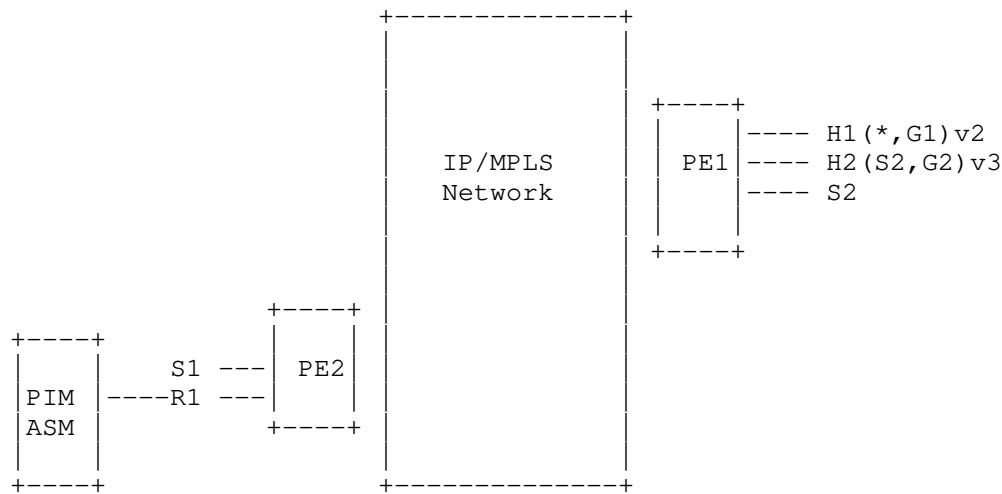


Figure 2: Multicast Router behind EVPN domain

Consider the EVPN network of Figure-2, where there is an EVPN instance configured across the PEs. Lets consider PE2 is connected to multicast router R1 and there is a network running PIM ASM behind R1. If there are receivers behind the PIM ASM network, the PIM Join would be forwarded to the PIM RP (Rendezvous Point). If receivers behind PIM ASM network are interested in a multicast flow originated by multicast source S2 (behind PE1), it is necessary for PE2 to receive multicast traffic. In this case PE2 MUST originate a  $(*,*)$  SMET route to receive all of the multicast traffic in the EVPN domain.

## 9.2 Multicast Join Synch Route

This EVPN route type is used to coordinate IGMP Join  $(x,G)$  state for a given BD between the PEs attached to a given ES operating in All-Active (or Single-Active) redundancy mode and it consists of following:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
Multicast Source Length (1 octet)
Multicast Source Address (variable)
Multicast Group Length (1 octet)
Multicast Group Address (Variable)
Originator Router Length (1 octet)
Originator Router Address (variable)
Flags (1 octet)

For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the one-octet Flags field, whose fields are defined as follows:

0	1	2	3	4	5	6	7
reserved				IE	v3	v2	v1

The least significant bit, bit 7 indicates support for IGMP version 1. The second least significant bit, bit 6 indicates support for IGMP version 2. The third least significant bit, bit 5 indicates support for IGMP version 3. The fourth least significant bit, bit 4 indicates whether the (S, G) information carried within the route-type is of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is not set.

The Flags field assists in distributing IGMP membership interest of a given host/VM for a given multicast route. The version bits help associate IGMP version of receivers participating within the EVPN domain. The include/exclude bit helps in creating filters for a

given multicast route.

If route is being prepared for IPv6 (MLD) then bit 7 indicates support for MLD version 1. The second least significant bit, bit 6 indicates support for MLD version 2. Since there is no MLD version 3, in case of IPv6 route third least significant bit MUST be 0. In case of IPv6 route, the fourth least significant bit MUST be ignored if bit 6 is not set.

#### 9.2.1 Constructing the Multicast Join Synch Route

This section describes the procedures used to construct the IGMP Join Synch route. Support for this route type is optional. If a PE does not support this route, then it MUST NOT indicate that it supports 'IGMP proxy' in the Multicast Flag extended community for the EVIs corresponding to its multi-homed Ethernet Segments (ESs).

An IGMP Join Synch route MUST carry exactly one ES-Import Route Target extended community, the one that corresponds to the ES on which the IGMP Join was received. It MUST also carry exactly one EVI-RT EC, the one that corresponds to the EVI on which the IGMP Join was received. See Section 9.5 for details on how to encode and construct the EVI-RT EC.

The Route Distinguisher (RD) SHOULD 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 (ESI) MUST be set to the 10-octet value defined for the ES.

The Ethernet Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0  
EVI is VLAN-Aware Bundle service without translation - set to the customer VID for the BD  
EVI is VLAN-Aware Bundle service with translation - set to the normalized Ethernet Tag ID - e.g., normalized VID

The Multicast Source length MUST be set to length of Multicast Source address in bits. If the Multicast Source field contains an IPv4 address, then the value of the Multicast Source Length field is 32. If the Multicast Source field contains an IPv6 address, then the value of the Multicast Source Length field is 128. In case of a (\*, G) Join, the Multicast Source Length is set to 0.

The Multicast Source is the Source IP address of the IGMP membership report. In case of a (\*, G) Join, this field does not exist.

The Multicast Group length MUST be set to length of multicast group address in bits. If the Multicast Group field contains an IPv4 address, then the value of the Multicast Group Length field is 32. If the Multicast Group field contains an IPv6 address, then the value of the Multicast Group Length field is 128.

The Multicast Group is the Group address of the IGMP membership report.

The Originator Router Length is the length of the Originator Router address in bits.

The Originator Router Address is the IP address of Router Originating the prefix.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had INCLUDE or EXCLUDE bit set.

### 9.3 Multicast Leave Synch Route

This EVPN route type is used to coordinate IGMP Leave Group (x,G) state for a given BD between the PEs attached to a given ES operating in All-Active (or Single-Active) redundancy mode and it consists of following:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
Multicast Source Length (1 octet)
Multicast Source Address (variable)
Multicast Group Length (1 octet)
Multicast Group Address (Variable)
Originator Router Length (1 octet)
Originator Router Address (variable)
Leave Group Synchronization # (4 octets)
Maximum Response Time (1 octet)
Flags (1 octet)

For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the Maximum Response Time and the one-octet Flags field, whose fields are defined as follows:

0	1	2	3	4	5	6	7
+	-	+	-	+	-	+	-
	reserved				IE		v3
+	-	+	-	+	-	+	-
					v2		v1
+	-	+	-	+	-	+	-

The least significant bit, bit 7 indicates support for IGMP version 1. The second least significant bit, bit 6 indicates support for IGMP version 2. The third least significant bit, bit 5 indicates support for IGMP version 3. The fourth least significant bit, bit 4 indicates whether the (S, G) information carried within the route-type is of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is not set.

The Flags field assists in distributing IGMP membership interest of a given host/VM for a given multicast route. The version bits help associate IGMP version of receivers participating within the EVPN domain. The include/exclude bit helps in creating filters for a given multicast route.

If route is being prepared for IPv6 (MLD) then bit 7 indicates support for MLD version 1. The second least significant bit, bit 6 indicates support for MLD version 2. Since there is no MLD version 3, in case of IPv6 route third least significant bit MUST be 0. In case of IPv6 route, the fourth least significant bit MUST be ignored if bit 6 is not set.

### 9.3.1 Constructing the Multicast Leave Synch Route

This section describes the procedures used to construct the IGMP Leave Synch route. Support for this route type is optional. If a PE does not support this route, then it MUST not indicate that it supports 'IGMP proxy' in Multicast Flag extended community for the EVIs corresponding to its multi-homed Ethernet Segments.

An IGMP Leave Synch route MUST carry exactly one ES-Import Route Target extended community, the one that corresponds to the ES on which the IGMP Leave was received. It MUST also carry exactly one EVI-RT EC, the one that corresponds to the EVI on which the IGMP Leave was received. See Section 9.5 for details on how to form the EVI-RT EC.

The Route Distinguisher (RD) SHOULD 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 (ESI) MUST be set to the 10-octet value defined for the ES.

The Ethernet Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0  
EVI is VLAN-Aware Bundle service without translation - set to the customer VID for the BD  
EVI is VLAN-Aware Bundle service with translation - set to the normalized Ethernet Tag ID - e.g., normalized VID

The Multicast Source length MUST be set to length of multicast source address in bits. If the Multicast Source field contains an IPv4

address, then the value of the Multicast Source Length field is 32. If the Multicast Source field contains an IPv6 address, then the value of the Multicast Source Length field is 128. In case of a (\*, G) Join, the Multicast Source Length is set to 0.

The Multicast Source is the Source IP address of the IGMP membership report. In case of a (\*, G) Join, this field does not exist.

The Multicast Group length MUST be set to length of multicast group address in bits. If the Multicast Group field contains an IPv4 address, then the value of the Multicast Group Length field is 32. If the Multicast Group field contains an IPv6 address, then the value of the Multicast Group Length field is 128.

The Multicast Group is the Group address of the IGMP membership report.

The Originator Router Length is the length of the Originator Router address in bits.

The Originator Router Address is the IP address of Router Originating the prefix.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had INCLUDE or EXCLUDE bit set.

#### 9.4 Multicast Flags Extended Community

The 'Multicast Flags' extended community is a new EVPN extended community. EVPN extended communities are transitive extended communities with a Type field value of 6. IANA will assign a Sub-Type from the 'EVPN Extended Community Sub-Types' registry.

A PE that supports IGMP proxy on a given BD MUST attach this extended community to the Inclusive Multicast Ethernet Tag (IMET) route it advertises for that BD and it MUST set the IGMP Proxy Support flag to 1. Note that an [RFC7432] compliant PE will not advertise this extended community so its absence indicates that the advertising PE does not support IGMP Proxy.

The advertisement of this extended community enables more efficient multicast tunnel setup from the source PE specially for ingress replication - i.e., if an egress PE supports IGMP proxy but doesn't have any interest in a given (x,G), it advertises its IGMP proxy capability using this extended community but it does not advertise





In EVPN, every EVI is associated with one or more Route Targets (RTs). These Route Targets serve two functions:

- Distribution control: RTs control the distribution of the routes. If a route carries the RT associated with a particular EVI, it will be distributed to all the PEs on which that EVI exists.
- EVI identification: Once a route has been received by a particular PE, the RT is used to identify the EVI to which it applies.

An IGMP Join Synch or IGMP Leave Synch route is associated with a particular combination of ES and EVI. These routes need to be distributed only to PEs that are attached to the associated ES. Therefore these routes carry the ES-Import RT for that ES.

Since an IGMP Join Synch or IGMP Leave Synch route does not need to be distributed to all the PEs on which the associated EVI exists, these routes cannot carry the RT associated with that EVI. Therefore, when such a route arrives at a particular PE, the route's RTs cannot be used to identify the EVI to which the route applies. Some other means of associating the route with an EVI must be used.

This document specifies four new Extended Communities (EC) that can be used to identify the EVI with which a route is associated, but which do not have any effect on the distribution of the route. These new ECs are known as the "Type 0 EVI-RT EC", the "Type 1 EVI-RT EC", the "Type 2 EVI-RT EC", and the "Type 3 EVI-RT EC".

A Type 0 EVI-RT EC is an EVPN EC (type 6) of sub-type 0xA.

A Type 1 EVI-RT EC is an EVPN EC (type 6) of sub-type 0xB.

A Type 2 EVI-RT EC is an EVPN EC (type 6) of sub-type 0xC.

A Type 3 EVI-RT EC is an EVPN EC (type 6) of sub-type TBD.

Each IGMP Join Synch or IGMP Leave Synch route MUST carry exactly one EVI-RT EC. The EVI-RT EC carried by a particular route is constructed as follows. Each such route is the result of having received an IGMP Join or an IGMP Leave message from a particular

BD. The route is said to be associated associated with that BD. For each BD, there is a corresponding RT that is used to ensure that routes "about" that BD are distributed to all PEs attached to that BD. So suppose a given IGMP Join Synch or Leave Synch route is associated with a given BD, say BD1, and suppose that the corresponding RT for BD1 is RT1. Then:

0. If RT1 is a Transitive Two-Octet AS-specific EC, then the EVI-RT EC carried by the route is a Type 0 EVI-RT EC. The value field of the Type 0 EVI-RT EC is identical to the value field of RT1.

1. If RT1 is a Transitive IPv4-Address-specific EC, then the EVI-RT EC carried by the route is a Type 1 EVI-RT EC. The value field of the Type 1 EVI-RT EC is identical to the value field of RT1.

2. If RT1 is a Transitive Four-Octet-specific EC, then the EVI-RT EC carried by the route is a Type 2 EVI-RT EC. The value field of the Type 2 EVI-RT EC is identical to the value field of RT1.

3. If RT1 is a Transitive IPv6-Address-specific EC, then the EVI-RT EC carried by the route is a Type 3 EVI-RT EC. The value field of the Type 3 EVI-RT EC is identical to the value field of RT1.

An IGMP Join Synch or Leave Synch route MUST carry exactly one EVI-RT EC.

Suppose a PE receives a particular IGMP Join Synch or IGMP Leave Synch route, say R1, and suppose that R1 carries an ES-Import RT that is one of the PE's Import RTs. If R1 has no EVI-RT EC, or has more than one EVI-RT EC, the PE MUST apply the "treat-as-withdraw" procedure of [RFC7606].

Note that an EVI-RT EC is not a Route Target Extended Community, is not visible to the RT Constrain mechanism [RFC4684], and is not intended to influence the propagation of routes by BGP.

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
Type=0x06										Sub-Type=n										RT associated with EVI											
										RT associated with the EVI (cont.)																					

Where the value of 'n' is 0x0A, 0x0B, 0x0C, or 0x0D corresponding to EVI-RT type 0, 1, 2, or 3 respectively.

#### 9.6 Rewriting of RT ECs and EVI-RT ECs by ASBRs

There are certain situations in which an ES is attached to a set of PEs that are not all in the same AS, or not all operated by the same provider. In some such situations, the RT that corresponds to a particular EVI may be different in each AS. If a route is propagated from AS1 to AS2, an ASBR at the AS1/AS2 border may be provisioned with a policy that removes the RTs that are meaningful in AS1 and replaces them with the corresponding (i.e., RTs corresponding to the same EVIs) RTs that are meaningful in AS2. This is known as RT-rewriting.

Note that if a given route's RTs are rewritten, and the route carries an EVI-RT EC, the EVI-RT EC needs to be rewritten as well.

#### 10 IGMP/MLD Immediate Leave

IGMP MAY be configured with immediate leave option. This allows the device to remove the group entry from the multicast routing table immediately upon receiving a IGMP leave message for (x,G). In case of all active multi-homing while synchronizing the IGMP Leave state to redundancy peers, Maximum Response Time MAY be filled in as Zero. Implementations SHOULD have identical configuration across multi-homed peers. In case IGMP Leave Synch route is received with Maximum Response Time Zero, irrespective of local IGMP configuration it MAY be processed as an immediate leave.

#### 11 IGMP Version 1 Membership Request

This document does not provide any detail about IGMPv1 processing. Multicast working group are in process of deprecating uses of IGMPv1 so it is RECOMMENDED that implementations only use IGMPv2 and above for IPv4 and MLDv1 and above for IPv6.

## 12 Security Considerations

Same security considerations as [RFC7432], [RFC2236], [RFC3376], [RFC2710], [RFC3810].

## 13 IANA Considerations

IANA has allocated the following codepoints from the EVPN Extended Community sub-types registry.

0x09	Multicast Flags Extended Community	[this document]
0x0A	EVI-RT Type 0	[this document]
0x0B	EVI-RT Type 1	[this document]
0x0C	EVI-RT Type 2	[this document]

IANA is requested to allocate a new codepoint from the EVPN Extended Community sub-types registry for the following.

0x0D	EVI-RT Type 3	[this document]
------	---------------	-----------------

IANA has allocated the following EVPN route types from the EVPN Route Type registry.

- 6 - Selective Multicast Ethernet Tag Route
- 7 - Multicast Join Synch Route
- 8 - Multicast Leave Synch Route

IANA is requested to create a registry, "Multicast Flags Extended Community Flags", in the BGP registry.

The Multicast Flags Extended Community contains a 16-bit Flags field. The bits are numbered 0-15, from high-order to low-order.

The registry should be initialized as follows:

Bit	Name	Reference
----	-----	-----
0 - 13	Unassigned	
14	MLD Proxy Support	This document
15	IGMP Proxy Support	This document

The registration policy should be "First Come First Served".

## 14 References

### 14.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4360] S. Sangli et al, "'BGP Extended Communities Attribute", February, 2006.
- [RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", February, 2015.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, October 1999.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015.
- [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) "

## 14.2 Informative References

- [RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for IGMP and MLD snooping PEs", 2006.

## 15 Acknowledgement

The authors would like to thank Stephane Litkowski, Jorge Rabadan, Anoop Ghanwani, Jeffrey Haas for reviewing and providing valuable comment.

## 16 Contributors

Mankamana Mishra  
Cisco systems  
Email: [mankamis@cisco.com](mailto:mankamis@cisco.com)

Derek Yeung  
Arrcus  
Email: [derek@arrcus.com](mailto:derek@arrcus.com)

## Authors' Addresses

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

Samir Thoria  
Cisco  
Email: [sthoria@cisco.com](mailto:sthoria@cisco.com)

Keyur Patel  
Arrcus  
Email: [keyur@arrcus.com](mailto:keyur@arrcus.com)

John Drake  
Juniper  
Email: [jdrake@juniper.net](mailto:jdrake@juniper.net)

Wen Lin  
Juniper  
Email: [wlin@juniper.net](mailto:wlin@juniper.net)

BESS WorkGroup  
Internet-Draft  
Intended status: Standards Track  
Expires: 22 October 2022

N. Malhotra, Ed.  
A. Sajassi  
A. Pattekar  
Cisco Systems  
J. Rabadan  
Nokia  
A. Lingala  
ATT  
J. Drake  
Juniper Networks  
20 April 2022

Extended Mobility Procedures for EVPN-IRB  
draft-ietf-bess-evpn-irb-extended-mobility-08

Abstract

Procedure to handle host mobility in a layer 2 Network with EVPN control plane is defined as part of RFC 7432. EVPN has since evolved to find wider applicability across various IRB use cases that include distributing both MAC and IP reachability via a common EVPN control plane. MAC Mobility procedures defined in RFC 7432 are extensible to IRB use cases if a fixed 1:1 mapping between VM IP and MAC is assumed across VM moves. Generic mobility support for IP and MAC that allows these bindings to change across moves is required to support a broader set of EVPN IRB use cases, and requires further consideration. EVPN all-active multi-homing further introduces scenarios that require additional consideration from mobility perspective. This document enumerates a set of design considerations applicable to mobility across these EVPN IRB use cases and defines generic sequence number assignment procedures to address these IRB use cases.

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 <https://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 22 October 2022.

## Copyright Notice

Copyright (c) 2022 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 (<https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Requirements Language and Terminology . . . . .	3
2. Introduction . . . . .	4
2.1. Document Structure . . . . .	6
3. Optional MAC only RT-2 . . . . .	6
4. Mobility Use Cases . . . . .	6
4.1. Host MAC+IP Move . . . . .	7
4.2. Host IP Move to new MAC . . . . .	7
4.2.1. VM Reload . . . . .	7
4.2.2. MAC Sharing . . . . .	7
4.2.3. Problem . . . . .	7
4.3. Host MAC move to new IP . . . . .	8
4.3.1. Problem . . . . .	9
5. EVPN All Active multi-homed ES . . . . .	10
6. Design Considerations . . . . .	11
7. Solution Components . . . . .	12
7.1. Sequence Number Inheritance . . . . .	12
7.2. MAC Sharing . . . . .	13
7.3. Multi-homing Mobility Synchronization . . . . .	14
8. Requirements for Sequence Number Assignment . . . . .	15
8.1. LOCAL MAC-IP learning . . . . .	15
8.2. LOCAL MAC learning . . . . .	15
8.3. Remote MAC OR MAC-IP Update . . . . .	16
8.4. REMOTE (SYNC) MAC update . . . . .	16
8.5. REMOTE (SYNC) MAC-IP update . . . . .	16
8.6. Inter-op . . . . .	17
8.7. MAC Sharing Race Condition . . . . .	17
8.8. Mobility Convergence . . . . .	18
8.8.1. Generalized Probing Logic . . . . .	18
9. Routed Overlay . . . . .	19
10. Duplicate Host Detection . . . . .	20



10.1.	Scenario A . . . . .	20
10.2.	Scenario B . . . . .	20
10.2.1.	Duplicate IP Detection Procedure for Scenario B . .	21
10.3.	Scenario C . . . . .	21
10.4.	Duplicate Host Recovery . . . . .	22
10.4.1.	Route Un-freezing Configuration . . . . .	22
10.4.2.	Route Clearing Configuration . . . . .	23
11.	Security Considerations . . . . .	23
12.	IANA Considerations . . . . .	23
13.	Acknowledgements . . . . .	23
14.	Normative References . . . . .	23
	Authors' Addresses . . . . .	24

## 1. Requirements Language 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 [RFC2119].

- \* EVPN-IRB: A BGP-EVPN distributed control plane based integrated routing and bridging fabric overlay discussed in [EVPN-IRB]
- \* Underlay: IP or MPLS fabric core network that provides IP or MPLS routed reachability between EVPN PEs.
- \* Overlay: VPN or service layer network consisting of EVPN PEs OR VPN provider-edge (PE) switch-router devices that runs on top of an underlay routed core.
- \* EVPN PE: A PE switch-router in a data-center fabric that runs overlay BGP-EVPN control plane and connects to overlay CE host devices. An EVPN PE may also be the first-hop layer-3 gateway for CE/host devices. This document refers to EVPN PE as a logical function in a data-center fabric. This EVPN PE function may be physically hosted on a top-of-rack switching device (ToR) OR at layer(s) above the ToR in the Clos fabric. An EVPN PE is typically also an IP or MPLS tunnel end-point for overlay VPN flow
- \* Symmetric EVPN-IRB: An overlay fabric first-hop routing architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed at both ingress and egress EVPN PEs.
- \* Asymmetric EVPN-IRB: An overlay fabric first-hop routing architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed and bridged at ingress PE and bridged at egress PEs.

- \* ARP: Address Resolution Protocol [RFC 826]. ARP references in this document are equally applicable to ND as well.
- \* ND: IPv6 Neighbor Discovery Protocol [RFC 4861].
- \* Ethernet-Segment: physical Ethernet or LAG port that connects an access device to an EVPN PE, as defined in [RFC 7432].
- \* ESI: Ethernet Segment Identifier as defined in [RFC 7432].
- \* LAG: Layer-2 link-aggregation, also known as layer-2 bundle port-channel, or bond interface.
- \* EVPN all-active multi-homing: PE-CE all-active multi-homing achieved via a multi-homed layer-2 LAG interface on a CE with member links to multiple PEs and related EVPN procedures on the PEs.
- \* RT-2: EVPN route type 2 carrying both MAC and IP reachability.
- \* RT-5: EVPN route type 5 carrying IP prefix reachability.
- \* MAC-IP: IP association for a MAC, referred to in this document may be IPv4, IPv6 or both.
- \* SYNC MAC route: In the context of EVPN multi-homing, this refers to a local MAC route SYNCed from another PE sharing the same ESI.
- \* SYNC MAC-IP route: In the context of EVPN multi-homing, this refers to a local MAC-IP route SYNCed from another PE sharing the same ESI.
- \* SYNC MAC sequence number: In the context of EVPN multi-homing, this refers to sequence number received with a SYNC MAC route.
- \* SYNC MAC-IP sequence number: In the context of EVPN multi-homing, this refers to sequence number received with a SYNC MAC-IP route.

## 2. Introduction

EVPN-IRB enables capability to advertise both MAC and IP routes via a single MAC+IP RT-2 advertisement. MAC is imported into local bridge MAC table and enables L2 bridged traffic across the network overlay. IP is imported into the local ARP table in an asymmetric IRB design OR imported into the IP routing table in a symmetric IRB design, and enables routed traffic across the layer 2 network overlay. Please refer to [EVPN-IRB] for more background on EVPN IRB forwarding modes.

To support EVPN mobility procedure, a single sequence number mobility attribute is advertised with the combined MAC+IP route. A single sequence number advertised with the combined MAC+IP route to resolve both MAC and IP reachability implicitly assumes a 1:1 fixed mapping between IP and MAC. While a fixed 1:1 mapping between IP and MAC is a common use case that could be addressed via existing MAC mobility procedure, additional IRB scenarios need to be considered, that don't necessarily adhere to this assumption. Following IRB mobility scenarios are considered:

- \* VM move results in VM IP and MAC moving together
- \* VM move results in VM IP moving to a new MAC association
- \* VM move results in VM MAC moving to a new IP association

While existing MAC mobility procedure can be leveraged for MAC+IP move in the first scenario, subsequent scenarios result in a new MAC-IP association. As a result, a single sequence number assigned independently per-[MAC, IP] is not sufficient to determine most recent reachability for both MAC and IP, unless the sequence number assignment algorithm is designed to allow for changing MAC-IP bindings across moves.

Purpose of this draft is to define additional sequence number assignment and handling procedures to adequately address generic mobility support across EVPN-IRB overlay use cases that allow MAC-IP bindings to change across VM moves and can support mobility for both MAC and IP components carried in an EVPN RT-2 for these use cases.

In addition, for hosts on an ESI multi-homed to multiple GW devices, additional procedure is proposed to ensure synchronized sequence number assignments across the multi-homing devices.

Content presented in this draft is independent of data plane encapsulation used in the overlay being MPLS or NVO Tunnels. It is also largely independent of the EVPN IRB solution being based on symmetric OR asymmetric IRB design as defined in [EVPN-INTER-SUBNET].

In addition to symmetric and asymmetric IRB, mobility solution for a routed overlay, where traffic to an end host in the overlay is always IP routed using EVPN RT-5 is also presented in this document.

To summarize, this draft covers mobility mobility for the following independent of the overlay encapsulation being MPLS or an NVO Tunnel:

- \* Symmetric EVPN IRB overlay

- \* Asymmetric EVPN IRB overlay
- \* Routed EVPN overlay

## 2.1. Document Structure

Following sections of the document should be considered informative:

- \* section 4 and 5 provide the necessary background and problem statement being addressed in this document.
- \* section 6 lists the resulting design considerations for the document.

Following sections of the document should be considered normative:

- \* section 8 describes the mobility and sequence number assignment procedures in an EVPN-IRB overlay required to address the scenarios described in section 4.
- \* section 9 describes the mobility procedures for a routed overlay network as opposed to an IRB overlay.
- \* section 10 describes corresponding duplicate detection procedures for EVPN-IRB and routed overlays.

## 3. Optional MAC only RT-2

In an EVPN IRB scenario, where a single MAC+IP RT-2 advertisement carries both IP and MAC routes, a MAC only RT-2 advertisement is redundant for host MACs that are advertised via MAC+IP RT-2. As a result, a MAC only RT-2 is an optional route that may not be advertised from or received at an EVPN PE. This is an important consideration for mobility scenarios discussed in subsequent sections.

MAC only RT-2 may still be advertised for non-IP host MACs that are not advertised via MAC+IP RT-2.

## 4. Mobility Use Cases

This section describes the IRB mobility use cases considered in this document. Procedures to address them are covered later in section 6 and section 7.

- \* Host move results in Host IP and MAC moving together
- \* Host move results in Host IP moving to a new MAC association

- \* Host move results in Host MAC moving to a new IP association

#### 4.1. Host MAC+IP Move

This is the baseline case, wherein a host move results in both host MAC and IP moving together with no change in MAC-IP binding across a move. Existing MAC mobility defined in RFC 7432 may be leveraged to apply to corresponding MAC+IP route to support this mobility scenario.

#### 4.2. Host IP Move to new MAC

This is the case, where a host move results in VM IP moving to a new MAC binding.

##### 4.2.1. VM Reload

A host reload or an orchestrated host move that results in host being re-spawned at a new location may result in host getting a new MAC assignment, while maintaining existing IP address. This results in a host IP move to a new MAC binding:

IP-a, MAC-a ---> IP-a, MAC-b

##### 4.2.2. MAC Sharing

This takes into account scenarios, where multiple hosts, each with a unique IP, may share a common MAC binding, and a host move results in a new MAC binding for the host IP.

As an example, hosts running on a single physical server, each with a unique IP, may share the same physical server MAC. In yet another scenario, an L2 access network may be behind a firewall, such that all hosts IPs on the access network are learnt with a common firewall MAC. In all such "shared MAC" use cases, multiple local MAC-IP ARP entries may be learnt with the same MAC. A host IP move, in such scenarios (for e.g., to a new physical server), could result in new MAC association for the host IP.

##### 4.2.3. Problem

In both of the above scenarios, a combined MAC+IP EVPN RT-2 advertised with a single sequence number attribute implicitly assumes a fixed IP to MAC mapping. A host IP move to a new MAC breaks this assumption and results in a new MAC+IP route. If this new MAC+IP route is independently assigned a new sequence number, the sequence number can no longer be used to determine most recent host IP reachability in a symmetric EVPN-IRB design OR the most recent IP to

MAC binding in an asymmetric EVPN-IRB design.

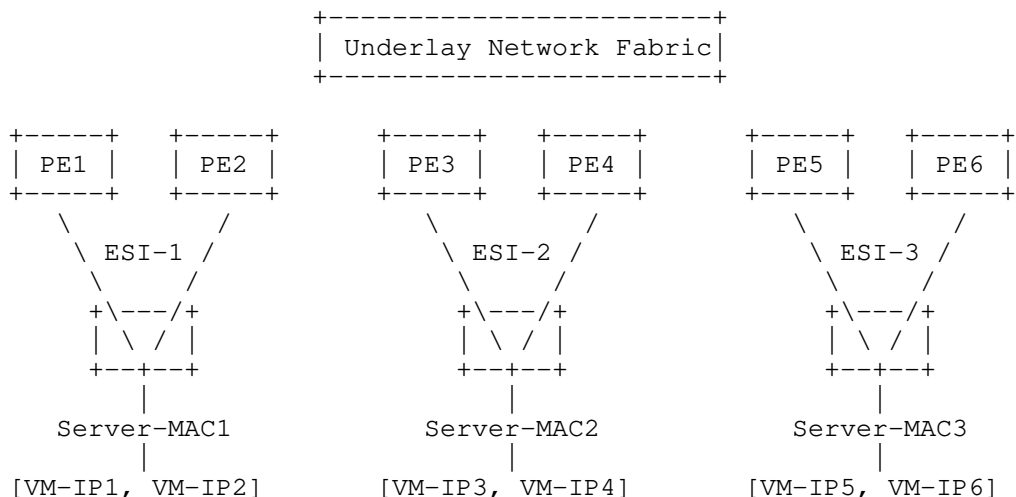


Figure 1

As an example, consider a topology shown in Figure 1, with host VMs sharing the physical server MAC. In steady state, [IP1, MAC1] route is learnt at [PE1, PE2] and advertised to remote PEs with a sequence number N. Now, VM-IP1 is moved to Server-MAC2. ARP or ND based local learning at [PE3, PE4] would now result in a new [IP1, MAC2] route being learnt. If route [IP1, MAC2] is learnt as a new MAC+IP route and assigned a new sequence number of say 0, mobility procedure for VM-IP1 will not trigger across the overlay network.

A sequence number assignment procedure needs to be defined to unambiguously determine the most recent IP reachability, IP to MAC binding, and MAC reachability for such a MAC sharing scenario.

#### 4.3. Host MAC move to new IP

This is a scenario where host move or re-provisioning behind a new gateway location may result in host getting a new IP address assigned, while keeping the same MAC.

## 4.3.1. Problem

Complication with this scenario is that MAC reachability could be carried via a combined MAC+IP route while a MAC only route may not be advertised at all. A single sequence number association with the MAC+IP route again implicitly assumes a fixed mapping between MAC and IP. A MAC move resulting in a new IP association for the host MAC breaks this assumption and results in a new MAC+IP route. If this new MAC+IP route independently assumes a new sequence number, this mobility attribute can no longer be used to determine most recent host MAC reachability.

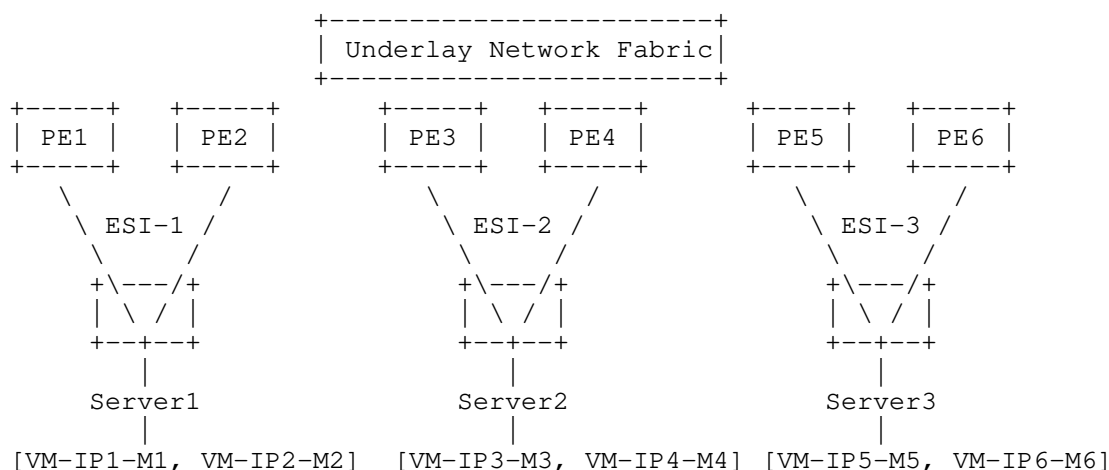


Figure 2

As an example, consider a host VM IP1-M1 that is learnt locally at [PE1, PE2] and advertised to remote hosts with a sequence number N. Consider a scenario where this VM with MAC M1 is re-provisioned at server 2, however, as part of this re-provisioning, assigned a different IP address say IP7. [IP7, M1] is learnt as a new route at [PE3, PE4] and advertised to remote PEs with a sequence number of 0. As a result, L3 reachability to IP7 would be established across the overlay, however, MAC mobility procedure for MAC1 will not trigger as a result of this MAC-IP route advertisement. If an optional MAC only route is also advertised, sequence number associated with the MAC only route would trigger MAC mobility as per [RFC7432]. However, in the absence of an additional MAC only route advertisement, a single sequence number advertised with a combined MAC+IP route may not be sufficient to update MAC reachability across the overlay.

A MAC-IP sequence number assignment procedure needs to be defined to unambiguously determine the most recent MAC reachability in such a scenario without a MAC only route being advertised.

Further, PE1/PE2, on learning new reachability for [IP7, M1] via PE3/PE4 MUST probe and delete any local IPs associated with MAC M1, such as [IP1, M1] in the above example.

Arguably, MAC mobility sequence number defined in [RFC7432], could be interpreted to apply only to the MAC part of MAC-IP route, and would hence cover this scenario. It could hence be interpreted as a clarification to [RFC7432] and one of the considerations for a common sequence number assignment procedure across all MAC-IP mobility scenarios detailed in this document.

#### 5. EVPN All Active multi-homed ES

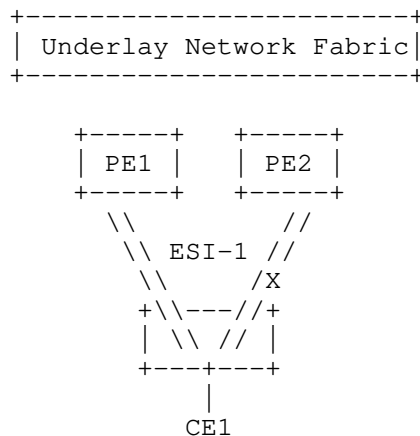


Figure 3

Consider an EVPN-IRB overlay network shown in Figure 2, with hosts multi-homed to two or more PE devices via an all-active multi-homed ES. MAC and ARP entries learnt on a local ES may also be synchronized across the multi-homing PE devices sharing this ES. This MAC and ARP SYNC enables local switching of intra and inter subnet ECMP traffic flows from remote hosts. In other words, local MAC and ARP entries on a given ES may be learnt via local learning and / or via sync from another PE device sharing the same ES.

For a host that is multi-homed to multiple PE devices via an all-active ES interface, local learning of host MAC and MAC-IP at each PE device is an independent asynchronous event, that is dependent on traffic flow and or ARP / ND response from the host hashing to a



directly connected PE on the MC-LAG interface. As a result, sequence number mobility attribute value assigned to a locally learnt MAC or MAC-IP route at each device may not always be the same, depending on transient states on the device at the time of local learning.

As an example, consider a host VM that is deleted from ESI-2 and moved to ESI-1. It is possible for host to be learnt on say, PE1 following deletion of the remote route from [PE3, PE4], while being learnt on PE2 prior to deletion of remote route from [PE3, PE4]. If so, PE1 would process local host route learning as a new route and assign a sequence number of 0, while PE2 would process local host route learning as a remote to local move and assign a sequence number of N+1, N being the existing sequence number assigned at [PE3, PE4].

Inconsistent sequence numbers advertised from multi-homing devices introduces:

- \* Ambiguity with respect to how the remote PEs should handle paths with same ESI and different sequence numbers. A remote PE may not program ECMP paths if it receives routes with different sequence numbers from a set of multi-homing PEs sharing the same ESI.
- \* Breaks consistent route versioning across the network overlay that is needed for EVPN mobility procedures to work.

As an example, in this inconsistent state, PE2 would drop a remote route received for the same host with sequence number N (as its local sequence number is N+1), while PE1 would install it as the best route (as its local sequence number is 0).

There is need for a mechanism to ensure consistency of sequence numbers advertised from a set of multi-homing devices for EVPN mobility to work reliably.

In order to support mobility for multi-homed hosts using the sequence number mobility attribute, local MAC and MAC-IP routes learnt on a multi-homed ES MUST be advertised with the same sequence number by all PE devices that the ES is multi-homed to. There is need for a mechanism to ensure consistency of sequence numbers assigned across these PEs.

## 6. Design Considerations

To summarize, sequence number assignment scheme and implementation must take following considerations into account:

- \* MAC+IP may be learnt on an ES multi-homed to multiple PE devices, hence requires sequence numbers to be synchronized across multi-homing PE devices.
- \* MAC only RT-2 is optional in an IRB scenario and may not necessarily be advertised in addition to MAC+IP RT-2.
- \* Single MAC may be associated with multiple IPs, i.e., multiple host IPs may share a common MAC.
- \* Host IP move could result in host moving to a new MAC, resulting in a new IP to MAC association and a new MAC+IP route.
- \* Host MAC move to a new location could result in host MAC being associated with a different IP address, resulting in a new MAC to IP association and a new MAC+IP route.
- \* LOCAL MAC-IP learn via ARP would always accompanied by a LOCAL MAC learn event resulting from the ARP packet. MAC and MAC-IP learning, however, could happen in any order.
- \* Use cases discussed earlier that do not maintain a constant 1:1 MAC-IP mapping across moves could potentially be addressed by using separate sequence numbers associated with MAC and IP components of MAC+IP route. Maintaining two separate sequence numbers however adds significant overhead with respect to complexity, debugability, and backward compatibility. Hence, this document addresses these requirements via a single sequence number attribute.

## 7. Solution Components

This section goes over main components of the EVPN IRB mobility solution proposed in this draft. Later sections will go over exact sequence number assignment procedures resulting from concepts described in this section.

### 7.1. Sequence Number Inheritance

Main idea presented here is to view a LOCAL MAC-IP route as a child of the corresponding LOCAL MAC only route that inherits the sequence number attribute from the parent LOCAL MAC only route:

Mx-IPx -----> Mx (seq# = N)

As a result, both parent MAC and child MAC-IP routes share one common sequence number associated with the parent MAC route. Doing so ensures that a single sequence number attribute carried in a combined

MAC+IP route represents sequence number for both a MAC only route as well as a MAC+IP route, and hence makes the MAC only route truly optional. As a result, optional MAC only route with its own sequence number is not required to establish most recent reachability for a MAC in the overlay network. Specifically, this enables a MAC to assume a different IP address on a move, and still be able to establish most recent reachability to the MAC across the overlay network via mobility attribute associated with the MAC+IP route advertisement. As an example, when Mx moves to a new location, it would result in LOCAL Mx being assigned a higher sequence number at its new location as per RFC 7432. If this move results in Mx assuming a different IP address, IPz, LOCAL Mx+IPz route would inherit the new sequence number from Mx.

LOCAL MAC and LOCAL MAC-IP routes would typically be sourced from data plane learning and ARP learning respectively, and could get learnt in control plane in any order. Implementation could either replicate inherited sequence number in each MAC-IP entry OR maintain a single attribute in the parent MAC by creating a forward reference LOCAL MAC object for cases where a LOCAL MAC-IP is learnt before the LOCAL MAC.

Arguably, this inheritance may be assumed from RFC 7432, in which case, the above may be interpreted as a clarification with respect to interpretation of a MAC sequence number in a MAC-IP route.

## 7.2. MAC Sharing

Further, for the shared MAC scenario, this would result in multiple LOCAL MAC-IP siblings inheriting sequence number attribute from a common parent MAC route:

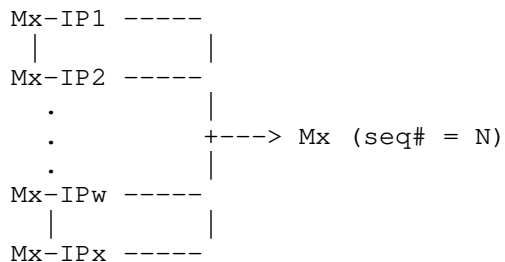


Figure 4

In such a case, a host-IP move to a different physical server would result in IP moving to a new MAC binding. A new MAC-IP route resulting from this move must now be advertised with a sequence number that is higher than the previous MAC-IP route for this IP,

advertised from the prior location. As an example, consider a route Mx-IPx that is currently advertised with sequence number N from PE1. IPx moving to a new physical server behind PE2 results in IPx being associated with MAC Mz. A new local Mz-IPx route resulting from this move at PE2 must now be advertised with a sequence number higher than N. This is so that PE devices, including PE1, PE2, and other remote PE devices that are part of the overlay can clearly determine and program the most recent MAC binding and reachability for the IP. PE1, on receiving this new Mz-IPx route with sequence number say, N+1, for symmetric IRB case, would update IPx reachability via PE2 in forwarding, for asymmetric IRB case, would update IPx's ARP binding to Mz. In addition, PE1 would clear and withdraw the stale Mx-IPx route with the lower sequence number.

This also implies that sequence number associated with local MAC Mz and all local MAC-IP children of Mz at PE2 must now be incremented to N+1, and re-advertised across the overlay. While this re-advertisement of all local MAC-IP children routes affected by the parent MAC route is an overhead, it avoids the need for two separate sequence number attributes to be maintained and advertised for IP and MAC components of MAC+IP RT-2. Implementation would need to be able to lookup MAC-IP routes for a given IP and update sequence number for it's parent MAC and its MAC-IP children.

### 7.3. Multi-homing Mobility Synchronization

In order to support mobility for multi-homed hosts, local MAC and MAC-IP routes learnt on a shared ES MUST be advertised with the same sequence number by all PE devices that the ES is multi-homed to. This also applies to local MAC only routes. LOCAL MAC and MAC-IP may be learnt natively via data plane and ARP/ND respectively as well as via SYNC from another multi-homing PE to achieve local switching. Local and SYNC route learning can happen in any order. Local MAC-IP routes advertised by all multi-homing PE devices sharing the ES must carry the same sequence number, independent of the order in which they are learnt. This implies:

- \* On local or SYNC MAC-IP route learning, sequence number for the local MAC-IP route MUST be compared and updated to the higher value.
- \* On local or SYNC MAC route learning, sequence number for the local MAC route MUST be compared and updated to the higher value.

If an update to local MAC-IP sequence number is required as a result of above comparison with SYNC MAC-IP route, it would essentially amount to a sequence number update on the parent local MAC, resulting in inherited sequence number update on the MAC-IP route.

## 8. Requirements for Sequence Number Assignment

Following sections summarize sequence number assignment procedure needed on local and SYNC MAC and MAC-IP route learning events in order to accomplish the above.

### 8.1. LOCAL MAC-IP learning

A local Mx-IPx learning via ARP or ND should result in computation OR re-computation of parent MAC Mx's sequence number, following which the MAC-IP route Mx-IPx would simply inherit parent MAC's sequence number. Parent MAC Mx Sequence number should be computed as follows:

- \* MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.
- \* MUST be at least equal to corresponding SYNC MAC sequence number if one is present.
- \* If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number.

Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

### 8.2. LOCAL MAC learning

Local MAC Mx Sequence number should be computed as follows:

- \* MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.
- \* MUST be at least equal to corresponding SYNC MAC sequence number if one is present.
- \* Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

Note that the local MAC sequence number might already be present if there was a local MAC-IP learnt prior to the local MAC, in which case the above may not result in any change in local MAC's sequence number.

### 8.3. Remote MAC OR MAC-IP Update

On receiving a remote MAC OR MAC-IP route update associated with a MAC Mx with a sequence number that is

- \* either higher than the sequence number assigned to a LOCAL route for MAC Mx,
- \* or equal to the sequence number assigned to a LOCAL route for MAC Mx, but the remote route is selected as best because of lower VTEP IP as per [RFC 7432],

following handling is required on the receiving PE:

- \* PE MUST trigger probe and deletion procedure for all LOCAL IPs associated with MAC Mx.
- \* PE MUST trigger deletion procedure for LOCAL MAC route for Mx.

### 8.4. REMOTE (SYNC) MAC update

Corresponding local MAC Mx (if present) sequence number should be re-computed as follows:

- \* If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.
- \* If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

### 8.5. REMOTE (SYNC) MAC-IP update

If this is a SYNCed MAC-IP on a local ES, it would also result in a derived SYNC MAC Mx route entry, as MAC only RT-2 advertisement is optional. Corresponding local MAC Mx (if present) sequence number should be re-computed as follows:

- \* If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.
- \* If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

## 8.6. Inter-op

In general, if all PE nodes in the overlay network follow the above sequence number assignment procedure, and the PE is advertising both MAC+IP and MAC routes, sequence number advertised with the MAC and MAC+IP routes with the same MAC would always be the same. However, an inter-op scenario with a different implementation could arise, where a PE implementation non-compliant with this document or with RFC 7432 assigns and advertises independent sequence numbers to MAC and MAC+IP routes. To handle this case, if different sequence numbers are received for remote MAC+IP and corresponding remote MAC routes from a remote PE, sequence number associated with the remote MAC route should be computed as:

- \* Highest of the all received sequence numbers with remote MAC+IP and MAC routes with the same MAC.
- \* MAC sequence number would be re-computed on a MAC or MAC+IP route withdraw as per above.

A MAC and / or IP move to the local PE would now result in the MAC (and hence all MAC-IP) sequence numbers incremented from the above computed remote MAC sequence number.

If MAC only routes are not advertised at all, and different sequence numbers are received with multiple MAC+IP routes for a given MAC, sequence number associated with the derived remote MAC route should still be computed as the highest of the all received MAC+IP sequence numbers with the same MAC.

## 8.7. MAC Sharing Race Condition

In a MAC sharing use case described in section 6.2, a race condition is possible with simultaneous host moves between a pair of PEs. As an example, consider PE1 with local host IPs I1 and I2 sharing MAC M1, and PE2 with local host IPs I3 and I4 sharing MAC M2. A simultaneous move of I1 from PE1 to PE2 and of I3 from PE2 to PE1, such that I3 is learnt on PE1 before I1's local entry has been probed out on PE1 and/or I1 is learnt on PE2 before I3's local entry has been probed out on PE2 may trigger a race condition. This race condition together with MAC sequence number assignment rules defined in section 7.1 can cause new mac-ip routes [I1, M2] and [I3, M1] to bounce a couple of times with an incremented sequence number until stale entries [I1, M1] and [I3, M2] have been probed out from PE1 and PE2 respectively. An implementation MUST ensure proper probing procedures to remove stale ARP, ND, and local MAC entries, following a move, on learning remote routes as defined in section 7.3 (and as per [EVPN-IRB]) to minimize exposure to this race condition.

## 8.8. Mobility Convergence

This sections is to be treated as optional and details ARP and ND probing procedures that MAY be implemented to achieve faster host re-learning and convergence on mobility events.

- \* Following a host move from PE1 to PE2, the host's MAC is discovered at PE2 as a local MAC via a data frames received from the host. If PE2 has a prior REMOTE MAC-IP host route for this MAC from PE1, an ARP/ND probe MAY be triggered at PE2 to learn the MAC-IP as a local adjacency and trigger EVPN RT-2 advertisement for this MAC-IP across the overlay with new reachability via PE2. This results in a reliable "event based" host IP learning triggered by a "MAC learning event" across the overlay, and hence faster convergence of overlay routed flows to the host.
- \* Following a host move from PE1 to PE2, once PE1 receives a MAC or MAC-IP route from PE2 with a higher sequence number, an ARP/ND probe MAY be triggered at PE1 to clear the stale local MAC-IP neighbor adjacency OR re-learn the local MAC-IP in case the host has moved back or is duplicate.
- \* Following a local MAC age-out, if there is a local IP adjacency with this MAC, an ARP/ND probe MAY be triggered for this IP to either re-learn the local MAC and maintain local l3 and l2 reachability to this host OR to clear the ARP/ND entry in case the host is indeed no longer local. Note that this accomplishes clearing of stale ARP entries, triggered by a MAC age-out event even when the ARP refresh timer was longer than the MAC age-out timer. Clearing of stale IP neighbor entries in-turn facilitates traffic convergence in the event that the host was silent and not discovered at its new location. Once stale neighbor entry for the host is cleared, routed traffic flow destined for the host can re-trigger ARP/ND discovery for this host at the new location.

### 8.8.1. Generalized Probing Logic

Above probing logic may be generalized as probing for an IP neighbor anytime a resolving parent MAC route is "inconsistent" with the MAC-IP neighbor route, where being inconsistent is defined as being not present OR conflicting in terms of the route source being local OR remote. MAC-IP to MAC parent relationship described earlier in this document in section 6.1 MAY be used to achieve this logic.



## 9. Routed Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

- \* A host MAC is never advertised in EVPN overlay control plane.
- \* Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI.
- \* An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged.
- \* Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN PE.

Please refer to [RFC 7814] for more details.

Host mobility within the stretched subnet would still need to be supported for this use. In the absence of any host MAC routes, sequence number mobility EXT-COMM specified in [RFC7432], section 7.7 may be associated with a /32 OR /128 host IP prefix advertised via EVPN route type 5. MAC mobility procedures defined in RFC 7432 can now be applied as is to host IP prefixes:

- \* On LOCAL learning of a host IP, on a new ESI, host IP MUST be advertised with a sequence number attribute that is higher than what is currently advertised with the old ESI.
- \* On receiving a host IP route advertisement with a higher sequence number, a PE MUST trigger ARP/ND probe and deletion procedure on any LOCAL route for that IP with a lower sequence number. A PE would essentially move the forwarding entry to point to the remote route with a higher sequence number and send an ARP/ND PROBE for the local IP route. If the IP has indeed moved, PROBE would timeout and the local IP host route would be deleted.

Note that there is still only one sequence number associated with a host route at any time. For earlier use cases where a host MAC is advertised along with the host IP, a sequence number is only associated with a MAC. Only if the MAC is not advertised at all, as in this use case, is a sequence number associated with a host IP.

Note that this mobility procedure would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

## 10. Duplicate Host Detection

Duplicate host detection scenarios across EVPN IRB can be classified as follows:

- \* Scenario A: where two hosts have the same MAC (host IPs may or may not be duplicate).
- \* Scenario B: where two hosts have the same IP but different MACs.
- \* Scenario C: where two hosts have the same IP and host MAC is not advertised at all.

Duplicate detection procedures for scenario B and C would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

### 10.1. Scenario A

For all use cases where duplicate hosts have the same MAC, MAC is detected as duplicate via duplicate MAC detection procedure described in RFC 7432. Corresponding MAC-IP routes with the same MAC do not require duplicate detection and MUST simply inherit the DUPLICATE property from the corresponding MAC route. In other words, if a MAC route is in DUPLICATE state, all corresponding MAC-IP routes MUST also be treated as DUPLICATE. Duplicate detection procedure need only be applied to MAC routes.

### 10.2. Scenario B

Due to misconfiguration, a situation may arise where hosts with different MACs are configured with the same IP. This scenario would not be detected by existing duplicate MAC detection procedure and would result in incorrect forwarding of routed traffic destined to this IP.

Such a situation, on LOCAL MAC-IP learning, would be detected as a move scenario via the following local MAC sequence number computation procedure described earlier in section 6.1:

- \* If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number.

Such a move that results in sequence number increment on local MAC because of a remote MAC-IP route associated with a different MAC MUST be counted as an "IP move" against the "IP" independent of MAC. Duplicate detection procedure described in RFC 7432 can now be applied to an "IP" entity independent of MAC. Once an IP is detected

as DUPLICATE, corresponding MAC-IP route should be treated as DUPLICATE. Associated MAC routes and any other MAC-IP routes associated with this MAC should not be affected.

#### 10.2.1. Duplicate IP Detection Procedure for Scenario B

Duplicate IP detection procedure for such a scenario is specified in [EVPN-PROXY-ARP]. What counts as an "IP move" in this scenario is further clarified as follows:

- \* On learning a LOCAL MAC-IP route Mx-IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different MAC association, say, Mz-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC.
- \* On learning a REMOTE MAC-IP route Mz-IPx, check if there is an existing LOCAL route for IPx with a different MAC association, say, Mx-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC.

A MAC-IP route SHOULD be treated as DUPLICATE if either of the following two conditions are met:

- \* Corresponding MAC route is marked as DUPLICATE via existing duplicate detection procedure.
- \* Corresponding IP is marked as DUPLICATE via extended procedure described above.

#### 10.3. Scenario C

For a purely routed overlay scenario described in section 8, where only a host IP is advertised via EVPN RT-5, together with a sequence number mobility attribute, duplicate MAC detection procedures specified in RFC 7432 can be intuitively applied to IP only host routes for the purpose of duplicate IP detection.

- \* On learning a LOCAL host IP route IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx.
- \* On learning a REMOTE host IP route IPx, check if there is an existing LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx.
- \* With configurable parameters "N" and "M", If "N" IP moves are detected within "M" seconds for IPx, treat IPx as DUPLICATE.

#### 10.4. Duplicate Host Recovery

Once a MAC or IP is marked as DUPLICATE and FROZEN, corrective action must be taken to un-provision one of the duplicate MAC or IP. Un-provisioning a duplicate MAC or IP in this context refers to a corrective action taken on the host side. Once one of the duplicate MAC or IP is un-provisioned, normal operation would not resume until the duplicate MAC or IP ages out, following this correction, unless additional action is taken to speed up recovery.

This section lists possible additional corrective actions that could be taken to achieve faster recovery to normal operation.

##### 10.4.1. Route Un-freezing Configuration

Unfreezing the DUPLICATE OR FROZEN MAC or IP via a CLI can be leveraged to recover from DUPLICATE and FROZEN state following corrective un-provisioning of the duplicate MAC or IP.

Unfreezing the frozen MAC or IP via a CLI at a PE should result in that MAC OR IP being advertised with a sequence number that is higher than the sequence number advertised from the other location of that MAC or IP.

Two possible corrective un-provisioning scenarios exist:

- \* Scenario A: A duplicate MAC or IP may have been un-provisioned at the location where it was NOT marked as DUPLICATE and FROZEN.
- \* Scenario B: A duplicate MAC or IP may have been un-provisioned at the location where it was marked as DUPLICATE and FROZEN.

Unfreezing the DUPLICATE and FROZEN MAC or IP, following the above corrective un-provisioning scenarios would result in recovery to steady state as follows:

- \* Scenario A: If the duplicate MAC or IP was un-provisioned at the location where it was NOT marked as DUPLICATE, unfreezing the route at the FROZEN location will result in the route being advertised with a higher sequence number. This would in-turn result in automatic clearing of local route at the PE location, where the host was un-provisioned via ARP/ND PROBE and DELETE procedure specified earlier in section 8 and in [RFC 7432].
- \* Scenario B: If the duplicate host is un-provisioned at the location where it was marked as DUPLICATE, unfreezing the route will trigger an advertisement with a higher sequence number to the other location. This would in-turn trigger re-learning of local

route at the remote location, resulting in another advertisement with a higher sequence number from the remote location. Route at the local location would now be cleared on receiving this remote route advertisement, following the ARP/ND PROBE.

Note that the probes referred to in the above scenarios are event driven probes resulting from receiving a route with a higher sequence number. Periodic probes resulting from refresh timers may also occur in addition as completely independent probes.

#### 10.4.2. Route Clearing Configuration

In addition to the above, route clearing CLIs may also be leveraged to clear the local MAC or IP route, to be executed AFTER the duplicate host is un-provisioned:

- \* clear mac CLI: A clear MAC CLI can be leveraged to clear a DUPLICATE MAC route, to recover from a duplicate MAC scenario.
- \* clear ARP/ND: A clear ARP/ND CLI may be leveraged to clear a DUPLICATE IP route to recover from a duplicate IP scenario.

Note that the route unfreeze CLI may still need to be run if the route was un-provisioned and cleared from the NON-DUPLICATE / NON-FROZEN location. Given that unfreezing of the route via the unfreeze CLI would any ways result in auto-clearing of the route from the "un-provisioned" location, as explained in the prior section, need for a route clearing CLI for recovery from DUPLICATE / FROZEN state is truly optional.

#### 11. Security Considerations

This document raises no new security issues for EVPN.

#### 12. IANA Considerations

None.

#### 13. Acknowledgements

Authors would like to thank Vibov Bhan and Patrice Brisset for feedback the process of design and implementation of procedures defined in this document. Authors would like to thank Wen Lin for a detailed review and valuable comments related to MAC sharing race conditions. Authors would also like to thank Saumya Dikshit for a detailed review and valuable comments across the document.

#### 14. Normative References

- [EVPN-IRB] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-inter-subnet-forwarding-13, 10 February 2021, <<http://www.ietf.org/internet-drafts/draft-ietf-bess-evpn-inter-subnet-forwarding-13.txt>>.
- [EVPN-PROXY-ARP] Rabadan, J., Sathappan, S., Nagaraj, K., Hankins, G., and T. King, "Operational Aspects of Proxy-ARP/ND in EVPN Networks", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-proxy-arp-nd-11, 7 January 2021, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-proxy-arp-nd-11.txt>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., and B. Fee, "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension Solution", RFC 7814, DOI 10.17487/RFC7814, March 2016, <<https://tools.ietf.org/html/rfc7814>>.

#### Authors' Addresses

Neeraj Malhotra (editor)  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
United States of America  
Email: [nmalhotr@cisco.com](mailto:nmalhotr@cisco.com)

Ali Sajassi  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
United States of America  
Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

Aparna Pattekar  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
United States of America  
Email: apjoshi@cisco.com

Jorge Rabadan  
Nokia  
777 E. Middlefield Road  
Mountain View, CA 94043  
United States of America  
Email: jorge.rabadan@nokia.com

Avinash Lingala  
ATT  
200 S. Laurel Avenue  
Middletown, CA 07748  
United States of America  
Email: ar977m@att.com

John Drake  
Juniper Networks  
Email: jdrake@juniper.net

BESS WorkGroup  
Internet-Draft  
Intended status: Standards Track  
Expires: 21 May 2022

N. Malhotra, Ed.  
A. Sajassi  
Cisco Systems  
J. Rabadan  
Nokia  
J. Drake  
Juniper  
A. Lingala  
ATT  
S. Thoria  
Cisco Systems  
17 November 2021

Weighted Multi-Path Procedures for EVPN Multi-Homing  
draft-ietf-bess-evpn-unequal-lb-15

Abstract

EVPN enables all-active multi-homing for a CE device connected to two or more PEs via a LAG, such that bridged and routed traffic from remote PEs to hosts attached to the Ethernet Segment can be equally load balanced (it uses Equal Cost Multi Path) across the multi-homing PEs. EVPN also enables multi-homing for IP subnets advertised in IP Prefix routes, so that routed traffic from remote PEs to those IP subnets can be load balanced. This document defines extensions to EVPN procedures to optimally handle unequal access bandwidth distribution across a set of multi-homing PEs in order to:

- \* provide greater flexibility, with respect to adding or removing individual multi-homed PE-CE links.
- \* handle multi-homed PE-CE link failures that can result in unequal PE-CE access bandwidth across a set of multi-homing PEs.

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 <https://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 21 May 2022.

## Copyright Notice

Copyright (c) 2021 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 (<https://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. Requirements Language and Terminology . . . . .	3
2. Introduction . . . . .	4
2.1. PE-CE Link Provisioning . . . . .	4
2.2. PE-CE Link Failures . . . . .	6
2.3. Design Requirement . . . . .	7
3. Solution Overview . . . . .	8
4. EVPN Link Bandwidth Extended Community . . . . .	8
4.1. Encoding and Usage of EVPN Link Bandwidth Extended Community . . . . .	9
4.2. Note on BGP Link Bandwidth Extended Community . . . . .	10
5. Weighted Unicast Traffic Load-balancing to an Ethernet Segment . . . . .	10
5.1. Egress PE Behavior . . . . .	10
5.2. Ingress PE Behavior . . . . .	10
6. Weighted BUM Traffic Load-Sharing across an Ethernet Segment . . . . .	12
6.1. The BW Capability in the DF Election Extended Community . . . . .	12
6.2. BW Capability and Default DF Election algorithm . . . . .	13
6.3. BW Capability and HRW DF Election algorithm (Type 1 and 4) . . . . .	14
6.3.1. BW Increment . . . . .	14
6.3.2. HRW Hash Computations with BW Increment . . . . .	15
6.4. BW Capability and Preference DF Election algorithm . . . . .	16
7. Cost-Benefit Tradeoff on Link Failures . . . . .	17

8. Real-time Available Bandwidth . . . . .	17
9. Weighted Load-balancing to Multi-homed Subnets . . . . .	17
10. Weighted Load-balancing without EVPN aliasing . . . . .	17
11. EVPN-IRB Multi-homing With Non-EVPN routing . . . . .	18
12. Operational Considerations . . . . .	18
13. Security Considerations . . . . .	18
14. IANA Considerations . . . . .	18
15. Acknowledgements . . . . .	19
16. Contributors . . . . .	19
17. References . . . . .	19
17.1. Normative References . . . . .	19
17.2. Informative References . . . . .	20
Authors' Addresses . . . . .	20

## 1. Requirements Language 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 [RFC2119].

"Local PE" in the context of an Ethernet Segment refers to a provider edge switch OR router that physically hosts the Ethernet Segment.

"Remote PE" in the context of an Ethernet Segment refers to a provider edge switch OR router in an EVPN overlay, whose overlay reachability to the Ethernet Segment is via the Local PE.

- \* BW: BandWidth
- \* LAG: Link Aggregation Group
- \* ES: Ethernet Segment
- \* ESI: Ethernet Segment ID
- \* vES: Virtual Ethernet Segment
- \* EVI: Ethernet virtual Instance, this is a mac-vrf.
- \* Path-List: A forwarding object used to load-balance routed or bridged traffic across multiple forwarding paths.
- \* Access Bandwidth: Bandwidth of PE-CE links in an Ethernet Segment
- \* Egress PE: In the context of an Ethernet Segment or a route, this is the PE that advertises a locally attached Ethernet Segment RT-1, or a locally attached host or prefix route (RT-2, RT-5).

- \* Ingress PE: In the context of an Ethernet Segment or a route, this is the receiving PE that learns remote Ethernet Segment RT-1 and/or host and prefix routes (RT-2, RT-5) from the Egress PE
- \* IMET: Inclusive Multicast Route
- \* DF: Designated Forwarder
- \* BDF: Backup Designated Forwarder
- \* DCI: Data Center Interconnect Router

## 2. Introduction

In an EVPN-IRB based network overlay, with a CE multi-homed via a EVPN all-active multi-homing, bridged and routed traffic from ingress PEs can be equally load balanced (ECMPed) across the multi-homing egress PEs:

- \* ECMP Load-balancing for bridged unicast traffic is enabled via aliasing and mass-withdraw procedures detailed in RFC 7432.
- \* ECMP Load-balancing for routed unicast traffic is enabled via existing L3 ECMP mechanisms.
- \* Load-sharing of bridged BUM traffic on local ports is enabled via EVPN DF election procedure detailed in RFC 7432

All of the above load balancing and DF election procedures implicitly assume equal bandwidth distribution between the CE and the set of egress PEs. Essentially, with this assumption of equal "access" bandwidth distribution across all egress PEs, ALL remote traffic is equally load balanced across the egress PEs. This assumption of equal access bandwidth distribution can be restrictive with respect to adding / removing links in a multi-homed LAG interface and may also be easily broken on individual link failures. A solution to handle unequal access bandwidth distribution across a set of egress PEs is proposed in this document. Primary motivation behind this proposal is to enable greater flexibility with respect to adding / removing member PE-CE links, as needed and to optimally handle PE-CE link failures.

### 2.1. PE-CE Link Provisioning

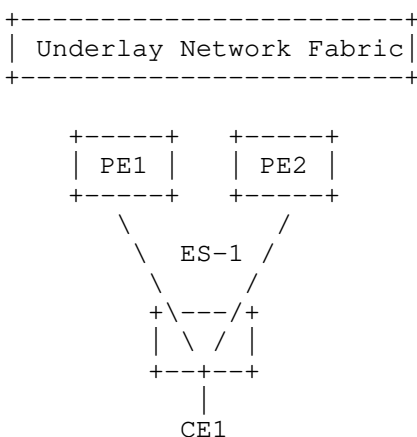


Figure 1

Consider CE1 that is dual-homed to egress PE1 and egress PE2 via EVPN all-active multi-homing with single member links of equal bandwidth to each PE (aka, equal access bandwidth distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it must add a link to both PE1 and PE2 in order to maintain equal access bandwidth distribution and inter-work with EVPN ECMP load balancing. In other words, for a dual-homed CE, total number of CE links must be provisioned in multiples of 2 (2, 4, 6, and so on). For a triple-homed CE, number of CE links must be provisioned in multiples of three (3, 6, 9, and so on). To generalize, for a CE that is multi-homed to "n" PEs, number of PE-CE physical links provisioned must be an integral multiple of "n". This is restrictive in case of dual-homing and very quickly becomes prohibitive in case of multi-homing.

Instead, a provider may wish to increase PE-CE bandwidth OR number of links in any link increments. As an example, for CE1 dual-homed to egress PE1 and egress PE2 in all-active mode, provider may wish to add a third link to only PE1 to increase total bandwidth for this CE by 50%, rather than being required to increase access bandwidth by 100% by adding a link to each of the two PEs. While existing EVPN based all-active load balancing procedures do not necessarily preclude such asymmetric access bandwidth distribution among the PEs providing redundancy, it may result in unexpected traffic loss due to congestion in the access interface towards CE. This traffic loss is due to the fact that PE1 and PE2 will continue to be treated as equal cost paths at remote PEs, and as a result may attract approximately equal amount of CE1 destined traffic, even when PE2 only has half the bandwidth to CE1 as PE1. This may lead to congestion and traffic loss on the PE2-CE1 link. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts must also be load balanced across PE1 and PE2 in 2:1 manner.

## 2.2. PE-CE Link Failures

More importantly, unequal PE-CE bandwidth distribution described above may occur during regular operation following a link failure, even when PE-CE links were provisioned to provide equal bandwidth distribution across multi-homing PEs.

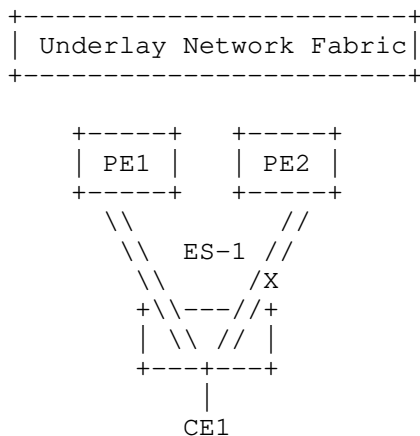


Figure 2

Consider a CE1 that is multi-homed to egress PE1 and egress PE2 via a LAG with two member links to each PE. On a PE2-CE1 physical link failure, LAG represented by an Ethernet Segment ES-1 on PE2 stays up, however, its bandwidth is cut in half. With existing ECMP procedures, both PE1 and PE2 may continue to attract equal amount of

traffic from remote PEs, even when PE1 has double the bandwidth to CE1. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts must also be load balanced across PE1 and PE2 in 2:1 manner to avoid unexpected congestion and traffic loss on PE2-CE1 links within the LAG. As an alternative, min-link on LAGs is sometimes used to bring down the LAG interface on member link failures. This however results in loss of available bandwidth in the network, and is not ideal.

### 2.3. Design Requirement

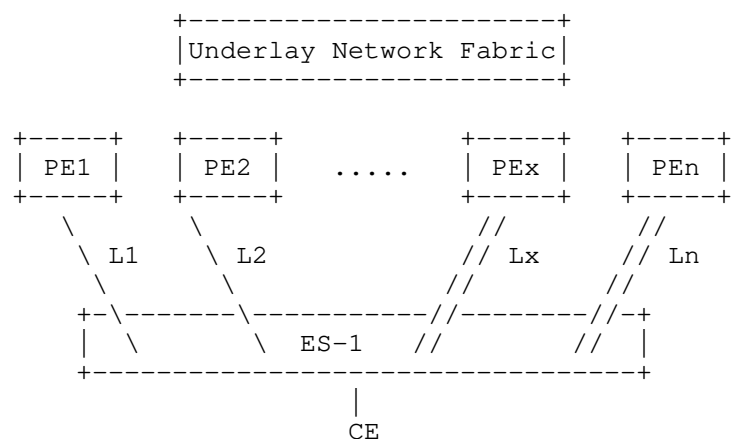


Figure 3

To generalize, if total link bandwidth to a CE is distributed across "n" egress PEs, with  $L_x$  being the total bandwidth to PEx across all links, traffic from ingress PEs to this CE must be load balanced unequally across egress PE set [PE1, PE2, ....., PEn] such that, fraction of total unicast and BUM flows destined for CE that are serviced by egress PEx is:

$$L_x / [L_1 + L_2 + \dots + L_n]$$

Figure 3 illustrates a scenario where egress PE1..PEn are attached to a multi-homed Ethernet Segment, however this document generalizes this requirement so that the unequal load balancing can be applied to PEs attached to a vES or to a multi-homed subnet advertised by EVPN IP Prefix routes.

The solution proposed below includes extensions to EVPN procedures to achieve the above. Following assumption apply to procedure described in this document:

- \* For procedures related to bridged unicast and BUM traffic, EVPN all active multi-homing is assumed.
- \* Procedures related to bridged unicast and BUM traffic are applicable to both aliasing and non-aliasing mode as defined in [RFC7432].

### 3. Solution Overview

In order to achieve weighted load balancing to an ES or vES for overlay unicast traffic, Ethernet A-D per ES route (EVPN Route Type 1) is leveraged to signal the Ethernet Segment weight to ingress PEs. Using Ethernet A-D per ES route to signal the Ethernet Segment weight provides a mechanism that reacts to changes in access bandwidth or number of access links in a service and host independent manner. Ingress PEs computing the MAC path-lists based on global and aliasing Ethernet A-D routes now have the ability to setup weighted load balancing path-lists based on the ES access bandwidth or number of links received from each egress PE that the ES is multi-homed to.

In order to achieve weighted load balancing of overlay BUM traffic, EVPN ES route (Route Type 4) is leveraged to signal the ES weight to egress PEs within an ES's redundancy group to influence per-service DF election. Egress PEs in an ES redundancy group now have the ability to do service carving in proportion to each egress PE's relative ES weight.

Unequal load balancing to multi-homed subnets is achieved by signaling the weight along with the IP Prefix routes advertised for the subnet.

Procedures to accomplish this are described in greater detail next.

### 4. EVPN Link Bandwidth Extended Community

A new EVPN Link Bandwidth extended community is defined for the solution specified in this document:

- \* This extended community is defined of type 0x06 (EVPN).
- \* IANA is requested to assign a sub-type value of 0x10 for the EVPN Link bandwidth extended community, of type 0x06 (EVPN).
- \* EVPN Link Bandwidth extended community is defined as transitive.

#### 4.1. Encoding and Usage of EVPN Link Bandwidth Extended Community

EVPN Link Bandwidth Extended Community value field is used to carry total bandwidth of egress PE's all physical links in an ethernet segment, expressed in Mbits/sec (MegabitsPerSecond) represented as an unsigned integer. Note however that the load balancing algorithm defined in this document uses ratio of Link Bandwidths. Hence, the operator may choose a different unit or use the community as a generalized weight that may be set to link count, locally configured weight, or a value computed based on something other than link bandwidth. In such case, the operator MUST ensure consistent usage of the unit across all egress PEs in an ethernet segment. This may involve multiple routing domains/Autonomous Systems.

In order to facilitate this, as well as avoid interop issues because of provisioning error, one octet in the extended community's six octet 'value' field is used to explicitly signal if the weight encoded in the remaining five octets is link bandwidth expressed in Mbps or a generalized weight value. This results in the following encoding for EVPN link bandwidth extended community:

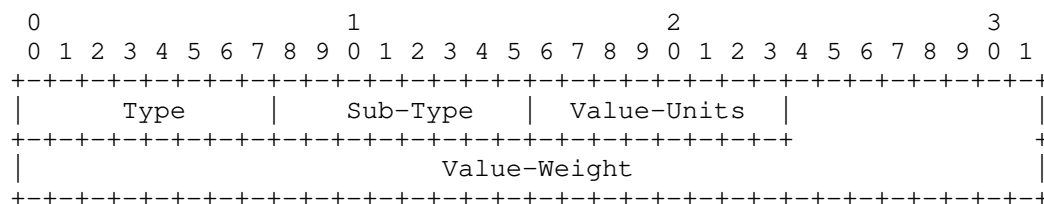


Figure 4

Value-Units is encoded as:

- \* 0x00: weight expressed using default units of Mbps
- \* 0x01: generalized weight expressed in something other than Mbps

Generalized weight units are intentionally left arbitrary to allow for flexibility in its usage for different applications without having to define new encoding for each non-default application. Implementations SHOULD support the default units of Mbps, while support of non-default generalized weight is considered optional.

Additionally, following considerations apply to handling of this extended community at the ingress PE:



- \* An ingress PE MUST check for consistent 'Value-Units' received in the EVPN link bandwidth extended community from each egress PE in an Ethernet Segment. In case of any inconsistency in 'Value-Units' across egress PEs in an Ethernet Segment, this EVPN Link Bandwidth extended community is to be ignored.
- \* An ingress PE MUST ensure that each route contains only a single instance of this extended community sub-type. In case of more than one instance, this EVPN Link Bandwidth extended community is to be ignored.

#### 4.2. Note on BGP Link Bandwidth Extended Community

Link bandwidth extended community described in [BGP-LINK-BW] for layer 3 VPNs was considered for re-use here. This Link bandwidth extended community is however defined in [BGP-LINK-BW] as optional non-transitive. Since it is not possible to change deployed behavior of extended community defined in [BGP-LINK-BW], it was decided to define a new one. In inter-AS scenarios, link-bandwidth needs to be signaled to eBGP neighbors. When signaled across AS boundary, this extended community can be used to achieve optimal load-balancing towards egress PEs in a different AS. This is applicable both when next-hop is changed or unchanged across AS boundaries.

### 5. Weighted Unicast Traffic Load-balancing to an Ethernet Segment

#### 5.1. Egress PE Behavior

A PE that is part of an Ethernet Segment's redundancy group SHOULD advertise an additional "EVPN link bandwidth" extended community with Ethernet A-D per ES route (EVPN Route Type 1), that carries total bandwidth of PE's physical links in an Ethernet Segment or a generalized weight. New EVPN link bandwidth extended community defined in this document is used for this purpose.

EVPN link bandwidth extended community SHOULD NOT be attached to per-EVI RT-1 or to EVPN RT-2.

#### 5.2. Ingress PE Behavior

An ingress PE MUST ensure that the EVPN link bandwidth extended community is received from all the egress PEs in an Ethernet Segment and check for consistent 'Value-Units' received from each egress PE in an Ethernet Segment. In case of missing EVPN Link Bandwidth extended community OR inconsistent 'Value-Units' from any of the egress PEs in an Ethernet Segment, this EVPN Link Bandwidth extended community is to be ignored by the ingress PE and ingress PE is to follow regular ECMP forwarding to that Ethernet Segment.

Once consistency of 'Value-Units' is validated, ingress PE SHOULD use the 'Value-Weight' received from each egress PE to compute a relative (normalized) weight for each egress PE, per ES, and then use this relative weight to compute a weighted path-list to be used for load balancing, as opposed to using an ECMP path-list for load balancing across the egress PE paths. Egress PE Weight and resulting weighted path-list computation at ingress PEs is a local matter. An example computation algorithm is shown below to illustrate the idea:

if,

$L(x,y)$  : link bandwidth advertised by egress PE-x for ES-y

$W(x,y)$  : normalized weight assigned to egress PE-x for ES-y

$H(y)$  : Highest Common Factor (HCF) of  $[L(1,y), L(2,y), \dots, L(n,y)]$

then, the normalized weight assigned to egress PE-x for ES-y may be computed as follows:

$$W(x,y) = L(x,y) / H(y)$$

For a MAC+IP route (EVPN Route Type 2) received with ES-y, ingress PE may compute MAC and IP forwarding path-list weighted by the above normalized weights.

As an example, for a CE multi-homed to PE-1, PE-2, PE-3 via 2, 1, and 1 GE physical links respectively, as part of a LAG represented by ES-10:

$$L(1, 10) = 2000 \text{ Mbps}$$

$$L(2, 10) = 1000 \text{ Mbps}$$

$$L(3, 10) = 1000 \text{ Mbps}$$

$$H(10) = 1000$$

Normalized weights assigned to each egress PE for ES-10 are as follows:

$$W(1, 10) = 2000 / 1000 = 2.$$

$$W(2, 10) = 1000 / 1000 = 1.$$

$$W(3, 10) = 1000 / 1000 = 1.$$

For a remote MAC+IP host route received with ES-10, forwarding load balancing path-list may now be computed as: [PE-1, PE-1, PE-2, PE-3] instead of [PE-1, PE-2, PE-3]. This now results in load balancing of all traffic destined for ES-10 across the three egress PEs in proportion to ES-10 bandwidth at each egress PE.

Weighted path-list computation must only be done for an ES if EVPN link bandwidth extended community is received from all of the egress PE's advertising reachability to that ES via Ethernet A-D per ES Route Type 1. In an unlikely event that EVPN link bandwidth extended community is not received from one or more egress PEs, forwarding path-list should be computed using regular ECMP semantics. Note that a default weight cannot be assumed for an egress PE that does not advertise its link bandwidth as the weight to be used in path-list computation is relative.

If per-ES RT-1 is not advertised or withdrawn from any of the egress PE(s), as per [RFC7432], egress PE is removed from the forwarding path-list for that [EVI, ES]. Hence, the weighted path-list MUST be re-computed.

In an unlikely scenario that per-[ES, EVI] RT-1 is not advertised from any of the egress PE(s), as per [RFC7432], egress PE is not included in the forwarding path-list for that [EVI, ES]. Hence, the weighted path-list for the [EVI, ES] MUST be computed based only on the weights received from egress PEs that advertised the per-[ES, EVI] RT-1.

## 6. Weighted BUM Traffic Load-Sharing across an Ethernet Segment

Optionally, load sharing of per-service DF role, weighted by individual egress PE's link-bandwidth share within a multi-homed ES may also be achieved.

In order to do that, a new DF Election Capability [RFC8584] called "BW" (Bandwidth Weighted DF Election) is defined. BW MAY be used along with some DF Election Types, as described in the following sections.

### 6.1. The BW Capability in the DF Election Extended Community

[RFC8584] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests IANA to allocate a bit in the "DF Election capabilities" registry setup by [RFC8584]:

Bit 4: BW (Bandwidth Weighted DF Election)

ES routes advertised with the BW bit set will indicate the desire of the advertising egress PE to consider the link-bandwidth in the DF Election algorithm defined by the value in the "DF Type".

As per [RFC8584], all the egress PEs in the ES MUST advertise the same Capabilities and DF Type, otherwise the PEs will fall back to Default [RFC7432] DF Election procedure.

The BW Capability MAY be advertised with the following DF Types:

- \* Type 0: Default DF Election algorithm, as in [RFC7432]
- \* Type 1: HRW algorithm, as in [RFC8584]
- \* Type 2: Preference algorithm, as in [EVPN-DF-PREF]
- \* Type 4: HRW per-multicast flow DF Election, as in [EVPN-PER-MCAST-FLOW-DF]

The following sections describe how the DF Election procedures are modified for the above DF Types when the BW Capability is used.

## 6.2. BW Capability and Default DF Election algorithm

When all the PEs in the Ethernet Segment (ES) agree to use the BW Capability with DF Type 0, the Default DF Election procedure as defined in [RFC7432] is modified as follows:

- \* Each PE advertises a "EVPN Link Bandwidth" extended community along with the ES route to signal the PE-CE link bandwidth (LBW) for the ES.
- \* A receiving egress PE MUST use the ES link bandwidth extended community received from each egress PE to compute a relative weight for each egress PE in an Ethernet Segment.
- \* The DF Election procedure MUST now use this weighted list of egress PEs to compute the per-VLAN Designated Forwarder, such that the DF role is distributed in proportion to this normalized weight. As a result, a single PE may have multiple ordinals in the DF candidate PE list and 'N' used in (V mod N) operation as defined in [RFC7432] is modified to be total number of ordinals instead of being total number of egress PEs in an Ethernet Segment.

Considering the same example as in Section 5.2, the candidate PE list for DF election is:

[PE-1, PE-1, PE-2, PE-3].

The DF for a given VLAN-a on ES-10 is now computed as  $(\text{VLAN-a} \% 4)$ . This would result in the DF role being distributed across PE1, PE2, and PE3 in portion to each PE's normalized weight for ES-10.

### 6.3. BW Capability and HRW DF Election algorithm (Type 1 and 4)

[RFC8584] introduces Highest Random Weight (HRW) algorithm (DF Type 1) for DF election in order to solve potential DF election skew depending on Ethernet tag space distribution. [EVPN-PER-MCAST-FLOW-DF] further extends HRW algorithm for per-multicast flow based hash computations (DF Type 4). This section describes extensions to HRW Algorithm for EVPN DF Election specified in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] in order to achieve DF election distribution that is weighted by link bandwidth.

#### 6.3.1. BW Increment

A new variable called "bandwidth increment" is computed for each [PE, ES] advertising the ES link bandwidth extended community as follows:

In the context of an ES,

$L(i)$  = Link bandwidth advertised by PE(i) for this ES

$L(\min)$  = lowest link bandwidth advertised across all PEs for this ES

Bandwidth increment, "b(i)" for a given PE(i) advertising a link bandwidth of  $L(i)$  is defined as an integer value computed as:

$$b(i) = L(i) / L(\min)$$

As an example,

with  $PE(1) = 10$ ,  $PE(2) = 10$ ,  $PE(3) = 20$

bandwidth increment for each PE would be computed as:

$b(1) = 1$ ,  $b(2) = 1$ ,  $b(3) = 2$

with  $PE(1) = 10$ ,  $PE(2) = 10$ ,  $PE(3) = 10$

bandwidth increment for each PE would be computed as:

$b(1) = 1$ ,  $b(2) = 1$ ,  $b(3) = 1$

Note that the bandwidth increment must always be an integer, including, in an unlikely scenario of a PE's link bandwidth not being an exact multiple of  $L(\min)$ . If it computes to a non-integer value (including as a result of link failure), it MUST be rounded down to an integer.

### 6.3.2. HRW Hash Computations with BW Increment

HRW algorithm as described in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] computes a random hash value for each  $PE(i)$ , where,  $(0 < i \leq N)$ ,  $PE(i)$  is the PE at ordinal  $i$ , and  $Address(i)$  is the IP address of  $PE(i)$ .

For ' $N$ ' PEs sharing an Ethernet segment, this results in ' $N$ ' candidate hash computations. The PE that has the highest hash value is selected as the DF.

We refer to this hash value as "affinity" in this document. Hash or affinity computation for each  $PE(i)$  is extended to be computed one per bandwidth increment associated with  $PE(i)$  instead of a single affinity computation per  $PE(i)$ .

$PE(i)$  with  $b(i) = j$ , results in  $j$  affinity computations:

$affinity(i, x)$ , where  $1 < x \leq j$

This essentially results in number of candidate HRW hash computations for each PE that is directly proportional to that PE's relative bandwidth within an ES and hence gives  $PE(i)$  a probability of being DF in proportion to it's relative bandwidth within an ES.

As an example, consider an ES that is multi-homed to two PEs,  $PE1$  and  $PE2$ , with equal bandwidth distribution across  $PE1$  and  $PE2$ . This would result in a total of two candidate hash computations:

$affinity(PE1, 1)$

$affinity(PE2, 1)$

Now, consider a scenario with  $PE1$ 's link bandwidth as  $2x$  that of  $PE2$ . This would result in a total of three candidate hash computations to be used for DF election:

$affinity(PE1, 1)$

$affinity(PE1, 2)$

$affinity(PE2, 1)$

which would give PE1 2/3 probability of getting elected as a DF, in proportion to its relative bandwidth in the ES.

Depending on the chosen HRW hash function, affinity function MUST be extended to include bandwidth increment in the computation.

For e.g.,

affinity function specified in [EVPN-PER-MCAST-FLOW-DF] MAY be extended as follows to incorporate bandwidth increment j:

```
affinity(S,G,V, ESI, Address(i,j)) =  
(1103515245.((1103515245.Address(i).j + 12345) XOR  
D(S,G,V,ESI))+12345) (mod 2^31)
```

affinity or random function specified in [RFC8584] MAY be extended as follows to incorporate bandwidth increment j:

```
affinity(v, Es, Address(i,j)) = (1103515245.((1103515245.Address(i).j  
+ 12345) XOR D(v,Es))+12345) (mod 2^31)
```

#### 6.4. BW Capability and Preference DF Election algorithm

This section applies to ES'es where all the PEs in the ES agree use the BW Capability with DF Type 2. The BW Capability modifies the Preference DF Election procedure [EVPN-DF-PREF], by adding the LBW value as a tie-breaker as follows:

Section 4.1, bullet (f) in [EVPN-DF-PREF] now considers the LBW value:

f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP bit, the LBW value and the lowest IP PE in that order. For instance:

- \* If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=1, LBW=2000] in PE2, PE2 would be elected due to the DP bit.
- \* If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=0, LBW=2000] in PE2, PE2 would be elected due to a higher LBW, even if PE1's IP address is lower.
- \* The LBW exchanged value has no impact on the Non-Revertive option described in [EVPN-DF-PREF].

## 7. Cost-Benefit Tradeoff on Link Failures

While incorporating link bandwidth into the DF election process provides optimal BUM traffic distribution across the ES links, it also implies that DF elections are re-adjusted on link failures or bandwidth changes. If the operator does not wish to have this level of churn in their DF election, then they should not advertise the BW capability. Not advertising BW capability may result in less than optimal BUM traffic distribution while still retaining the ability to allow an ingress PE to do weighted ECMP for its unicast traffic to a set of egress PEs.

## 8. Real-time Available Bandwidth

PE-CE link bandwidth availability may sometimes vary in real-time disproportionately across PE-CE links within a multi-homed ES due to various factors such as flow based hashing combined with fat flows and unbalanced hashing. Reacting to real-time available bandwidth is at this time outside the scope of this document.

## 9. Weighted Load-balancing to Multi-homed Subnets

EVPN Link bandwidth extended community may also be used to achieve unequal load-balancing of prefix routed traffic by including this extended community in EVPN Route Type 5. When included in EVPN RT-5, its value is to be interpreted as egress PE's relative weight for the prefix included in this RT-5. Ingress PE will then compute the forwarding path-list for the prefix route using weighted paths received from each egress PE.

## 10. Weighted Load-balancing without EVPN aliasing

[RFC7432] defines per-[ES, EVI] RT-1 based EVPN aliasing procedure as an optional procedure. In an unlikely scenario where an EVPN implementation does not support EVPN aliasing procedures, MAC forwarding path-list at the ingress PE is computed based on per-ES RT-1 and RT-2 routes received from egress PEs, instead of per-ES RT-1 and per-[ES, EVI] RT-1 from egress PEs. In such a case, only the weights received via per-ES RT-1 from the egress PEs included in the MAC path-list are to be considered for weighted path-list computation.



## 11. EVPN-IRB Multi-homing With Non-EVPN routing

EVPN-LAG based multi-homing on an IRB gateway may also be deployed together with non-EVPN routing, such as global routing or an L3VPN routing control plane. Key property that differentiates this set of use cases from EVPN IRB use cases discussed earlier is that EVPN control plane is used only to enable LAG interface based multi-homing and NOT as an overlay VPN control plane. Applicability of weighted ECMP procedures proposed in this document to these set of use cases is an area of further consideration beyond the scope of this document.

## 12. Operational Considerations

None

## 13. Security Considerations

This document raises no new security issues for EVPN.

## 14. IANA Considerations

[RFC8584] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests IANA to allocate a bit in the "DF Election capabilities" registry setup by [RFC8584]:

Bit 4: BW (Bandwidth Weighted DF Election)

A new EVPN Link Bandwidth extended community is defined to signal local ES link bandwidth to ingress PEs. This extended community is defined of type 0x06 (EVPN). IANA is requested to assign a sub-type value of 0x10 for the EVPN Link bandwidth extended community, of type 0x06 (EVPN). EVPN Link Bandwidth extended community is defined as transitive.

IANA is requested to set up a registry called "Value-Units" for the 1-octet field in the EVPN Link Bandwidth Extended Community. New registrations will be made through the "RFC Required" procedure defined in [RFC8126]. The following initial values in that registry exist:

Value	Name	Reference
----	-----	-----
0	Weight in units of Mbps	This document
1	Generalized Weight	This document
2-255	Unassigned	

## 15. Acknowledgements

Authors would like to thank Satya Mohanty for valuable review and inputs with respect to HRW and weighted HRW algorithm refinements proposed in this document. Authors would also like to thank Bruno Decraene and Sergey Fomin for valuable review and comments.

## 16. Contributors

Satya Ranjan Mohanty  
Cisco Systems  
US  
Email: satyamoh@cisco.com

## 17. References

### 17.1. Normative References

#### [EVPN-DF-PREF]

Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., Mohanty, S., and , "Preference-based EVPN DF Election", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-pref-df-06, 19 June 2020, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-pref-df-06.txt>>.

#### [EVPN-PER-MCAST-FLOW-DF]

Sajassi, A., mishra, m., Thoria, S., Rabadan, J., and J. Drake, "Per multicast flow Designated Forwarder Election for EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-per-mcast-flow-df-election-04, 31 August 2020, <<http://www.ietf.org/internet-drafts/draft-ietf-bess-evpn-per-mcast-flow-df-election-04.txt>>.

#### [EVPN-VIRTUAL-ES]

Sajassi, A., Brissette, P., Schell, R., Drake, J., Rabadan, J., and , "EVPN Virtual Ethernet Segment", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-virtual-eth-segment-06, 9 March 2020, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-virtual-eth-segment-06.txt>>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., and B. Fee, "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension Solution", RFC 7814, DOI 10.17487/RFC7814, March 2016, <<https://tools.ietf.org/html/rfc7814>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, R., Sajassi, N., Drake, A., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

## 17.2. Informative References

- [BGP-LINK-BW]  
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-07, March 2019, <<https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-07.txt>>.

## Authors' Addresses

Neeraj Malhotra (editor)  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
United States of America

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

Ali Sajassi  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
United States of America

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

Jorge Rabadan  
Nokia  
777 E. Middlefield Road

Mountain View, CA 94043  
United States of America

Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

John Drake  
Juniper

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

Avinash Lingala  
ATT  
200 S. Laurel Avenue  
Middletown, CA 07748  
United States of America

Email: [ar977m@att.com](mailto:ar977m@att.com)

Samir Thoria  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
United States of America

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

BESS  
Internet-Draft  
Intended status: Standards Track  
Expires: May 4, 2020

W. Lin  
Juniper Networks, Inc.  
B. Wen  
V. Kozak  
Comcast  
J. Rabadan  
Nokia  
November 1, 2019

EVPN and BGP-based L2VPN Seamless Integration  
draft-lin-bess-evpn-bgp-based-l2vpn-seamless-integ-00

Abstract

This document presents a seamless integration solution for BGP-based Layer-2 VPN (L2VPN) and EVPN to provide point-to-point Virtual Private Wire Service (VPWS). In addition, this document also extends the existing seamless integration for multipoint Ethernet VPN service with all-active multihoming support. The specified solution allows the coexistence of EVPN and L2VPN services under the same point-to-point or multipoint VPN instance. By using this seamless integration solution, a service provider can introduce EVPN into their existing L2VPN network or migrate from an existing L2VPN based network to EVPN.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 <https://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 May 4, 2020.

#### Copyright Notice

Copyright (c) 2019 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	4
3. L2VPN PE, EVPN PE and Composite PE . . . . .	5
4. Requirements . . . . .	6
5. Model of Operation for Seamless Integration of Point-to-point Ethernet VPN . . . . .	6
5.1. Point-to-point Ethernet VPN . . . . .	6
5.2. Operation Model for Seamless Integration . . . . .	7
5.3. Seamless Integration for Single Homing or Multihoming . . . . .	7
5.4. Control Plane Overview . . . . .	8
5.5. Data Plane Overview . . . . .	8
6. Seamless Integration Solution for Point-to-point Ethernet VPN . . . . .	8
6.1. Local ID and Remote ID . . . . .	9
6.2. Composite PE Control Plane Procedure . . . . .	9
6.2.1. Auto-Discovery . . . . .	9
6.2.2. Control Plane Signaling . . . . .	10
6.2.3. Status of an Attachment Circuit . . . . .	11
6.2.4. Layer 2 Extended Community . . . . .	11
6.2.5. Port-active Multihoming and DF election . . . . .	12
6.2.6. Optimization . . . . .	12
6.3. Composite PE Forwarding Procedure . . . . .	13
6.4. Composite PE Procedures for All-Active Multi-Homing . . . . .	15
7. Extended Seamless Integration Solution for Multipoint Ethernet VPN . . . . .	16
7.1. All-Active Multi-Homing and Seamless Integration for BGP-VPLS services . . . . .	16
7.2. Extensions for MAC Flush . . . . .	18
8. IANA Considerations . . . . .	19
9. Security Considerations . . . . .	19

10. Acknowledgements . . . . .	19
11. Contributors . . . . .	19
12. References . . . . .	19
12.1. Normative References . . . . .	19
12.2. Informative References . . . . .	20
Authors' Addresses . . . . .	20

## 1. Introduction

[RFC6624] specifies a point-to-point L2VPN solution by using BGP auto-discovery and signaling. This BGP-based L2VPN service may offer point-to-point service using different types of L2 encapsulation, such as Ethernet, frame relay, etc., and with single home or active-standby redundancy.

EVPN VPWS leverages the latest EVPN technology and brings extra functions to Layer 2 point-to-point Ethernet service, such as all-active redundancy, load balancing and mass withdrawal. All-active redundancy also makes it easier to achieve fast convergence on an access link or node failure.

When expanding an existing L2VPN network with Ethernet encapsulation, service provider may want to deploy EVPN VPWS to provide additional Layer 2 point to point Ethernet services, and at the same time some of the customer traffic may still need to be terminated on the existing L2VPN PEs within the service provider network.

This document describes a seamless-integration solution that allows the co-existence of point-to-point Ethernet services using BGP-based L2VPN procedure per [RFC6624] or EVPN VPWS procedure per [RFC8214] under the same VPN network and over the same MPLS/IP network. Service providers may also use the seamless integration solution for migration a traditional L2VPN network to EVPN VPWS based network.

For the multipoint Ethernet VPN service, [RFC8560] specifies a seamless integration solution for VPLS and EVPN with single home and single-active redundancy support. This document extends the seamless integration solution defined in [RFC8560] with all-active multihoming support for PEs that can support both VPLS per [RFC4761] and EVPN procedures. In the extended solution, VPLS [RFC4761] procedure is used to establish PWs to the rest of VPLS PEs in the same VPN network. Support for using VPLS [RFC4762] procedure to set up PWs to the rest of VPLS PEs is outside the scope of this document.

In this document, section 5 and 6 describe the requirements and operation model for the seamless integration solution for point-to-point Ethernet VPN. Section 6 covers the solution and procedure in more detail.

The extended seamless integration solution for multipoint Ethernet VPN is covered in Section 7.

## 2. Terminology

**AC: Attachment Circuit.** In EVPN VPWS, an attachment circuit for EVPN is also referred to as an Ethernet Segment (ES).

**L2: Layer 2**

**VPWS: Virtual Private Wire Service**

**Point to point: P2P**

**P2P Ethernet Service:** a point-to-point L2 service where the hand-off between a Provide Edge (PE) and a Customer Edge (CE) is based on L2 Ethernet. In this document a P2P Ethernet service is established based on control plane procedure specified in this document or EVPN VPWS [RFC8214] or BGP based L2VPN [RFC6624]. Forwarding is based on using an MPLS label as the demultiplexer.

**L2VPN PE:** a PE supports L2VPN services based on the procedures specified in [RFC6624]

**EVPN VPWS PE:** a PE supports EVPN VPWS based on the procedures specified in [RFC8214]. In this document an EVPN VPWS PE may also be referred to as an EVPN PE for short. An EVPN PE may or may not support seamless integration solution specified in this document.

**BGP VPLS PE:** a PE supports VPLS procedure and multipoint Ethernet VPN service defined in [RFC4761].

**Composite PE:** In the context of a point-to-point Ethernet VPN, a composite PE is a PE that can provide seamless integration solution specified in this document based on both L2VPN procedure per [RFC6624] and EVPN VPWS procedure per [RFC8214] under the same VPN instance. In the context of a multipoint Ethernet VPN, a composite PE is a PE that can provide seamless integration solution based on [RFC8560] as well as the extended procedure specified in this document under section 7.

**L2VPN Route:** a BGP NLRI used for auto-discovery and signaling for L2VPN per [RFC6624]. [RFC6624], in turns, uses BGP VPLS NRLI defined in [RFC4761] for L2VPN. Through out this document, the terms "L2VPN A-D route" and "L2VN route" are used exchangeable.

**BGP-VPLS route:** a BGP NLRI used for auto-discovery and signaling for BGP-based VPLS per [RFC4761].



EVPN E-AD per EVI Route: an EVPN Ethernet A-D per EVI route used for auto-discovery and signaling for EVPN VPWS per [RFC8214].

This document does not distinguish between "all-active" and "active-active" and they are used interchangeably. The same applies to "single-active" and "active-standby".

This document also uses the terms "P2P Ethernet service" and "VPWS" interchangeably. For simplicity, this document may refer to a P2P Ethernet service as a P2P service for short.

This document also makes frequent use of the terminologies specified in [RFC4761], [RFC6624], [RFC7432] and [RFC8214]

### 3. L2VPN PE, EVPN PE and Composite PE

There are three types of PEs defined in this seamless integration solution: L2VPN PE, EVPN PE and composite PE. Under a given Layer 2 Ethernet VPN, the type of PE is categorized by the technology it is provisioned for. For instance, a PE that is provisioned to use L2VPN and EVPN on the same VPN service is considered a composite PE. A L2VPN PE that provides BGP-VPLS service per [RFC4761] is also referred to as BGP-VPLS or VPLS PE for short.

Also in this document in the context of a given Layer 2 Ethernet VPN, an EVPN PE is a PE that is provisioned to provide only the EVPN solution per [RFC8214], or [RFC7432] or both, but not seamless integration solution. It is irrelevant whether an EVPN PE is capable to support seamless integration solution.

For example, for a non-L2VPN PE, a network administrator may know a-priori that the PE does not need to establish any P2P Ethernet service that involves L2VPN PE under a given Layer 2 Ethernet VPN instance. In this case, the PE can be provisioned to act only as an EVPN PE for that VPN even though it is capable of providing seamless integration procedure. If such a prior knowledge is unavailable, then a PE SHALL be provisioned to act as a composite PE if it is capable of. Otherwise, it is unable to establish a P2P Ethernet service with a L2VPN PE.

The term "homogeneous PEs" refers to PEs that are of the same types. Unless explicitly specified in this specification, a PE's type applies to a given Layer 2 Ethernet VPN instance. A PE may act as an EVPN PE for one VPN, but as a composite PE for another VPN.

#### 4. Requirements

The seamless integration solution for point-to-point Ethernet VPN meets the following requirements:

- o It must allow L2VPN, EVPN and composite PEs to participate in the same Layer 2 Ethernet VPN instance.
- o The composite PE, the PE that supports the seamless integration solution, must be backward compatible to support both EVPN VPWS and L2VPN when Ethernet is used as the hand-off between the PE and CE. The composite PE must support the establishment of a layer 2 P2P Ethernet service with a L2VPN PE or an EVPN PE.
- o No change should be required for any exiting L2VPN PEs beyond what are already specified in [RFC6624].
- o The seamless integration solution must support a CE single homed to PEs of different types: L2VPN, EVPN and composite PEs.
- o The seamless solution must support active-standby, also known as single-active, redundancy for L2VPN PEs or EVPN PEs or composite PEs, as long as PEs connecting to the same multihomed CE are of the same type.
- o Composite PEs provisioned for all-active multihoming for their multithemed CE(s) MUST work with L2VPN PE(s) working in single home or active-standby multihoming.
- o The solution SHALL support control word forwarding procedure defined in [RFC4448].
- o The solution SHALL support staged migration to EVPN VPWS network when all L2VPN PEs are upgraded to support EVPN VPWS.

The requirements for the seamless integration solution for multipoint Ethernet VPN are specified in [RFC8560] and they are also reiterated in section 7.

#### 5. Model of Operation for Seamless Integration of Point-to-point Ethernet VPN

##### 5.1. Point-to-point Ethernet VPN

In the seamless integration solution described in this document, PEs participating in a VPN offer point-to-point Layer 2 connections between different customer sites, and Ethernet is used as the Layer 2 hand-off between a PE and a CE. Under the seamless integration

solution, two different techniques can be used to establish P2P Ethernet services under the same VPN: some P2P Ethernet services may use the technique specified per [RFC6624], while others may use the technique specified per [RFC8214]. [RFC6624] uses the terminology of "Layer 2 VPN (L2VPN)". [RFC8214] uses the terminology of "Ethernet VPN (EVPN)". In this document, we refer to a VPN that is capable of offering Layer 2 Ethernet services by using both L2VPN and EVPN VPWS technologies as a point-to-point Ethernet VPN.

## 5.2. Operation Model for Seamless Integration

A PE participating in a point-to-point Ethernet VPN offers P2P Ethernet services with different remote PEs. By nature of point-to-point service, there is no requirement for full mesh among all the PEs participating in the same point-to-point Ethernet VPN instance.

The seamless integration solution allows the coexistence of composite PE, L2VPN PE and EVPN PE under the same VPN instance. It allows the establishment of P2P Ethernet services over the same MPLS/IP core: (a) between two homogenous PEs, or (b) between a composite PE and a L2VPN PE, or (c) between a composite PE and a EVPN PE.

A composite PE can establish a P2P Ethernet service with a L2VPN PE and different a P2P service with an EVPN PE. It is the sole responsibility of a composite PE to seamlessly integrate with L2VPN PEs and EVPN PEs.

There will be no P2P service between an EVPN PE and a L2VPN PE in the same L2 Ethernet VPN as an EVPN PE is provisioned only to provide the procedure/function per EVPN VPWS.

## 5.3. Seamless Integration for Single Homing or Multihoming

L2VPN offers single home as well as active-standby multihoming support, but not active-active multihoming support. Under the seamless integration solution, a composite PE can integrate with L2VPN PE(s) working in:

Case 1: single home

Case 2: active-standby multihoming with its peer L2VPN PE(s)

A composite PE supports seamless integration with EVPN PE(s) working in:

Case 1: single home

Case 2: single-active multihoming with its peer EVPN PE(s)

Case 3: all-active multihoming with its peer EVPN PE(s)

While providing seamless integration solution, a composite PE may provide single home support as well as single-active or all-active multihoming support support to its locally attached CE.

For single-active multihoming, there are two options that a multihomed CE may connect to a redundant set of composite PEs:

1. Through a LAG interface while the composite PEs working in a port-active for single-active multihoming, and the DF or non-DF role on the composite PE is elected on a per port basis.
2. Through a separate interface to each composite PE working in single-active multihoming, and the DF or non-DF role on the composite PE is elected on a per access interface basis.

#### 5.4. Control Plane Overview

In the seamless integration solution, a L2VPN PE continues to use the standard procedure per [RFC6624] without any change or additional new procedure. An EVPN PE also continues to use procedure per [RFC8214] without any change or additional new procedure.

A composite PE follows the seamless integration procedure defined in this document.

A composite PE uses EVPN VPWS procedure per [RFC8214] to establish a P2P Ethernet service with an EVPN PE.

#### 5.5. Data Plane Overview

Regardless of the type of a PE, data traffic continues to be carried over a MPLS/IP tunnel from an ingress PE to an egress PE. At the egress PE, an MPLS label is used as the demultiplexer to identify the attachment circuit for a P2P Ethernet service.

#### 6. Seamless Integration Solution for Point-to-point Ethernet VPN

It is the sole responsibility of a composite PE to provide seamless integration solution with a L2VPN PE. So the focus of the solution is the composite PE. This section and its sub-sections follow specify the solution and procedures a composite PE provides.

## 6.1. Local ID and Remote ID

Similar to other PEs, a composite PE is provisioned for the VPN it participates through Route Target(s) and a Route Distinguisher (RD). For each P2P Ethernet service, the PE involved is provisioned with a pair of local and remote IDs. The local ID identifies an local attachment circuit associated with a P2P service, while the remote ID identifies an attachment circuit attached to a remote PE.

For a given P2P Ethernet service, a local ID for a PE is the remote ID for its corresponding remote PE. It is required that that both PEs involved in a P2P Ethernet service must have a matching pair of local/remote IDs correspondingly. In the BGP signaling procedure for auto-discovery, only local ID is signaled in the control plane, but not remote ID.

In L2VPN, the ID used to identify an attachment circuit associated with a P2P service is referred to as a VE ID or site ID which is a 16-bit integer. A valid VE-ID for L2VPN is in the range of 1 to 0xFFFFE.

In EVPN VPWS, the ID used to identify an attachment associated with a P2P service is referred as an EVPN VPWS service instance identifier which is a 24-bit integer. A valid service instance identifier for EVPN VPWS is in the range of 1 to 0xFFFFF.

A p2p Ethernet service using L2VPN procedure MUST keep its local/remote ID within the range of 0x1 to 0xFFFFE.

## 6.2. Composite PE Control Plane Procedure

This section and the sub-sections under it cover the control plane procedure of a composite PE to interact with other types of PEs.

### 6.2.1. Auto-Discovery

All three types of PEs defined in this document continue to use MP-BGP for auto-discovery. An auto-discovery procedure involves two parts: A PE needs to identify itself for other PEs to discover it, and a PE needs to auto discover other PEs. Auto-discovery is only meaningful to PEs participating in the same VPN.

A composite PE needs to identify itself and discover other PE(s) participating in the same point-to-point Ethernet VPN. If a composite PE does not know a-priori the type of remote PE for a given P2P Ethernet service it tries to establish, a composite PE MUST participate in both L2VPN and EVPN auto-discovery procedures per

[RFC6624] and [RFC8214] except in the cases specified in section 6.2.5.

Similar to a L2VPN or EVPN PE, a composite PE uses Route Target community to identify itself as a part of a point-to-point Ethernet VPN instance. A composite PE announces itself through both BGP L2VPN A-D route and EVPN E-AD per EVI route, and with the RT(s) belong to the VPN it participates. A network operator may choose to use different RT(s) to identify L2VPN PEs and EVPN PEs participating in the same VPN. In this case, A composite PE needs to be provisioned with RTs used by L2VPN PEs and EVPN PEs.

A composite PE discovers other L2VPN PEs by processing L2VPN A-D routes that have route target(s) matching its import RT(s). At the same time, a composite PE discovers other EVPN or composite PEs by processing EVPN E-AD per EVI routes that have the RT(s) matching its import RT(s).

At the end of discovery procedure, a L2VPN PE discovers all L2VPN PEs and all composite PEs participating in the same VPN. However a L2VPN cannot distinguish a L2VPN from a composite PE. From a point of L2VPN PE, all composite PEs are L2VPN PEs.

Also at the end of discovery procedure, an EVPN PE discovers all EVPN PEs and all composite PEs participating in the same VPN. Similarly, an EVPN PE cannot distinguish an EVPN PE from a composite PE. From a point of EVPN PE, all composite PEs are EVPN PEs.

#### 6.2.2. Control Plane Signaling

In the seamless integration solution, a composite PE relies on MP-BGP signaling to exchange information for each of its P2P Ethernet service. A composite PE uses the procedures defined in [RFC6624] and [RFC8214] for control plane signaling, and by default it originates both a L2VPN route and an EVPN E-AD per EVI route for each of its P2P Ethernet service. Note that these are the same routes used for auto-discovery.

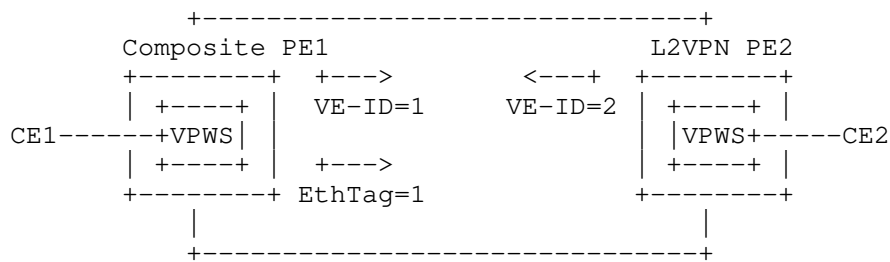


Figure 1: BGP-L2VPN and EVPN-VPWS integration

As it is shown in Figure 1 above, PE1 is provisioned to be a composite PE. PE1 originates both L2VPN A-D route with local VE-ID (1) as well as EVPN E-AD per EVI route with local Ethernet Tag ID (1) in their corresponding NLRIs. PE2 is a L2VPN PE and it originates a L2VPN A-D route with a local VE-ID (2) in its NLRI. A p2p Ethernet service is established between PE1 and PE2 based the L2VPN procedure per [RFC4761] when both PEs have the matching local/remote VE-IDs.

A composite PE may be optimized to originate either L2VPN route or EVPN E-AD per EVI route, but not both based on its provisioning model. Please see section 6.2.6 for detail.

If a CE is multihomed to composite PEs in multihoming mode, each composite PE also originates an EVPN E-AD per ES route and EVPN Ethernet segment per [RFC8214].

### 6.2.3. Status of an Attachment Circuit

A composite PE uses status vector TLV to notify other L2VPN PEs through its L2VPN route the status of its attachment circuit per . A composite PE updates the corresponding L2VPN route with an updated status vector when there is a status change in its attachment circuit.

A composite PE withdraws its corresponding EVPN E-AD route per procedure defined in [RFC8214] when its locally attached Ethernet segment goes down.

#### 6.2.4. Layer 2 Extended Community

A composite PE uses L2VPN info extended community for L2VPN per [RFC6624]. It shall support L2 encapsulation of type 4 and type 5.

A composite PE uses EVPN Layer 2 attribute extended community specified in [RFC8214] for EVPN, and it attaches the Layer 2 extended community in the EVPN A-D route it originates.

#### 6.2.5. Port-active Multihoming and DF election

For the seamless migration, it is desirable that a multihomed CE uses a LAG interface to connect to a redundant set of composite PEs, such that when L2VPN PE involved in a Layer 2 P2P Ethernet service is migrated to support EVPN-VPWS, there is no need to touch the multihomed CE device if at that stage the redundant set of composite PEs are changed to provide all-active multihoming.

In addition, if the LACP protocol is running for the interface and while in single-active scenario, it is recommended a non-DF composite PE sends out-of-sync state for the interface instead of operational down. To that end, each composite PE is required to play a DF or non-DF role on a per port basis instead of per VLAN or per (ES, VLAN) basis.

To support multihomed CE connecting to the composite PEs working in a single-active multihoming scenario through a LAG interface, each composite PE must support port-active load-balancing, similar as it is specified in section 3 of [EVPN-MH-PA] except that a composite PE must also provide L2VPN functionality per [RFC6624].

Please note that per port DF/non-DF role can be achieved by using one of the standard based DF election algorithms, as long as the algorithm can be easily carried out on a per port basis, such as the preference based DF election when both the ESI and preference are configured on a per port basis.

Supporting port based single-active multihoming on the composite PEs with its multihomed CE using LAG interface does not change the control plane signaling, and it is oblivious to L2VPN PE. Since we cannot change the behavior of a L2VPN PE, a composite PE will continue to signal the preference for L2VPN on a per access interface basis through the Layer 2 extended community alongside its corresponding L2VPN A-D route. A L2VPN PE continues to carry the DF election based on its normal L2VPN process.

#### 6.2.6. Optimization

With the simplest provisioning model, if a composite PE does not know a-priori whether the remote PE for a given P2P service is a L2VPN PE or an EVPN PE, the composite needs to participate in the auto-discovery and signaling procedures for both L2VPN and EVPN. This works well as it allows a composite to establish a P2P service with



different types of PEs composite PE, and to switch from using a L2VPN PW to EVPN VPWS dynamically during the migration process.

The simplest provisioning model may not be optimal though, in that a composite PE originates twice as many A-D routes as they are required to establish the number of P2P services it is provisioned to. Therefore in some scenario, it is desirable that a composite PE be optimized to perform either L2VPN or EVPN VPWS procedure for a given P2P service, but not both.

For a composite PE, if a Service Provider has the prior knowledge about the types of remote PEs for some or all of its P2P Ethernet services, reducing the number of routes a composite PE originates can be achieved through the configuration. Based on the configuration, a composite may advertise EVPN route but not L2VPN A-D route for a P2P Ethernet service, or vice versa. It is up to the Service Provider to decide based on the network requirement.

### 6.3. Composite PE Forwarding Procedure

A composite PE supports forwarding procedure defined in [RFC6624] and [RFC8214].

When a packet arrives at an ingress composite PE, the composite PE adds a VPN service label based on the AC that packet arrives at, and it encapsulates the packet and sends it through a tunnel to the egress PE.

- o A composite PE will not forward customer traffic to the L2VPN PE playing a non-DF role
- o If a composite PE detects that two or more EVPN PEs are attached to the same ES and they are working in all-active mode, it will load balance the traffic among the EVPN PEs.
- o If a composite PE detects that two or more EVPN PEs are attached to the same ES and they are working in single-active mode, it will only forward the traffic to the EVPN PE playing a DF role.
- o If a set of composite PEs work in all-active multihoming mode for the same multihomed CE, then regardless of DF or Non-DF role each composite PE plays, it must forward the packet received from its multihomed CE to the remote L2VPN DF PE.
- o If a composite PE receives both L2VPN and EVPN A-D routes from a remote PE for the same p2p Ethernet service, the composite should install forwarding routes in a make-before-break fashion:

- a. For the traffic coming from the remote PE to its local access interface direction, to achieve a fast failover, the composite may install forwarding routes based on both L2VPN and EVPN A-D routes. However, to save system resource in a scaled setup, the composite may choose to install only the forwarding route for the EVPN A-D route and it should do so before it deletes the forwarding route for the L2VPN A-D route if it was installed beforehand.
- b. For traffic coming from its local access interface to the remote PE direction, only one route can be installed for the same local access interface. Forwarding should be based on the EVPN A-D route. The composite PE should update the forwarding route in a make-before-break fashion if the forwarding route for L2VPN A-D route has already been installed before the processing of the incoming EVPN A-D route.
- o If a composite PE receives both L2VPN and EVPN A-D routes from a remote PE for the same p2p Ethernet service, and later on the remote PE is reverted back to a L2VPN only PE and withdraws its EVPN A-D route, the composite PE should also update the forwarding route accordingly in a make-before-break fashion:
  - a. For the traffic coming from the remote PE to its local access interface direction, if the forwarding route for the L2VPN A-D route is not there, the composite PE should install the forwarding route for the L2VPN A-D route before it tears down the forwarding route for the EVPN A-D route.
  - b. For the traffic coming from its local access interface to the remote PE direction, only one route can be installed for the same local access interface. The composite PE should update the forwarding route based on the L2VPN A-D route in a make-before-break fashion.
- o Upon reception of an A-D per EVI route and an L2VPN route for the same P2P service, if both routes match the configured IDs, a composite PE MUST select the EVPN route and forward the traffic only to the EVPN PE, and not the L2VPN PE.

When a packet arrives at an egress PE, the VPN service label carried in the packet is used as the demultiplexer to identify the AC connecting to the destination CE.

#### 6.4. Composite PE Procedures for All-Active Multi-Homing

Two or more Composite PEs MAY be attached to the same All-Active multi-homed CE and still be seamlessly integrated in an L2VPN network. This is illustrated in Figure 2.

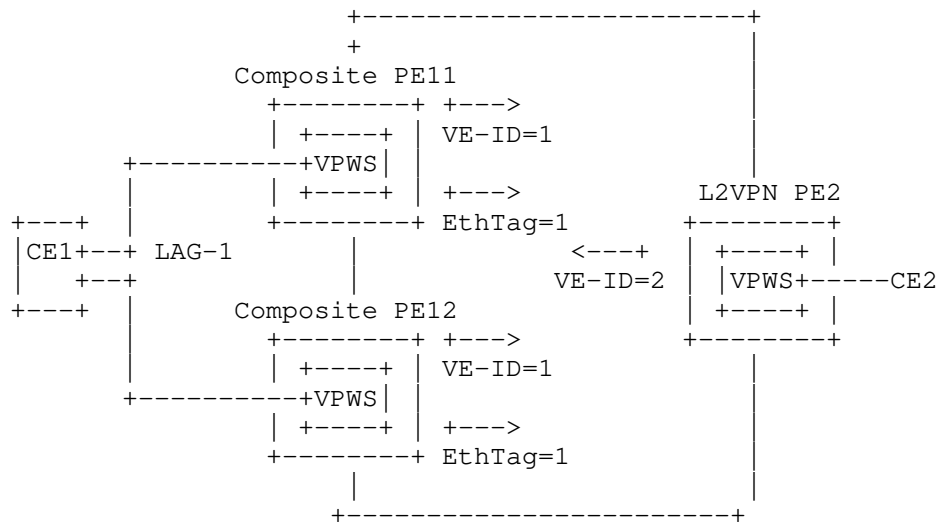


Figure 2: L2VPN and EVPN-VPWS integration with all-active multi-homed CEs

In the example of Figure 2, PE11 and PE12 are configured as composite PEs with the same local CE identifiers. That is, both PEs are configured with local VE-ID (1) and the same remote VE-ID (2). Also, both will be configured with the same EVPN local Ethernet Tag (1) and remote Ethernet Tag (2). As long as PE2 does not become a composite PE or an EVPN PE, it will not import the A-D per-EVI routes and will import the L2VPN routes only. PE2 will make a selection to use either PE11 or PE12 to setup an L2VPN VPWS service.

For example, assuming PE11 is selected, PE2 forwards the traffic coming from CE2 to PE11 (there is no per-flow load-balancing). In case of failure, upon receiving the L2VPN route withdraw from PE11, PE2 will change its forwarding next-hop to PE12.

In the reverse direction, CE1 will perform per-flow load-balancing to PE11 and PE12. Both PEs will program their forwarding paths to send CE1 traffic to PE2.

The benefit of this solution is that when PE2 can be upgraded to an EVPN or composite PE, the P2P service can be migrated to EVPN VPWS with no changes on CE1 or PE11/PE12.

## 7. Extended Seamless Integration Solution for Multipoint Ethernet VPN

[RFC8560] specifies how EVPN and VPLS PEs can be seamlessly integrated into the same network, assuming the VPLS PEs use [RFC4761] or [RFC4762] procedures to setup the pseudowires to the remote VPLS PEs or composite PEs. [RFC8560] procedures consider that CEs can be multi-homed to two VPLS PEs, or to a group of composite PEs in a single-active or port-active Ethernet Segment. All-active multi-homing is out of scope.

This specification updates [RFC8560] in case All-Active multi-homing is used on two or more composite PEs of the same multipoint VPN service and the composite PEs and VPLS PEs use the BGP-VPLS [RFC4761] control and data plane procedures. Seamless integration and All-active multi-homing procedures for [RFC4762] VPLS is out of scope. This document also updates [RFC8560] to clarify the required MAC Flush procedures in case single-active/all-active/port-active multi-homing is used on the composite PEs.

### 7.1. All-Active Multi-Homing and Seamless Integration for BGP-VPLS services

All-active Ethernet Segments MAY be used in a VPLS service composed of composite and BGP-VPLS PEs. Ethernet Segments are an EVPN construct, hence only supported in composite PEs and not BGP-VPLS PEs. Figure 3 illustrates an example of the use of All-active Ethernet Segments and seamless integration between BGP-VPLS and EVPN.

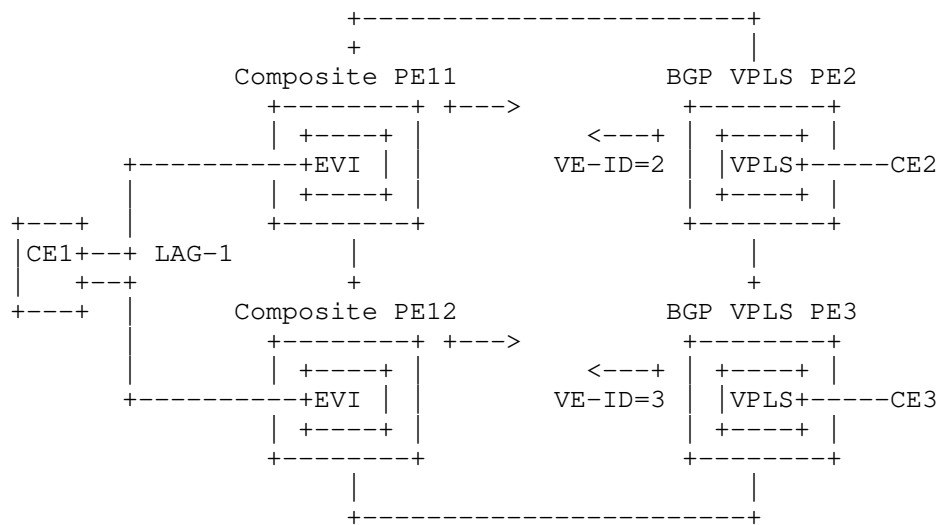


Figure 3: BGP-VPLS and EVPN PEs with all-active multi-homing

In Figure 3, the composite PEs will be provisioned for EVPN and All-active Multi-homing as specified in [RFC7432]. In addition, BGP-VPLS is enabled on the same services. PE11 and PE12 will therefore advertise the corresponding EVPN and BGP-VPLS routes. The EVPN routes are only imported by PE11 and PE12, whereas the BGP-VPLS routes are imported by all the PEs in the service.

In this case, the PEs MUST follow the procedures in [RFC8560] that are repeated below for the reader's benefit:

- o The composite PEs MUST place their EVPN MP2P service tunnels and BGP-VPLS PWs in the same Split Horizon Group. I.e., traffic coming from a BGP-VPLS PW MUST NOT be forwarded to an EVPN tunnel.
- o If two composite PEs successfully attempt to setup a BGP-VPLS PW and an EVPN tunnel, the BGP-VPLS pseudowire will be brought operationally down.
- o The composite PEs will not advertise any MAC/IP routes for MAC address learned on a BGP-VPLS PW that is part of the Split Horizon Group assigned to the EVPN tunnels.

In addition, this document updates [RFC8560] so that All-active multi-homing Ethernet Segments MAY exist in the composite PEs. If an all-active multi-homing ES is defined in a group of composite PEs, all the BDs associated to the LAG MUST support and follow the EVPN multi-homing procedures.

If a group of composite PEs work in all-active multihoming and another group of composite PEs work in single-active, normal EVPN procedure will be used between these two group of composite PEs.

If a group of composite PEs work in all-active multihoming and remote BGP-VPLS PEs work in single-active, BGP-VPLS procedure will be used between composite PEs and BGP-VPLS PEs.

When all-active multi-homing is used, a MAC flip-flopping effect will exist on the BGP-VPLS PEs. In Figure 3, this effect results in CE1's MAC moving between two different PWs in PE2 and PE3. E.g., at first CE1 may hash the traffic to PE11, and PE2 may learn CE1's MAC on its pseudowire PE2-PE11. Later, if CE1 hashes the traffic (for a different flow) to PE12, PE2 will move CE1's MAC to its PW PE2-PE12. This MAC move or "flip-flopping" can happen continuously and may have harmful consequences for the BGP-VPLS PEs. In some cases, the BGP-VPLS PEs will consider this to be a loop.

The solution to avoid the MAC "flip-flopping" is based on the support of "MAC Pinning" on the BGP-VPLS PEs, as follows:

- o In Figure 3, the composite PEs and BGP-VPLS PEs will setup their PWs normally.
- o The MAC flip-flopping effect would be avoided by enabling MAC Pinning on the PE2 and PE3 pseudowires.
- o With MAC Pinning enabled, PE2 and PE3 will learn CE1's MAC on only one PW and will not be relearned in the same or different PW until the MAC ages out. E.g., consider CE1 hashes the first flow to PE11 and PE11 forwards to PE2. PE2 learns CE1's MAC on PW PE2-PE11. Since MAC Pinning is applied on that PW, subsequent frames arriving at PW PE2-PE12 with CE1's MAC will not trigger a relearn process on PE2.

MAC Pinning is assumed to be supported by all the BGP-VPLS PEs in the network, therefore no upgrade is required on the BGP-VPLS PEs to support this specification.

## 7.2. Extensions for MAC Flush

Irrespective of the type of multi-homing used on the composite PEs, in case of a failure on the Ethernet Segment (node or link failure) the composite PEs MUST indicate the need to flush MAC addresses to the remote BGP-VPLS PEs.

E.g., in Figure 3, consider CE1's MAC is learned on PW PE2-PE11 (on PE2). If the link CE1-PE11 fails, PE2 will continue sending the

unicast traffic to CE1 using the PW to PE11, and therefore causing a blackhole until CE1's MAC ages out.

A MAC flush mechanism is required in order to speed up the convergence in case of ES failures. This requires some extensions to [RFC8560] and it will be added in future versions.

## 8. IANA Considerations

This document raises no new IANA request. There is no IANA actions.

## 9. Security Considerations

This document does not introduce any new security concern. This document inherits the same security as they are specified in [RFC6624] and [RFC8214].

## 10. Acknowledgements

The authors would like to thank Hitesh Mali and Vasu Venkatraman for their valuable comments and feedbacks. They would also like to thank John Drake for his review and support.

## 11. Contributors

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

Vinod Prabhu, Nokia

## 12. References

### 12.1. Normative References

- [EVPN-MH-PA]       Brissette, P., Thoria, S., and A. Sajassi, "EVPN multi-homing port-active load-balancing", internet-draft draft-brissette-bess-evpn-mh-pa-04, March 2019.
- [RFC2119]       Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4761]       Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.

- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, DOI 10.17487/RFC6624, May 2012, <<https://www.rfc-editor.org/info/rfc6624>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <<https://www.rfc-editor.org/info/rfc8214>>.
- [RFC8560] Sajassi, A., Ed., Salam, S., Del Regno, N., and J. Rabadan, "Seamless Integration of Ethernet VPN (EVPN) with Virtual Private LAN Service (VPLS) and Their Provider Backbone Bridge (PBB) Equivalents", RFC 8560, DOI 10.17487/RFC8560, May 2019, <<https://www.rfc-editor.org/info/rfc8560>>.

## 12.2. Informative References

- [RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, DOI 10.17487/RFC4448, April 2006, <<https://www.rfc-editor.org/info/rfc4448>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <<https://www.rfc-editor.org/info/rfc4762>>.

## Authors' Addresses

Wen Lin  
Juniper Networks, Inc.  
  
EMail: [wlin@juniper.net](mailto:wlin@juniper.net)



Bin Wen  
Comcast

EMail: bin\_wen@comcast.com

Voiteck Kozak  
Comcast

EMail: voitek\_kozak@comcast.com

Jorge Rabadan  
Nokia

EMail: jorge.rabadan@nokia.com

BESS Working Group  
Internet-Draft  
Intended status: Informational  
Expires: May 7, 2020

S. Litkowski  
S. Agrawal  
Cisco  
K. Patel  
Arrcus  
S. Zhuang  
Huawei  
November 4, 2019

Modifying RFC5549 VPNv4 over IPv6 next hop handling procedures  
draft-litkowski-bess-vpnv4-ipv6-nh-handling-00

Abstract

RFC4364 and RFC4659 define respectively BGP extensions to provide VPN-IPv4 and VPN-IPv6 services. When defined RFC5549 has brought up an inconsistency in how the next hop is encoded when a VPN-IPv4 NLRI carries an IPv6 next hop compared to RFC4364 and RFC4659. For some reasons, existing and deployed implementations of RFC5549 haven't followed the specification and are using an VPN-IPv6 next hop as in RFC4364 and RFC4659. Moving these implementations to be compliant with RFC5549 may break existing network deployments. This document proposes a modification of RFC5549 to enable compliancy of these implementations. These document also proposes additional modifications of RFC5549 to address missing points.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 <https://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 May 7, 2020.

#### Copyright Notice

Copyright (c) 2019 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 (<https://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
2. Requested modifications . . . . .	3
2.1. Modifying next hop encoding . . . . .	3
2.2. Handling of VPN IPv4 multicast SAFI . . . . .	4
3. Deployment considerations . . . . .	6
4. Security Considerations . . . . .	7
5. Acknowledgements . . . . .	7
6. IANA Considerations . . . . .	7
7. References . . . . .	7
7.1. Normative References . . . . .	7
7.2. Informative References . . . . .	8
Authors' Addresses . . . . .	8

#### 1. Problem statement

[RFC4364] and [RFC4659] define respectively BGP extensions to provide VPN-IPv4 and VPN-IPv6 services.

[RFC4364] defines only VPN-IPv4 carried with an IPv4 next hop. For historical reasons, as per Section 4.3.2 of [RFC4364], the next hop address is encoded as a VPN-IPv4 address with an RD of 0. The expected next hop length value is 12 bytes. As stated in Section 4.3.2 of [RFC4364], the justification of using a VPN-IPv4 address in the next hop field came from [RFC2858] which required the next hop address to be in the same address family as the Network Layer Reachability Information.

[RFC4659] defines VPN-IPv6 carried over with either an IPv4 or IPv6 next hop. When IPv4 transport is used, the next hop is encoded as a VPN-IPv6 address with an RD set to 0 followed by the IPv4-mapped IPv6 address of the advertising BGP speaker. The expected next hop length is 24 bytes. When IPv6 transport is used, the next hop is encoded as one or two VPN-IPv6 address(es) still using an RD set to 0. Section 3.2.1.1 of [RFC4659] clearly states: "the BGP speaker SHALL advertise a Next Hop Network Address field containing a VPN-IPv6 address (...) This is potentially followed by another VPN-IPv6 address".

[RFC5549] specifies, among other, the extensions to allow advertising VPN-IPv4 NLRI with an IPv6 protocol next hop. In such a case the next hop of the NLRI is encoded as one or two IPv6 addresses. [RFC5549] does not use VPN-IPv6 addresses but regular IPv6 addresses (no RD) in the next hop field. Refer to Section 4 and Section 6.2 of [RFC5549] for more details. As a consequence, [RFC5549] brings an inconsistency in how the next hop is encoded for VPN SAFIs compared to [RFC4364] and [RFC4659].

While, from a pure specification point of view, this inconsistency between next hop encodings does not create any issue, several existing implementations are using a consistent encoding of the next hop using VPN-IPv4 format (using RD set to 0) for all the cases listed above. Authors have looked at nine implementations including the major ones deployed in the market, and all these implementations are encoding the next hop using a VPN-IPv4 format (with RD set to 0), except two which does not support [RFC5549] at all. Although these multiple implementations are not compliant with [RFC5549], modifying these implementations may create backward compatibility issues as well as operation pain for operators who have deployed.

In addition, [RFC5549] only deals with VPN-IPv4 unicast address-family (AFI=1, SAFI=128), but does not handle the case of the VPN-IPv4 multicast address family (AFI=1, SAFI=129).

This document proposes a modification of [RFC5549] for the VPN-IPv4 family to address these problems.

## 2. Requested modifications

### 2.1. Modifying next hop encoding

While authors agree that nowadays using a VPN-IP address in a BGP next hop field does not make any sense, to accomodate running codes, deployments and bring consistency with legacy, authors propose [RFC5549] next hop encoding rules to be modified when IPVPN SAFIs are used.

This document proposes to add the following text as part of Section 4 of [RFC5549]:

- o When the AFI=1 and when the SAFI is an IPVPN SAFI (128 or 129), a BGP speaker MUST encode the next hop as VPN-IPv6 address(es) with an RD set to zero.

To accomodate this text, the example provided in Section 6.2 of [RFC5549] must also be modified as follows:

- o Section 6.2 title must be changed to "IPv4 VPN unicast over IPv6 Core"

- o OLD TEXT:

The MP\_REACH\_NLRI is encoded with:

- + AFI = 1
- + SAFI = 128
- + Length of Next Hop Network Address = 16 (or 32)
- + Network Address of Next Hop = IPv6 address of Next Hop
- + NLRI = IPv4-VPN routes

- o NEW TEXT:

The MP\_REACH\_NLRI is encoded with:

- + AFI = 1
- + SAFI = 128
- + Length of Next Hop Network Address = 24 (or 48)
- + Network Address of Next Hop = IPv6 address of Next Hop encoded as a VPN-IPv6 address with RD set to 0
- + NLRI = IPv4-VPN routes

## 2.2. Handling of VPN IPv4 multicast SAFI

VPN IPv4 multicast SAFI (AFI=1, SAFI=129) must be handled in the same way as the VPN IPv4 unicast SAFI (AFI=1, SAFI=128).

This document proposes to modify [RFC5549] as follows to accomodate this change:

- o Section 3:

OLD TEXT:

The following current AFI/SAFI definitions for the IPv4 NLRI or VPN-IPv4 NLRI (<1/1>, <1/2>, <1/4>, and <1/128>) only have provisions for advertising a Next Hop address that belongs to the IPv4 protocol.

NEW TEXT:

The following current AFI/SAFI definitions for the IPv4 NLRI or VPN-IPv4 NLRI (<1/1>, <1/2>, <1/4>, <1/128>, and <1/129>) only have provisions for advertising a Next Hop address that belongs to the IPv4 protocol.

- o Section 3:

OLD TEXT:

This is in addition to the current mode of operation allowing advertisement of NLRI for <AFI/SAFI> of <1/1>, <1/2> and <1/4> with a next hop address of IPv4 type and advertisement of NLRI for <AFI/SAFI> of <1/128> with a next hop address of VPN-IPv4 type.

NEW TEXT:

This is in addition to the current mode of operation allowing advertisement of NLRI for <AFI/SAFI> of <1/1>, <1/2> and <1/4> with a next hop address of IPv4 type and advertisement of NLRI for <AFI/SAFI> of <1/128> and <1/129> with a next hop address of VPN-IPv4 type.

- o Section 3 line "SAFI = 1, 2, 4, or 128" must be changed to "SAFI = 1, 2, 4, 128, or 129".
- o Section 4 line "NLRI SAFI = 1, 2, 4, or 128" must be changed to "NLRI SAFI = 1, 2, 4, 128, or 129".
- o A Section 6.3 named "IPv4 VPN multicast over IPv6 Core" may be added to provide an example with the following text:

The extensions defined in this document may be used for support of IPv4 VPNs for multicast over an IPv6 backbone. In this application, PE routers would advertise VPN-IPv4 multicast NLRI in the MP\_REACH\_NLRI along with an IPv6 Next Hop.

The MP\_REACH\_NLRI is encoded with:

- o AFI = 1
- o SAFI = 129
- o Length of Next Hop Network Address = 24 (or 48)
- o Network Address of Next Hop = IPv6 address of Next Hop encoded as a VPN-IPv6 address with RD set to 0
- o NLRI = IPv4-VPN routes

During BGP Capability Advertisement, the PE routers would include the following fields in the Capabilities Optional Parameter:

- o Capability Code set to "Extended Next Hop Encoding"
- o Capability Value containing <NLRI AFI=1, NLRI SAFI=129, next hop AFI=2>

### 3. Deployment considerations

As most of the vendors and deployments today are already implementing the VPN-IPv6 address in the next hop field, interoperability in these deployments will not be broken when modifying [RFC5549]. Even if authors have polled multiple vendors including all the major players of the market, there is still a possibility that an existing implementation strictly follows [RFC5549] as it is today. While it should be unlikely, it could happen. In case such a situation exists today, this compliant implementation is not interoperable with most of the implementations of the market and code changes are required (at one side or the other) to get the interoperability. It should be noted that no interoperability issue has been brought to vendors by customers or during interoperability testing between vendors at EANTC for example. By modifying [RFC5549] as we propose, this hypothetical compliant implementation will not be compliant anymore and will require code change to become compliant. This code change can simply be a knob on a per-neighbor basis to accommodate the behavior of the neighbor without breaking any hypothetical deployment between RFC5549 compliant implementations.

As IETF is driven by running code, authors think that changing the existing standard to accomodate running codes and deployments will help the overall industry without causing damages. In case a compliant implementation exists today (but again it is really unlikely), this implementation can add a knob to provide new compliancy and interoperability. This approach will require fewer code changes within the whole industry and then keep most of the existing deployments more stable.

#### 4. Security Considerations

This document does not introduce any additional security issue compared to [RFC4364], [RFC4659] and [RFC5549].

#### 5. Acknowledgements

#### 6. IANA Considerations

IANA has no action.

#### 7. References

##### 7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <<https://www.rfc-editor.org/info/rfc4659>>.
- [RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <<https://www.rfc-editor.org/info/rfc5549>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.



## 7.2. Informative References

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz,  
"Multiprotocol Extensions for BGP-4", RFC 2858,  
DOI 10.17487/RFC2858, June 2000,  
<<https://www.rfc-editor.org/info/rfc2858>>.

## Authors' Addresses

Stephane Litkowski  
Cisco

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

Swadesh Agrawal  
Cisco

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

Keyur Patel  
Arrcus

Email: [keyur@arrcus.com](mailto:keyur@arrcus.com)

Shunwan Zhuang  
Huawei

Email: [zhuangshunwan@huawei.com](mailto:zhuangshunwan@huawei.com)

BESS Working Group  
Internet Draft  
Intended Status: Proposed Standard

N. Malhotra, Ed.  
Individual

K. Patel  
Arrcus

J. Rabadan  
Nokia

Expires: May 5, 2020

Nov 2, 2019

PE-CE Control Plane for EVPN  
draft-malhotra-bess-evpn-pe-ce-00

Abstract

In an EVPN network, EVPN PE's provide VPN bridging and routing service to connected CE devices based on BGP EVPN control plane. At present, there is no PE-CE control plane defined for an EVPN PE to learn CE MAC, IP, and any other routes from a CE that may be distributed in EVPN control plane to enable unicast flows between CE devices. As a result, EVPN PE's rely on data plane based gleaning of source MACs for CE MAC learning, ARP/ND snooping for CE IPv4/IPv6 learning, and in some cases, local configuration for learning prefix routes behind a CE. A PE-CE control plane alternative to this traditional learning approach, where applicable, offers certain distinct advantages that in turn result in simplified EVPN operation.

This document defines a PE-CE control plane as an optional alternative to traditional non-control-plane based PE-CE learning in an EVPN network. It defines PE-CE control plane procedures and TLVs based on L3DL as the base protocol, enumerates advantages that may be achieved by using this PE-CE control plane, and discusses in detail EVPN use cases that are simplified as a result.

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) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1	Introduction . . . . .	4
1.1	Terminology . . . . .	5
2.	PE <-> CE Control Plane Overview . . . . .	7
3.	TLVs . . . . .	9
3.1	Overlay IPv4 Encapsulation PDU . . . . .	9
3.2	Overlay IPv6 Encapsulation PDU . . . . .	11
3.3	Overlay IPv4 Prefix Encapsulation PDU . . . . .	13
3.4	Overlay IPv6 Prefix Encapsulation PDU . . . . .	14
4.	CE MAC/IP Learning on a PE AC . . . . .	15
4.1	PE <-> CE L3DL Session Establishment . . . . .	15
4.2	CE MAC/IP Learning . . . . .	15
5.	PE Any-cast GW MAC/IP Learning on CE . . . . .	15
6.	Remote CE MAC/IP Learning on CE . . . . .	16
7.	PE <-> CE Control Plane with EVPN All-active Multi-Homing . . . . .	17
7.1	All-active Multi-Homing Mode . . . . .	17
7.2	Source MAC . . . . .	18
7.3	CE MAC/IP Learning with EVPN All-active Multi-Homing . . . . .	18
7.4	LAG Member Link Failure . . . . .	19
7.4.1	Session Re-establishment . . . . .	19

7.4.2 TLV Retention . . . . .	19
7.4 LAG Failure . . . . .	19
7.5 Example PE <-> CE Control Plane Flow with All-active Multi-Homing . . . . .	20
8. Software Neighbor Tables . . . . .	22
9. MAC/IP Learning Conflict Resolution . . . . .	22
10. EVPN SLA Signaling . . . . .	23
11. PE-CE Overlay Prefix Learning . . . . .	23
12. Asymmetric EVPN-IRB . . . . .	23
13. Centralized Gateway EVPN-IRB . . . . .	24
14. Use Cases . . . . .	24
14.1 CE Application SLA . . . . .	24
14.2 Simplified EVPN Operations . . . . .	24
14.2.1 EVPN All-active Multi-Homing . . . . .	25
14.2.2 Convergence on CE Host Moves . . . . .	26
14.2.2.1 Silent Hosts . . . . .	26
14.2.2.2 Probing . . . . .	27
14.2.3 ARP Gleaning Latency . . . . .	28
14.3 Applicability to non-EVPN Use Cases . . . . .	28
15. Summary . . . . .	28
16. References . . . . .	30
16.1 Normative References . . . . .	30
16.2 Informative References . . . . .	30
17. Acknowledgements . . . . .	31
Contributors . . . . .	31
Authors' Addresses . . . . .	31

## 1 Introduction

In an EVPN network, CE devices typically connect to an EVPN PE via layer-2 interfaces that terminate in a BD on the PE. Multi-homed LAG interfaces together with EVPN all-active multi-homing procedures are used to achieve PE-CE link and PE node redundancy for fault-tolerance and load-balancing. PEs provide overlay bridging and, optionally, first-hop routing service for these CE devices based on an EVPN control plane that is used to distribute CE MAC, IP, and prefix reachability across PEs.

At present, there is no PE-CE control plane defined for an EVPN PE to learn connected CE host MACs and IPs. As a result, EVPN PEs rely on:

- o data plane based gleaning of source MAC for MAC learning,
- o ARP snooping for IPv4 + MAC learning, and
- o ND snooping for IPv6 + MAC learning.

A PE-CE control plane alternative to this traditional learning approach, where applicable, can offer some distinct advantages across various boot-up, mobility, and convergence scenarios:

- o PE-CE learning is decoupled from non-deterministic hashing of data, ARP, and ND packets from CEs over all-active multi-homed LAG interfaces.
- o PE-CE learning is decoupled from non-deterministic periodicity of data traffic from CEs or, in an extreme scenario, from CE device being silent for an extended period.
- o PE-CE learning is decoupled from non-deterministic CE behavior with respect to unsolicited ARPs and NAs following boot-up and moves.
- o PE-CE learning is decoupled from latencies associated with data packet triggered ARP and ND gleaning.

This results in simplification of certain EVPN operations such as aliasing, MAC and IP syncing across multi-homing PEs, and probing on MAC/IP moves. It also helps achieve a deterministic convergence behavior across various boot-up, mobility, and failure scenarios.

Beside simplification of existing EVPN procedures, PE-CE protocol is also leveraged to enable new use cases that would not be possible otherwise:

- o Signal application SLA requirements to an EVPN PE that may in-turn be used by the PE to influence overlay and underlay routing policies for a host.
- o Signal prefix routes behind a CE for cases where a CE does not run a dynamic routing protocol on the PE-CE link.

This document defines a new PE-CE control plane as an alternative to traditional data-plane and ARP/ND snooping based PE-CE host learning and to local configuration-based PE-CE prefix learning. It defines PE-CE control plane procedures and TLVs based on [L3DL] as the base protocol, enumerates advantages that may be achieved by using this PE-CE control plane, and discusses in detail EVPN operations that are simplified as a result. Use of PE-CE control plane defined in this document is intended to be optional and backwards compatible with CEs that use traditional PE-CE learning within the same BD. While the protocol is discussed using L3DL as the base protocol, signaling described in this document may also, in future, be extended to use LLDP as the base protocol.

### 1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The following terms are used in this document:

- o L3DL: Layer 3 Discovery and Liveness Protocol defined in [L3DL]
- o EVPN-IRB: A BGP-EVPN distributed control plane based integrated routing and bridging fabric overlay discussed in [EVPN-IRB]
- o Underlay: IP or MPLS fabric core network that provides IP or MPLS routed reachability between EVPN PEs.
- o Overlay: VPN or service layer network consisting of EVPN PEs OR VPN provider-edge (PE) switch-router devices that runs on top of an underlay routed core.
- o EVPN PE: A PE switch-router in a data-center fabric that runs overlay BGP-EVPN control plane and connects to overlay CE host devices. An EVPN PE may also be the first-hop layer-3 gateway for CE/host devices. This document refers to EVPN PE as a logical function in a data-center fabric. This EVPN PE function may be physically hosted on a top-of-rack switching device (ToR) OR at layer(s) above the ToR in the Clos fabric. An EVPN PE is typically also an IP or MPLS tunnel end-point for overlay VPN flows.
- o CE: A tenant host device that has layer 2 connectivity to an EVPN PE switch-router, either directly OR via intermediate switching device(s).
- o Symmetric EVPN-IRB: An overlay fabric first-hop routing architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed at both ingress and egress EVPN PEs.
- o Asymmetric EVPN-IRB: An overlay fabric first-hop routing

architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed and bridged at ingress PE and bridged at egress PEs.

- o Centralized EVPN-IRB: An overlay fabric first-hop routing architecture, wherein, overlay host-to-host routed inter-subnet flows are routed at a centralized gateway, typically at the one of the spine layers, and where EVPN PEs are pure bridging devices.
- o ARP: Address Resolution Protocol [RFC 826].
- o ND: IPv6 Neighbor Discovery Protocol [RFC 4861].
- o Ethernet-Segment: physical Ethernet or LAG port that connects an access device to an EVPN PE, as defined in [RFC 7432].
- o ESI: Ethernet Segment Identifier as defined in [RFC 7432].
- o LAG: Layer-2 link-aggregation, also known as layer-2 bundle port-channel, or bond interface.
- o EVPN all-active multi-homing: PE-CE all-active multi-homing achieved via a multi-homed layer-2 LAG interface on a CE with member links to multiple PEs and related EVPN procedures on the PEs.
- o EVPN Aliasing: multi-homing procedure as defined in [RFC 7432].
- o BD: Broadcast Domain.
- o Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.
- o AC: A PE Attachment Circuit. This may be an access (untagged) or trunk (tagged) layer-2 interface that is a member of a local VLAN or a BD.
- o SLA: Service Layer Agreement

## 2. PE <-> CE Control Plane Overview

Layer 3 Discovery and Liveness (L3DL) protocol is defined in [L3DL] as a protocol over Ethernet links to auto-discover connected neighbor's layer 2, layer 3 attributes, and encapsulations for the purpose of bringing up upper layer routing protocols. This document leverages L3DL as a PE-CE protocol in an EVPN network fabric on access links between an EVPN PE and CE. Specifically,

- o PE-CE control plane based on L3DL protocol is proposed for CE MAC learning as an alternative to data-plane based source MAC learning.
- o PE-CE control plane based on L3DL protocol is proposed for CE MAC-IP adjacency learning as an alternative to MAC-IP learning based on ARP/ND snooping.
- o PE-CE control plane based on L3DL is proposed for learning of IP Prefixes and associated overlay indexes, as an alternative to local configuration on the PE for use case defined in section 4.1 of [EVPN-PREFIX-ADV].

Note that any specification related to base L3DL protocol itself is considered out of scope for this document and will continue to be covered in the base protocol spec. This document will instead focus on procedures and TLV extensions needed to achieve the above learning on PE-CE links in an EVPN network. Any text that relates to the base protocol included in this document is simply background information in the context of use cases covered in this document. The reader should refer to the base L3DL protocol document for the exact L3DL protocol specification.

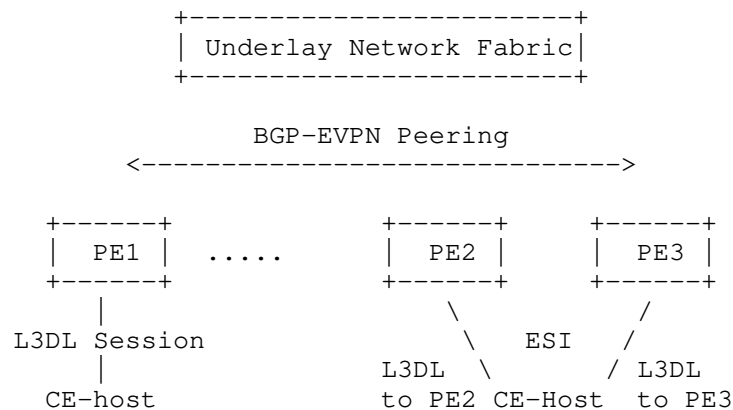


Figure 1



An L3DL session is established on layer-2 logical interfaces between the EVPN PE and each connected CE host device. A session end-point on a local logical interface is identified by peer Logical Link Endpoint Identifier (LLEI) as defined in [L3DL]. L3DL HELLO messages are used for end-point discovery and OPEN messages are exchanged between two end-points to establish an L3DL peering. Once L3DL peering is established, encapsulation TLVs are exchanged for learning.

In the context of an EVPN network, CE Attachment Circuits (AC logical interfaces) typically terminate in a BD on the PE, with multi-homed LAG interfaces used for EVPN all-active multi-homing. CE hosts may be directly connected to EVPN PEs via access ports, or may be connected on trunk-ports via another switch. In a common EVPN-IRB design, EVPN PEs also function as distributed first-hop gateways for hosts in a BD. While symmetric and asymmetric IRB designs are possible as discussed in [EVPN IRB], procedures described in subsequent sections assume symmetric IRB with distributed any-cast gateways on EVPN PEs. Any deviations from these procedures for asymmetric IRB design or a centralized IRB design will be covered in future updates to this document.

The next few sections will focus on additional L3DL TLVs and procedures needed for PE-CE learning on EVPN PE ACs without and with all-active multi-homing.

### 3. TLVs

This section defines new TLVs that are used by PE-CE control plane defined in this document.

#### 3.1 Overlay IPv4 Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv4 and MAC bindings. Alternatively, it may also be used to carry an overlay MAC with a NULL IPv4 address in a non-IRB use case.

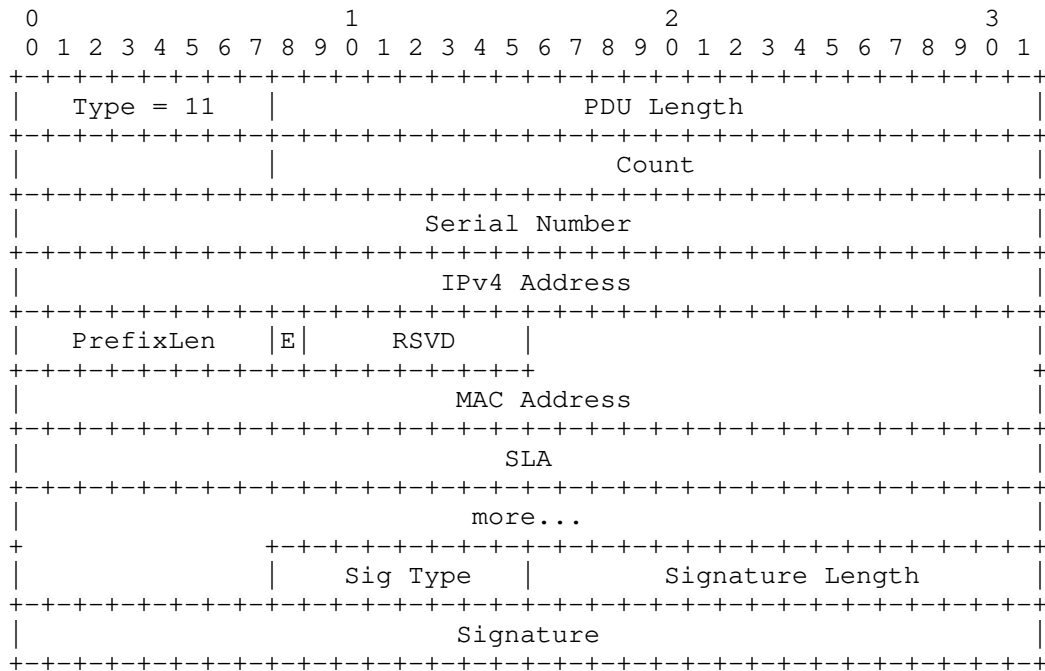


Figure 2

- o A new L3DL PDU type (11) is requested for this PDU.
- o The IPv4 Address is that of an overlay.
- o MAC address carries the MAC binding for the particular IPv4 address if one is set in the PDU. If an IPv4 address is not set, it simply signals an overlay MAC address.
- o EVPN flag 'E' indicates if this encapsulation is being sent on behalf of a remote host learnt via EVPN. Use of this flag is covered in a later section.
- o A 32 bit 'SLA' word is used to signal SLA requirements of a CE host to the EVPN PE. An EVPN PE may use these to implement

routing policies needed to fulfil the CE SLA requirement. As an example, if a CE indicates a minimum delay requirement for the applications it runs, EVPN provider network may route or bridge traffic destined to this host over traffic engineered paths that implement a minimum delay routing policy.

In addition to carrying CE host IP and MAC to a PE, this PDU may also be used to carry PE's any-cast gateway IPv4 address and MAC bindings to a CE host device. Optionally, it may also be used to relay a remote CE's IPv4 address and MAC bindings to a local CE host within a subnet. Procedures related to use of this PDU are discussed in subsequent sections.

The encapsulation list in this PDU MUST follow full replace semantics as in the L3DL protocol specification.

### 3.2 Overlay IPv6 Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv6 and MAC bindings:

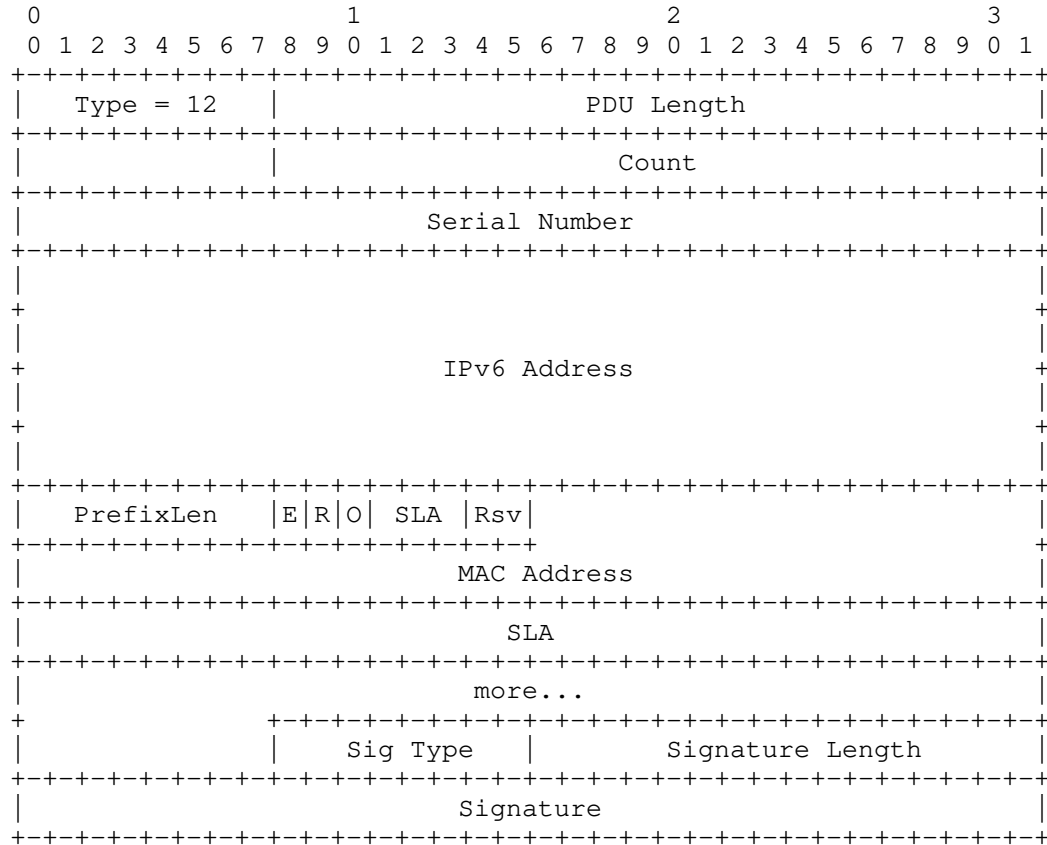


Figure 3

- o A new L3DL PDU type (12) is requested for this PDU.
- o The IPv6 Address is that of an overlay.
- o MAC address carries the MAC binding for IPv6 address in the PDU.
- o An EVPN flag 'E' indicates if this encapsulation is being sent on behalf of a remote host learnt via EVPN. Usage of this flag is covered in a later section.
- o A Router flag 'R' is used to carry "Router Flag" or "R-bit" as defined in [RFC4861]. Usage of this flag for the purpose of installing ND cache entries based on learning via this TLV is as defined in [RFC4861]

- o An Override flag 'O' is used to carry "Override Flag" or "O-bit" as defined in [RFC4861]. Usage of this flag for the purpose of installing ND cache entries based on learning via this TLV is as defined in [RFC4861]
- o A 32 bit 'SLA' word is used to signal SLA requirements of a CE host to the EVPN PE. An EVPN PE may use these to implement routing policies needed to fulfil the CE SLA requirement. As an example, if a CE indicates a minimum delay requirement for the applications it runs, EVPN provider network may route or bridge traffic destined to this host over traffic engineered paths that implement a minimum delay routing policy.

In addition to carrying CE host IP and MAC to a PE, this PDU may also be used to carry PE's any-cast gateway IPv6 address and MAC bindings to a CE host device. Optionally, it may also be used to relay a remote CE's IPv6 address and MAC bindings to a local CE host within a subnet. Procedures related to use of this PDU are discussed in subsequent sections.

The encapsulation list contained in this PDU MUST follow full replace semantics as in the L3DL protocol specification.

### 3.3 Overlay IPv4 Prefix Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv4 prefix routes for prefixes behind a CE that does not run a dynamic routing protocol for use-case as defined in section 4.1 of [EVPN-PREFIX-ADV]:

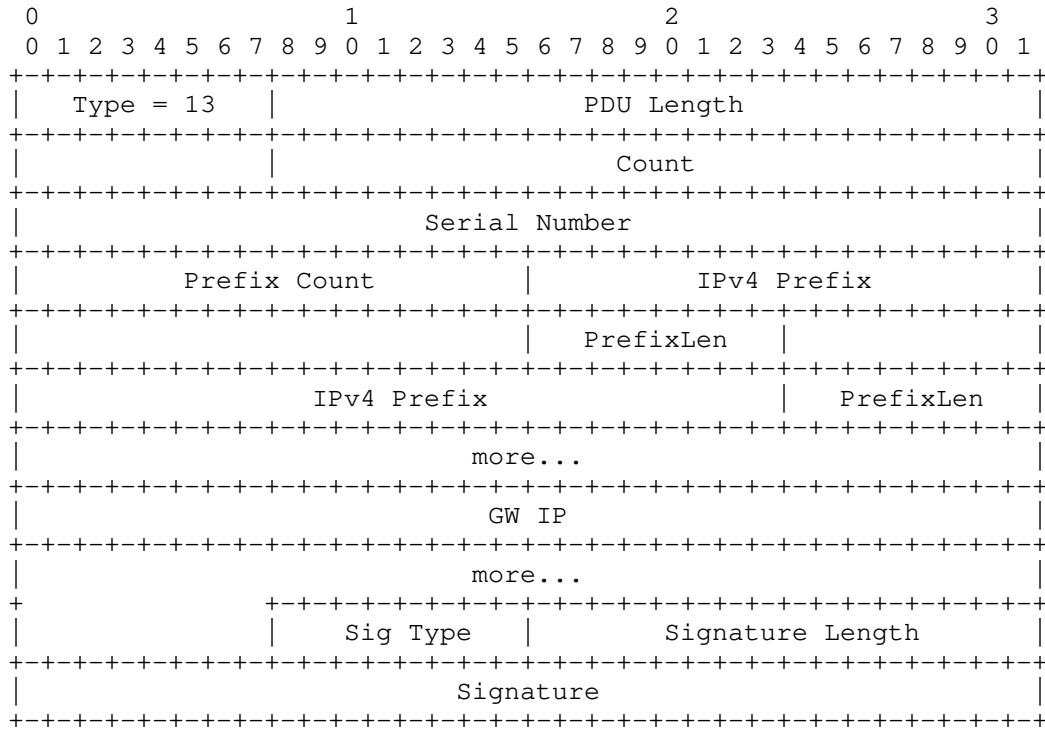


Figure 4

A CE device as defined in [EVPN-PREFIX-ADV], with prefixes behind it MAY use the above PDU to send these prefixes to an EVPN PE with itself as the GW. An EVPN PE MAY then advertise prefixes received via this PDU as RT-5, with TS as the GW, as defined in [EVPN-PREFIX-ADV].

- o A new L3DL PDU type (10) is requested for this PDU.
- o IPv4 Prefix is set to a prefix behind a CE.
- o PrefixLen is set to IPv4 prefix length for the advertised prefix.
- o GW-IP is set to the CE IPv4 address (advertised via Type 8 PDU).

Multiple prefixes may be set for a single GW IP. The encapsulation list contained in this PDU MUST follow full replace semantics as in the L3DL protocol specification.

### 3.4 Overlay IPv6 Prefix Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv6 prefix routes for prefixes behind a CE that does not run a dynamic routing protocol for use-case as defined in section 4.1 of [EVPN-PREFIX-ADV]:

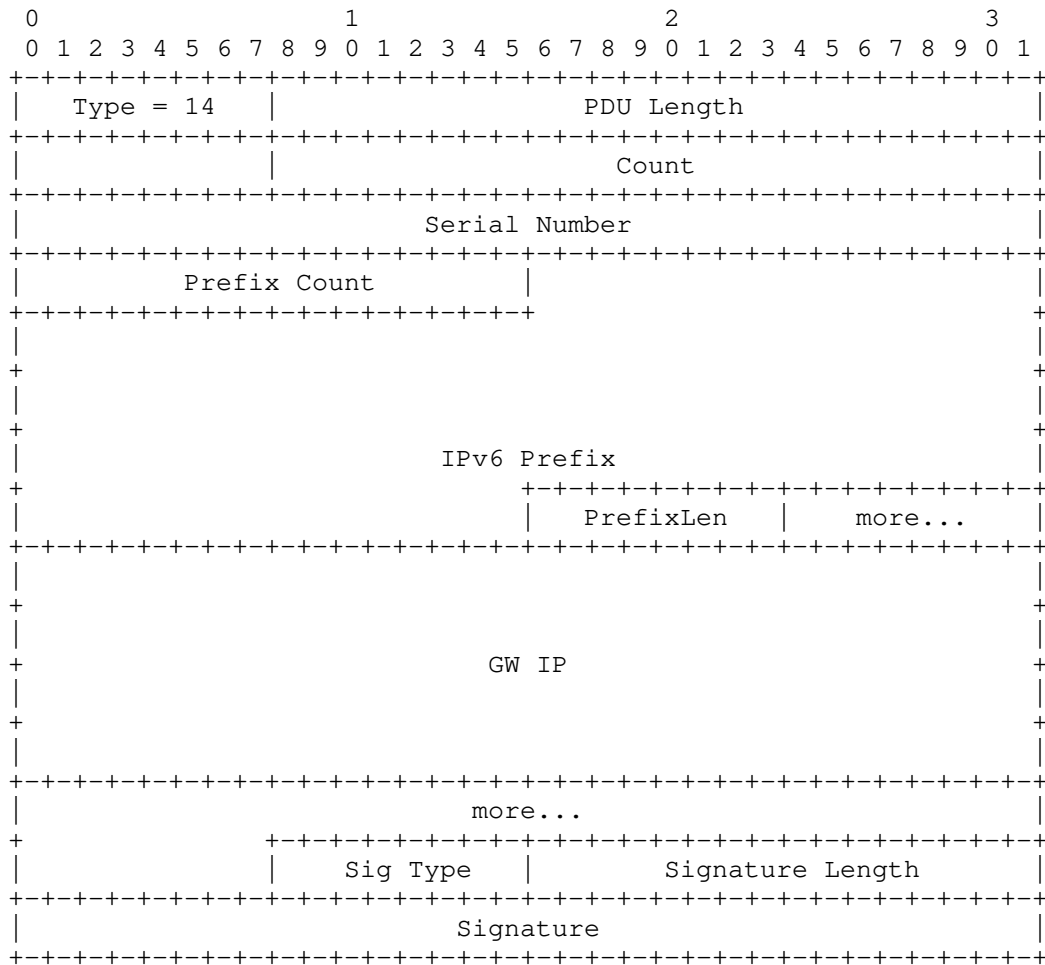


Figure 5

A CE device as defined in [EVPN-PREFIX-ADV], with prefixes behind it MAY use the above PDU to send these prefixes to an EVPN PE with itself as the GW. An EVPN PE MAY then advertise prefixes received via this PDU as RT-5, with TS as the GW, as defined in [EVPN-PREFIX-ADV].

- o A new L3DL PDU type (14) is requested for this PDU.
- o IPv6 Prefix is set to an IPv6 prefix behind a CE.
- o PrefixLen is set to IPv6 prefix length for the advertised prefix.
- o GW-IP is set to the CE IPv6 address (advertised via Type 9 PDU).

Multiple prefixes may be set for a single GW IP. The encapsulation list contained in this PDU MUST follow full replace semantics as in the L3DL protocol specification.

#### 4. CE MAC/IP Learning on a PE AC

This section defines procedures for learning a connected CE MAC and IP on a PE local attachment circuit (AC).

##### 4.1 PE <-> CE L3DL Session Establishment

On an EVPN PE,

- o A HELLO and/or OPEN PDU sent from a CE host source MAC is received on a tagged or untagged interface that is member of a local BD, referred here to as an AC.
- o OPEN messages are exchanged with the host on the AC.
- o L3DL session is established to the host source MAC and bound to a local AC.

##### 4.2 CE MAC/IP Learning

Overlay IPv4 and IPv6 encapsulation PDU types 8/9 from a CE are used for the purpose of CE MAC/IP learning on a PE:

- o The EVPN flag 'E' MUST NOT be set in type 8/9 PDU from a CE.
- o A MAC entry for the MAC received in a type 8/9 PDU MUST be installed in the MAC-VRF table pointing to the AC to which the session is bound.
- o If an IPv4/IPv6 address is set in the PDU, an IPv4/IPv6 neighbor binding MUST be established for the IPv4/IPv6 address in the PDU to the MAC address in the PDU. In other words, a next-hop re-write for these IPv4/IPv6 neighbor entries MUST be installed using the MAC address in the PDU, and if required by forwarding logic, bound to the AC associated with the L3DL session.
- o Note that an IPv4/IPv6 address MAY NOT be set in a type 8/9 PDU received from a CE, in which case this PDU is only used for MAC learning. This MAY be the case in a non-IRB EVPN network, wherein, an EVPN PE is not a first-hop router for the attached CEs.

#### 5. PE Any-cast GW MAC/IP Learning on CE

If L3DL based host learning is enabled on a PE with a distributed



any-cast gateway on the EVPN PE,

- o EVPN PE MUST send type 8/9 Overlay Encapsulation PDUs on associated ACs with L3DL sessions toward CE hosts.
- o Type 8/9 PDUs from an EVPN PE MUST be encoded with the any-cast gateway IPv4/IPv6 address and any-cast gateway MAC address.
- o EVPN flag 'E' MUST NOT be set in this PDU.
- o A CE MAY process type 8/9 PDUs to establish GW IP to MAC bindings and learn gateway MAC to LAG AC bindings, similar to handling of type 8/9 PDUs on the PE described above.

Handling of type 8/9 PDUs for the purpose of gateway learning on the host is desirable but optional. A CE MAY continue to use ARP and ND for this purpose.

#### 6. Remote CE MAC/IP Learning on CE

For CE to CE intra-subnet flows across the overlay, CE needs to learn and install a neighbor IP to MAC binding for remote CEs. This is handled today either by flooding ARP/ND requests across the overlay bridge and optionally implementing an ARP/ND suppression cache on the PE that is populated via MAC+IP EVPN route-type 2. ARP/ND request frames are trapped on the PE that does a local ARP/ND reply on behalf of the remote CE. If L3DL based learning is enabled in the fabric, L3DL may be used for this purpose to avoid overlay ARP/ND flooding, data frame triggered ARP learning, and to avoid maintaining an ARP suppression cache on the PE.

- o Remote MAC-IP routes learned via BGP EVPN route-type 2 that are imported to a local MAC-VRF MAY also be sent as type 8/9 PDUs on L3DL sessions to CEs over local ACs in that BD.
- o EVPN flag 'E' MUST be set in this encapsulation in the PDU.
- o A CE MAY install IPv4/IPv6 neighbor MAC bindings for remote CEs within a subnet based on 'E' flagged type 8/9 PDUs received from the PE.

Handling of type 8/9 PDUs for this purpose is optional but desirable to get full benefit of a fabric that is completely setup on boot-up, avoids overlay flooding, and is decoupled from latencies associated with data plane driven ARP and ND learning.

## 7. PE &lt;-&gt; CE Control Plane with EVPN All-active Multi-Homing

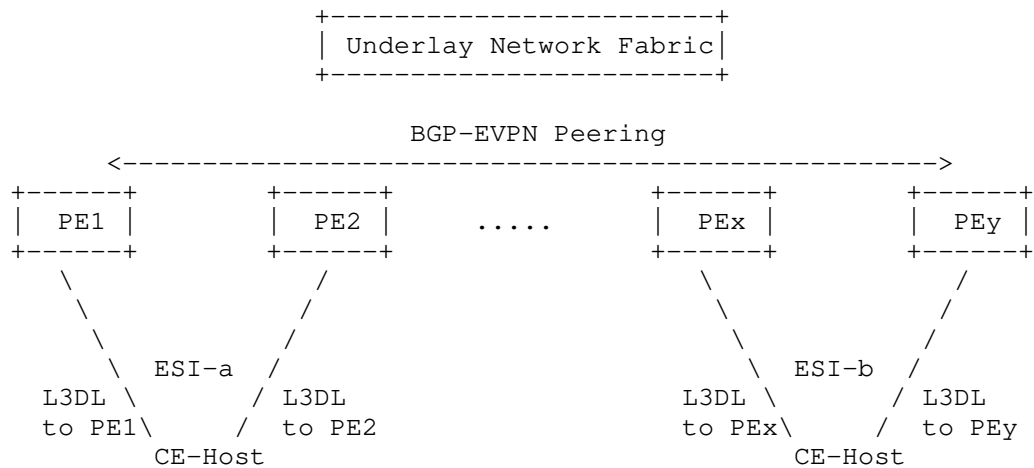


Figure 6

In an EVPN all-active multi-homing setup, a LAG interface on the CE includes member physical ports that connect to multiple PE devices. A subset of these member ports that terminate at a PE are configured as members of a local LAG interface at that PE. A LAG AC at the PE is a logical interface in a BD, identified by this LAG interface and optionally, an Ethernet Tag in case of trunk ports.

In order for L3DL based learning to work with EVPN all-active multi-homing, a separate L3DL peering MUST be established between the CE host and each PE device. For this reason, while an EVPN PE MAY form an L3DL peering to a CE host on its local LAG AC, the CE host MUST form an L3DL peering to a PE on a local LAG "member physical port".

A configurable All-active Multi-Homing mode is defined below in order to be able to bind an L3DL peering to a LAG member-port as opposed to a LAG interface.

## 7.1 All-active Multi-Homing Mode

When configured to run on a local LAG port in this mode,

- o L3DL HELLO messages MUST be replicated on ALL LAG member ports.
- o An L3DL OPEN message sent in response to a HELLO MUST be sent on the LAG member port on which the HELLO was received.
- o An L3DL session MUST be bound to the local LAG member port on

which the OPEN message was received.

- o L3DL encapsulation PDUs MUST be sent on the local LAG member port on which the session was bound.
- o L3DL Keep-Alives MUST be sent on the local LAG member port on which the session was bound.

Note that this may result in a PE receiving multiple HELLO PDUs from a CE end-point. This however is harmless, as per the [L3DL] specification. A PE simply drops redundant HELLOs from a MAC that it has already replied to with an OPEN, within a retry time window.

## 7.2 Source MAC

L3DL relies on the source MAC address in the Ethernet frame to establish a peering. When running L3DL on a LAG port (in all-active multi-homing mode or regular mode), L3DL frames MUST use the LAG interface MAC as the source MAC address in the Ethernet frame.

## 7.3 CE MAC/IP Learning with EVPN All-active Multi-Homing

In order to accomplish MAC/IP learning of CE host devices multi-homed to EVPN fabric PEs via EVPN All-active Multi-Homing:

- o A multi-homed CE device MUST be configured to run L3DL on a local LAG interfaces in All-active Multi-Homing mode defined above.
- o EVPN PE MAY run L3DL on local LAG interfaces to multi-homed CE devices in regular mode.
- o EVPN PEs that share the same Ethernet Segment MUST use unique source MACs (that of the local LAG) in HELLO/OPEN messages to establish separate L3DL sessions to a CE.

With the above rules in place,

- o An L3DL session on the CE is bound to a local LAG member-port.
- o An L3DL session on the PE is bound to a local LAG AC port.
- o A single L3DL session is established at the PE to a CE on the local LAG AC.
- o 'N' L3DL sessions are established at the CE, one to each PE on a local LAG member interface, where N = number of multi-homing PEs in an Ethernet Segment.

Once an L3DL session is established as above, all other host learning procedures defined earlier for CE MAC/IP learning on a PE's AC port apply as is to a LAG AC in an EVPN all-active multi-homing setup.

#### 7.4 LAG Member Link Failure

On a CE that is running in all-active multi-homing mode, an L3DL session to a PE is bound to a LAG member interface. If the link that the L3DL session is bound to fails, L3DL session will get torn down at the CE by virtue of the session interface going down. If the CE has additional active member link(s) to this PE, a new L3DL session must be established on one of the active member links via HELLO PDUs sent by the CE on its remaining active member links to the PE.

##### 7.4.1 Session Re-establishment

L3DL session at the CE is torn down immediately following the session interface failure. While the LAG interface at the PE is still operationally UP, L3DL session at the PE is subject to Keep Alive PDUs received from the CE. Once the session expires at the PE because of missed Keep Alive PDUs from the CE, PE will respond to HELLO on one of the active member link with an OPEN to re-establish a new session. Note that the new session is still bound to the LAG AC at the PE and to a new member link at the CE.

##### 7.4.2 TLV Retention

TLVs learnt from a CE over a failed session MUST be retained at the PE if the PE LAG AC is still operationally up following a member link failure because of active member link(s) in the LAG. TLV retention logic at the PE MAY be based on an age-out time, that is a local matter at the PE. TLV age-out time MUST be higher than the missed Keep Alive duration, after which the session is considered closed. Once a new L3DL session is established, PE MUST implement a mark and sweep logic to reconcile retained TLVs from the CE peer with the new set of TLVs received from this CE.

#### 7.4 LAG Failure

When a LAG member link failure results in the LAG interface being operationally down, TLV age-out logic discussed above MUST NOT be in effect. L3DL session MAY be considered as DOWN immediately on the LAG being down at the PE. This is so that, in the event of a total connectivity loss between a PE and CE, CE learnt routes can be withdrawn immediately.

## 7.5 Example PE &lt;-&gt; CE Control Plane Flow with All-active Multi-Homing

An example L3DL over all-active multi-homing session flow is discussed below for clarity.

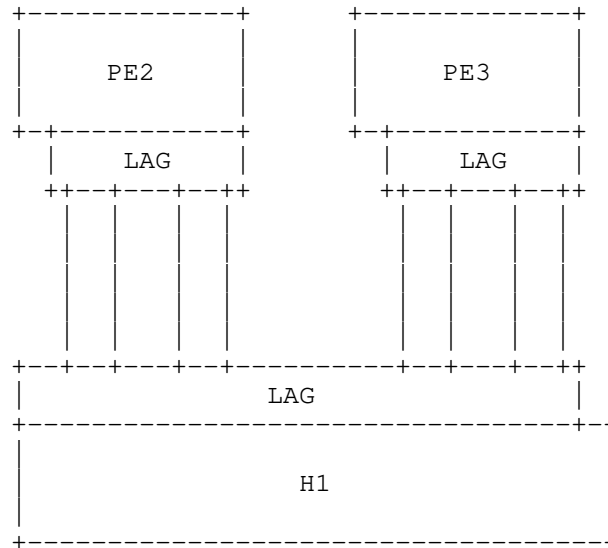


Figure 7

Example topology with CE H1 multi-homed to PE2 and PE3 via EVPN all-active multi-homing LAG with four member ports to each PE:

H1 member ports to PE2: i121, i122, i123, i124

PE2 member ports to H1: i211, i212, i213, i214

H1 member ports to PE3: i131, i132, i133, i134

PE3 member ports to H1: i311, i312, i313, i314

H1 LAG port to PE2/PE3: MLAG1

PE2 LAG port to H1: LAG2

PE3 LAG port to H1: LAG3

H1 LAG MAC: LMAC1

PE2 LAG MAC: LMAC2

PE3 LAG MAC: LMAC3

H1 running L3DL on MLAG1 in All-active Multi-Homing mode

PE2 running L3DL on LAG2 in regular mode

PE3 running L3DL on LAG3 in regular mode

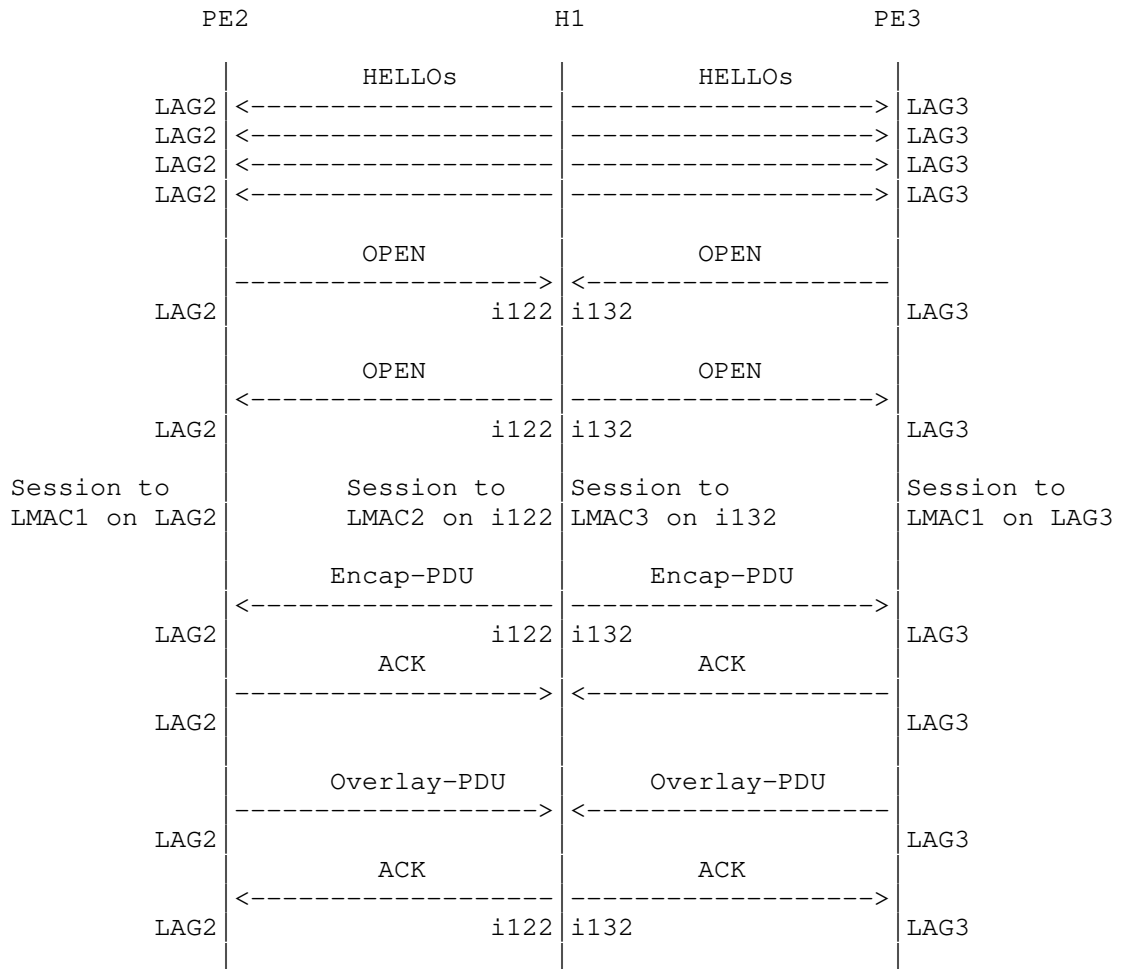


Figure 8

In an example flow shown above:

- o H1: originates HELLO(SMAC=LMAC2) on all MLAG member ports
- o PE2: Multiple HELLO(SMAC=LMAC2) copies received on port LAG2
- o PE3: Multiple HELLO(SMAC=LMAC2) copies received on port LAG3
- o PE2: A single OPEN(SMAC=LMAC2, DMAC=LMAC1) sent on port LAG2
- o PE3: A single OPEN(SMAC=LMAC3, DMAC=LMAC1) sent on port LAG3
- o PE2/PE3: duplicate HELLOs from same source LMAC2 are ignored
- o H1: OPEN(SMAC=LMAC2, DMAC=LMAC1) received on member port i122
- o H1: OPEN(SMAC=LMAC1, DMAC=LMAC2) sent on member port i122
- o H1: Session established to LMAC2 on MLAG1 member port i122

- o PE2: Session established to LMAC1 on LAG AC LAG2
- o H1: OPEN(SMAC=LMAC3, DMAC=LMAC1) received on member port i132
- o H1: OPEN(SMAC=LMAC1, DMAC=LMAC3) sent on member port i132
- o H1: Session established to LMAC3 on MLAG member port i132
- o PE3: Session established to LMAC1 on LAG AC LAG3
- o H1: IP encapsulation PDUs (type 4/5) sent to LMAC2 and LMAC3
- o PE2/PE3: H1 MAC and IP are learned
- o PE2/PE3: overlay IP encapsulation PDUs (type 8/9) sent to LMAC1
- o H1: Any-cast GW MAC and IP are learned
- o H1: Remote host MAC and IP are learned

## 8. Software Neighbor Tables

Some networking stack implementations rely on ARP and ND populated neighbor tables for software forwarding. In order to inter-work with such an implementation, an L3DL learned IPv4/IPv6 neighbor entry MAY also be installed in ARP and ND neighbor table as a static / permanent entry.

In addition,

- o Pre-installing L3DL learned neighbor entries may help reduce potential conflict with ARP or ND learned neighbor entries.
- o Pre-installing L3DL learned neighbor entries may help reduce reliance on data traffic triggered ARP requests / ND solicitations and associated learning latency.

With respect to installing IPv6 entries learnt via LSoE in IPv6 ND cache, Router flag (R-bit) and Override flag (O-bit) received in LSoE PDU should be handled as defined in [RFC4861].

## 9. MAC/IP Learning Conflict Resolution

If L3DL learned neighbor entries are not already installed as static entries in ARP/ND neighbor table, it is possible that a neighbor IPv4/IPv6 adjacency may be learned both via L3DL and ARP/ND. Even if L3DL learned entries were pre-installed in neighbor table, a race condition is still possible leading to a potential conflict between ARP/ND learned and L3DL learned neighbor IP adjacency. In such scenarios, L3DL learned entry should be preferred for the purpose of programming neighbor IP adjacencies in forwarding.

With respect to MAC-VRF entries, it is recommended that data plane learning be turned off when L3DL based learning is enabled. However, if it is not, data plane learned entries MUST be reconciled with L3DL learned entries in software and, in case of a conflict, L3DL learned entries preferred if L3DL based learning is enabled.

## 10. EVPN SLA Signaling

Application SLA requirements received from a CE need to be signaled by the local PE to remote PEs in order for remote PEs to route or bridge overlay traffic destined to this CE via traffic engineered paths that meet the SLA. As an example, if SLA requirement for a CE is specified to be "minimum delay", remote PEs need to direct overlay bridged and routed traffic to this CE over traffic engineered underlay paths that implement a "minimum delay" routing policy.

Overlay SLA may also be required to be implemented at different levels of granularity:

- o per-host: [RT-2]
- o per-EVI
- o per-[ESI, EVI]: [RT-1]

Exact signaling specification and handling procedures for the above would be detailed either in future revisions of this document or in a separate document.

## 11. PE-CE Overlay Prefix Learning

[EVPN-PREFIX-ADV] section 4.1 defines a use case, wherein, a PE may advertise IP prefixes and subnets behind a CE. In this use case, CE device does not run a dynamic routing protocol. Instead, these prefixes are learnt on the PE via local policy or configuration. Prefixes are then advertised by PE as RT-5 with the CE as the GW.

PE-CE control plane defined in this document MAY be used to learn these prefixes from a CE as an alternative to local configuration on the PE. Once an L3DL session is established between a CE and a PE, as discussed earlier,

- o A CE MAY send type 10/11 PDUs with these IPv4/IPv6 prefixes over an L3DL session to a PE with the CE IP as the GW IP.
- o A PE MAY advertise prefixes learnt via type 10/11 PDUs as RT-5 with CE IP as the GW IP.

To summarize, A PE would advertise:

- o RT-2 for the CE MAC-IP learnt via type 8/9 PDU
- o RT-5 for Prefixes learnt via type 10/11 PDU with GW IP = CE IP

## 12. Asymmetric EVPN-IRB

Any deviations from the above procedures proposed in this document for asymmetric IRB design will be covered in subsequent updates to



this document.

### 13. Centralized Gateway EVPN-IRB

Any deviations from the above procedures proposed in this document for centralized GW based IRB design will be covered in subsequent updates to this document.

### 14. Use Cases

#### 14.1 CE Application SLA

Application SLA requirements signaled by a CE to an EVPN PE provide a mechanism for EVPN provider network to provide overlay routing and bridging services in accordance with customer application requirements. As an example, a CE may specify an SLA requirement to tunnel overlay application traffic destined to this CE over the lowest delay path. An EVPN PE may signal this SLA requirement to remote PEs along with CE MAC-IP route that in-turn result in the remote PEs bridging and routing traffic destined to this CE over traffic engineered underlay paths that are setup using the lowest delay metric.

Future revisions of this document will specify the exact encoding of SLA bits to achieve different SLA requirements.

#### 14.2 Simplified EVPN Operations

This section will discuss in detail, benefits and simplifications that may be achieved in the context of an EVPN network, if one chooses to implement PE-CE control plane defined in this document as opposed to using traditional data-plane and ARP/ND snooping based PE-CE learning.

## 14.2.1 EVPN All-active Multi-Homing

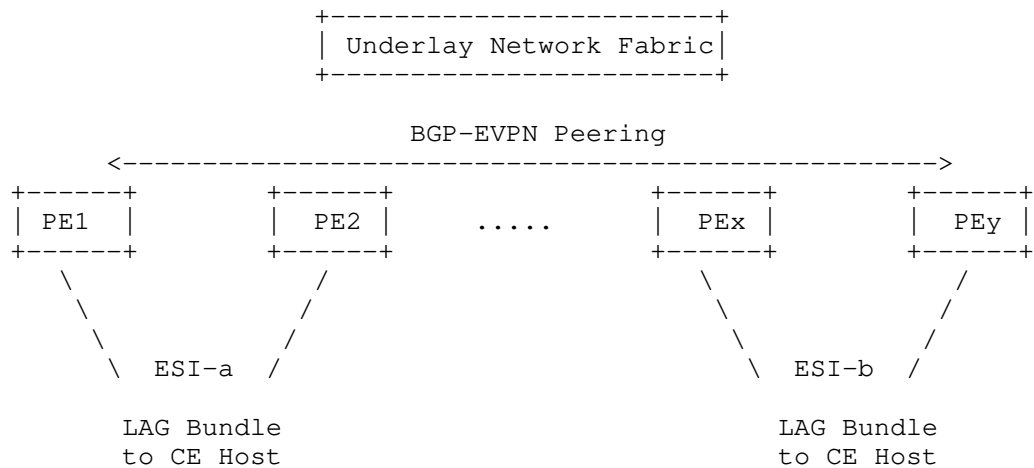


Figure 9

Data plane and ARP/ND snooping based MAC/IP learning on PE-CE all-active multi-homed LAG ports is subject to unpredictable hashing of ARP, ND, and data frames from host to PE. As an example, an ARP request for a connected host might originate at PE1 but the resulting ARP response from the host might be received at PE2. Redundant EVPN PEs in all-active multi-homing mode typically handle this unpredictability via combination of methods below:

- o PEs can handle unsolicited ARP and ND response frames.
- o PEs can implement additional mechanism to SYNC ARP, ND, and MAC tables across all PEs in a redundancy group for optimal forwarding to locally connected hosts.
- o PEs can implement EVPN aliasing procedures discussed in [RFC 7432] OR re-originate SYNCed MAC-IP adjacencies as local RT-2 to achieve MAC ECMP across the overlay.
- o PEs can also re-originate SYNCed MAC-IP adjacencies as local RT-2 to achieve IP ECMP across the overlay OR implement IP aliasing procedures discussed in [EVPN-IP-ALIASING].
- o PEs can also ensure EVPN sequence number SYNC for local MAC entries for EVPN mobility procedures to work correctly, as discussed in [EVPN-IRB-MOBILITY].

The PE-CE control plane learning alternative defined in this document fully decouples MAC and IP learning over MLAG ports from unpredictable hashing of data, AR, ND frames on all-active multi-

homed LAG member links. As a result, above procedures that essentially result from data-plane PE-CE learning on all-active multi-homed LAGs can be simplified via the PE-CE control plane alternative defined in this document.

#### 14.2.2 Convergence on CE Host Moves

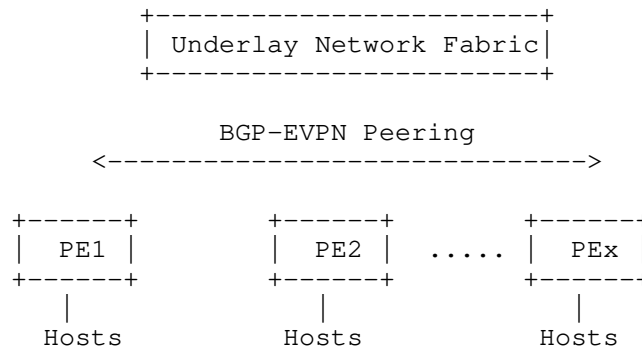


Figure 10

Host mobility across EVPN PE switches is a common occurrence in a data center fabric for flexibility in work load placement across a DC. Further, a host move must result in minimal, if any, disruption to traffic flows / services to / from the device.

Data plane and ARP/ND snooping based PE-CE learning may result in unpredictable convergence times, following host moves for the following cases:

- o A host may or may not send any data packet immediately following a move.
- o A host may or may not send an unsolicited ARP following a move.

While probing procedures, discussed in the next sub-sections are typically used to minimize convergence time, certain scenarios discussed below may still result in extended convergence times and flooding.

##### 14.2.2.1 Silent Hosts

If a host is silent for an extended period following a move from PE1 to PE2, any bridged traffic flow destined to this host will continue to be black-holed by PE1 until the MAC ages out at PE1. Once the the MAC ages out at PE1, any bridged traffic flow destined to the host is

flooded across the overlay bridge. Flooding of unknown unicast traffic on the overlay is enabled for this purpose. In summary, PE-CE learning that is based on data-plane and AR/ND snooping may be subject to non-deterministic convergence time and flooding following host moves because of being heavily dependent on unpredictable CE behavior.

PE-CE control plane based learning defined in this document fully decouples convergence in such scenarios from non-deterministic data flows and unsolicited ARP/ND behavior on a CE.

#### 14.2.2.2 Probing

ARP and ND probing procedures are typically used to achieve host re-learning and convergence following host moves across the overlay:

- o Following a host move from PE1 to PE2, the host's MAC is discovered at PE2 as a local MAC via a data frames received from the host. If PE2 has a prior REMOTE MAC-IP host route for this MAC from PE1, an ARP probe is typically triggered at PE2 to learn the MAC-IP as a local IP adjacency and triggers EVPN RT-2 advertisement for this MAC-IP across the overlay with new reachability via PE2.
- o Following a host move from PE1 to PE2, once PE1 receives a MAC or MAC-IP route from PE2 with a higher sequence number, an ARP probe is triggered at PE1 to clear the stale local MAC-IP neighbor adjacency OR re-learn the local MAC-IP in case the host has moved back or is duplicate.
- o Following a local MAC age-out, if there is a local IP adjacency with this MAC, an ARP probe is triggered for this IP to either re-learn the local MAC and maintain local l3 and l2 reachability to this host OR to clear the ARP entry in case the host is indeed no longer local. Note that clearing of stale ARP entries, following a move is required for traffic to converge in the event that the host was silent and not discovered at its new location. Once stale ARP entry for the host is cleared, routed traffic flow destined for the host can re-trigger ARP discovery for this host at the new location. ARP flooding on the overlay MUST also be done to enable ARP discovery via routed flows.
- o Alternatively, ARP probing timer may be tuned to be smaller than the MAC aging timer to avoid MAC age-out.

PE-CE control plane learning alternative defined in this document decouples host learning following moves from unpredictable host behavior with respect to sending data traffic and unsolicited ARPs,

and as a result from ARP probing and MAC aging timer settings. Host move handling is hence greatly simplified to a very predictable and deterministic behavior.

#### 14.2.3 ARP Gleaning Latency

If a CE's ARP binding is not already learned on a PE via an unsolicited ARP sent by the CE following events such as boot-up, flaps, and moves, a data frame that needs to be routed to the CE triggers ARP or ND discovery process on the PE. On a typical hardware switching platform, an IP packet that does not resolve to a link layer re-write would be punted to host stack that delivers packets with incomplete link-layer resolution to ARP or ND for resolution. An ARP request / ND Solicitation is generated for the CE IP and an ARP response or NA results in installing a link-layer re-write for the CE IP. In an EVPN multi-homing environment, this procedure is further complicated as the response is only received by one of the PEs that may or may not be the one that generated the ARP or ND request. Learned neighbor binding is SYNCed to other PEs that share the multi-homed Ethernet Segment. Routed flows can now be forwarded to the host via all PEs. Latency associated with such data frame driven ARP discovery may result in significant initial convergence hit, following triggers that warrant re-gleaning of CE IP to MAC binding.

PE-CE control plane learning alternative defined in this document results in proactive host learning following these scenarios, potentially avoiding a convergence hit on initial data packets.

#### 14.3 Applicability to non-EVPN Use Cases

While the L3DL based host learning procedure described in this document focuses on EVPN-IRB overlay fabric use case, it may also have benefits and applicability in non-EVPN use cases. Applicability of procedures described in this document to non-EVPN use cases is a topic for further study.

#### 15. Summary

PE-CE control plane is proposed as an alternative to data plane and ARP/ND snooping based PE-CE host MAC/IP learning and for PE-CE prefix learning. With a PE-CE control plane, CE host MAC and IP are deterministically learned on host boot-up, on host configuration, across host moves, on convergence triggers such as link failures, flaps, and PE re-boots and on all-active multi-homing LAG links. A PE-CE control plane decouples CE MAC and IP learning from traffic flows sourced by a CE, from varying CE behavior with respect to sending unsolicited ARP/ND frames, and from hashing of CE sourced frames over all-active multi-homed LAG links. As a result, it helps

achieve a predictable and reliable convergence behavior across these triggers and helps simplify certain EVPN procedures that are otherwise needed with a data-plane and ARP/ND snooping based PE-CE learning. In addition, it may also be used for non-host learning use cases such as prefix learning.

## 16. References

### 16.1 Normative References

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.
- [L3DL] Bush, R., Austein R., Patel, K., "Layer 3 Discovery and Liveness", Feb 2019, <<https://tools.ietf.org/html/draft-ietf-lsvr-l3dl-01>>.
- [EVPN-IRB] Sajassi, A., Salem, S., Thoria S., Drake J., Rabadan J., "Integrated Routing and Bridging in EVPN", July 2018, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-05>>.
- [EVPN-PREFIX-ADV] Rabadan J., Henderickx W., Drake J., Lin W., Sajassi, A., "IP Prefix Advertisement in EVPN", May 2018, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-11>>.
- [EVPN-IRB-MOBILITY] Malhotra, N., Sajassi, A., Rabadan, J., Drake J., Lingala A., Patekar A., "Extended Mobility Procedures for EVPN-IRB", June 2019, <<https://datatracker.ietf.org/doc/draft-ietf-bess-evpn-irb-extended-mobility>>.
- [EVPN-IP-ALIASING] Sajassi, A., Badoni, G., "L3 Aliasing and Mass Withdrawal Support for EVPN", July 2017, <<https://tools.ietf.org/html/draft-sajassi-bess-evpn-ip-aliasing-00>>.
- [RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", March 1997, <<https://tools.ietf.org/html/rfc2119>>.
- [RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", May 2017, <<https://tools.ietf.org/html/rfc8174>>.

### 15.2 Informative References

## 17. Acknowledgements

Authors would like to thank Randy Bush and Rob Austein for detailed review and feedback to ensure consistency with base L3DL protocol specification, as well as for helping build detailed L3DL flows included in this document.

Authors would like to thank Ali Sajassi and John Drake for detailed review and very valuable input on PE-CE protocol design for EVPN use cases as well as structuring this document for EVPN use cases.

## Contributors

Randy Bush  
Arrcus & IIJ  
5147 Crystal Springs  
Bainbridge Island, WA 98110  
United States of America  
Email: randy@psg.com

## Authors' Addresses

Neeraj Malhotra (Editor)  
Individual  
Email: neeraj.ietf@gmail.com

Keyur Patel  
Arrcus  
2077 Gateway Place, Suite #400  
San Jose, CA 95119, USA  
Email: keyur@arrcus.com

Jorge Rabadan  
Nokia  
777 E. Middlefield Road  
Mountain View, CA 94043, USA  
Email: jorge.rabadan@nokia.com



BESS WorkGroup  
Internet-Draft  
Intended status: Informational  
Expires: May 5, 2020

S. Mohanty  
M. Ghosh  
A. Sajassi  
Cisco Systems  
S. Breeze  
Claranet  
J. Uttaro  
ATT  
November 2, 2019

BGP EVPN Flood Traffic Optimization at EVPN Gateways  
draft-mohanty-bess-evpn-bum-opt-01

Abstract

In EVPN, the Broadcast, Unknown Unicast and Multicast (BUM) traffic is sent to all the routers participating in the EVPN instance. In a multi-homing scenario, when more than one PEs share the same Ethernet Segment, i.e. there are more than one PEs in a redundancy group, only the PE that is the Designated-Forwarder (DF) for the ES will forward that packet on the access interface whereas all non-DF PEs will drop the packet. In deployments such as EVPN Gateways (EVPN GW) or Data Center Interconnect (DCI) routers, this can be quite wasteful. This is especially true if there are significantly more EVPN GW or DCI PEs all participating in the same sets of ES and vES. This draft explores the problem and provides solutions for the same.

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 <https://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 May 5, 2020.

## Copyright Notice

Copyright (c) 2019 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 (<https://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. Requirements Language and Terminology . . . . .	2
2. Introduction . . . . .	3
3. Problem Description . . . . .	4
4. Solutions . . . . .	5
4.1. DF Election per-mcast-flow . . . . .	5
4.2. Suppress the advertisement of the IMET route . . . . .	5
4.3. Advertisement of the IMET route from the BDF . . . . .	7
5. Protocol Considerations . . . . .	7
6. Operational Considerations . . . . .	8
7. Security Considerations . . . . .	8
8. Acknowledgements . . . . .	8
9. Contributors . . . . .	8
10. References . . . . .	8
10.1. Normative References . . . . .	8
10.2. Informative References . . . . .	9
Authors' Addresses . . . . .	9

## 1. Requirements Language 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 [RFC2119].

- o ES: Ethernet Segment
- o vES: Virtual Ethernet Segment
- o EVI: Ethernet virtual Instance, this is a mac-vrf.
- o IMET: Inclusive Multicast Route

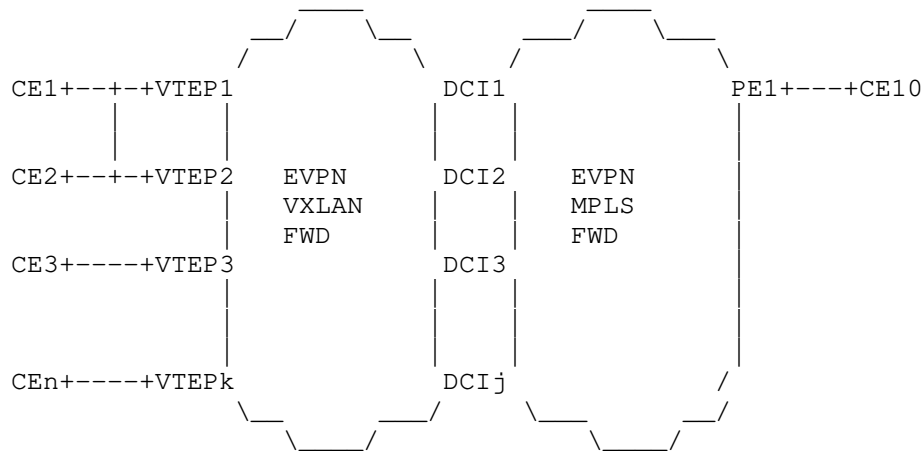
- o DF: Designated Forwarder
- o BDF: Backup Designated Forwarder
- o DCI: Data Center Interconnect Router

## 2. Introduction

EVPN [RFC7432] describes a solution for disseminating mac addresses over an mpls core via the Border Gateway Protocol. In EVPN, data plane learning is confined to the access, and the control plane flooding happens via BGP in the core. This prevents unnecessary flooding in the data plane as the traffic is directed to where the destination is learnt from. However, in case of Broadcast, Unknown Unicast and Multicast (BUM) traffic, the PE needs to do a flooding to all the other PEs in the domain.

PEs elect a Designated Forwarder (DF) amongst themselves, for a given ES, by exchanging type-4 routes via BGP. The role of a DF is to forward BUM traffic received from the core, towards its access facing interface. A PE in a non-DF role will drop flood traffic received on its core-facing interface. Note that the DF election process is only confined to the set of PEs who host the same Ethernet Segment. Remote PEs are not interested in type-4 routes for Ethernet Segments that they do not host. Hence remote PEs are ignorant of the DFs for segments which is not local to them. Consequently, when the remote PE needs to do a BUM flooding using ingress replication, it will flood the frames to all participating PEs, irrespective of whether DFs or not. The key to creating a list of PEs with which to flood to, is the Inclusive multicast ethernet tag route which is described below.

The IMET route (type-3) in EVPN advertises the BUM label for the EVI to all the other PEs who are interested in the same EVI. For ingress replication the label is encapsulated in the PMSI attribute. The label is used to encapsulate the BUM traffic at the ingress entity. This label is inserted just above the split-horizon label in the BUM frame. When the BUM packet is received by a PE that is multi-homed to the same Ethernet segment as the PE that originated the BUM packet, and, is the DF for that (EVI, ES) pair, after popping the transport label, the receiving PE is going to check if the split-horizon label is its own. If so, it will drop the packet if no other ES is configured. Otherwise it will forward the frame on all other Segments that are part of the same EVI. if the PE is not the DF, it will drop the packet immediately.



An EVPN Datacenter network with VXLAN forwarding joined to a traditional EVPN network with MPLS forwarding. Adjoining DCI routers are said to be EVPN GW's. A DCI will have a single vES (ESI) per BD, with multiple VTEP next-hops.

Figure 1

### 3. Problem Description

In the Figure 1. above, DCI1, DCI2 and DCI3 are all multi-homed EVPN GW's for multiple VTEPs serving the same vES, say vES1. PE1 has a single host which is not multi-homed.

The same EVPN instance (Bridge-Domain) exists on all the PEs and DCIs. For this EVPN instance, DCI1 is the Designated Forwarder on vES1 and DCI2 is the backup DF [RFC8584]. When PE1 sends the BUM traffic, the flooded frames are received by DCI1, DCI2, DCI3 up to DCIj. DCI1 is going to forward the flood traffic on its vES towards all VTEPs participating in vES1. DCI2, DCI3 and all DCIs up to DCIj will drop the flooded frames that they receive from the core.

Here it is wasteful for DCI2, DCI3 and DCIj to receive the flooded frames. Whilst the majority of deployments usually have two DCIs as part of the redundancy group, in some cases, there may be more than two on the same vES. An example being when capacity demands of the DCI are close to the hardware limits of the DCI. In this scenario, operators may chose to protect their investments and increase their resilience by installing additional DCIs, instead of replacing them or further segmenting the datacenter network. Further, increasing

the number of DCIs results in more efficient load-balancing across VNIs.

We can now formally describe the issue. In general, consider an EVPN instance, EVIi, that exists in a DCI, say DCIj. As per existing EVPN behavior, even if DCIj is not the DF for any of its virtual Ethernet Segments and also there are no other single-homed Ethernet Segments that are part of EVIi in DCIj, then DCIj will still receive BUM traffic meant for EVIi from a remote PE, PEk. This traffic is simply dropped as PEk is not a DF for any of these virtual Ethernet Segments.

1. This is an unnecessary usage of bandwidth in the EVPN Core.
2. DCIj receives traffic which it drops which is non-optimal usage of the L2 Forwarding engine.
3. PEk replicates a copy of the Ethernet Frame to DCIj which is only to be dropped. This consumes cycles at PEk.

In this draft we address the above problem and give possible solutions.

#### 4. Solutions

##### 4.1. DF Election per-mcast-flow

Solving the bandwidth in the EVPN core is an operators primary concern. Given the majority of traffic volume in BUM comes from large multicast flows, adopting the mechanisms described in :["I-D.draft-ietf-bess-evpn-per-mcast-flow-df-election-00"](#) not only improves the distribution of multicast traffic amongst DCI1...DCIj for a given vES, techniques such as not advertising the SMET from a non-DF DCI ensure that only DCIs who've won the election for the group, receive multicast traffic for the group.

This solution explicitly requires IGMP snooping in the BD where the vES resides.

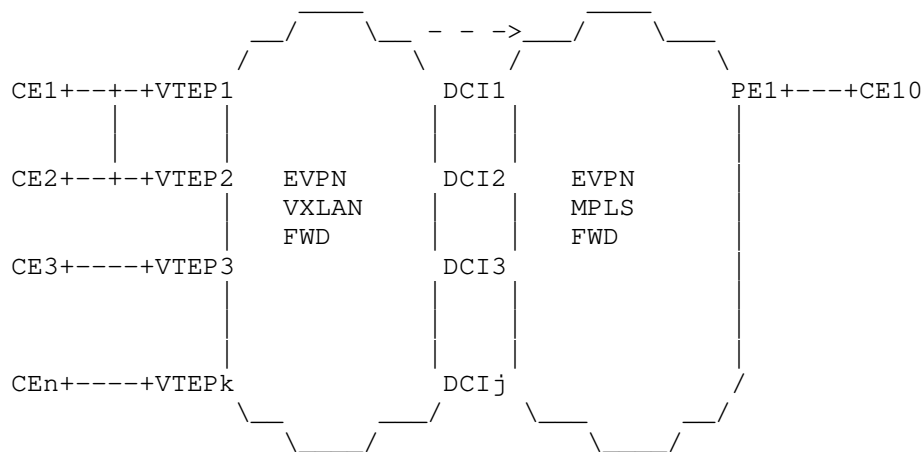
This solution does not solve the problem of unnecessary Broadcast and Unknown Unicast being replicated to nDFs, but it solves the most prominent problem of bandwidth.

##### 4.2. Suppress the advertisement of the IMET route

The next solution is for a DCI not to advertise the IMET route if the outcome is to drop the flooded traffic

- o DCIj only needs to advertise "Inclusive Multicast Ethernet Tag route" (Type-3 route) for an EVPN Instance, EVIi if and only if EVIi is configured on at least one Ethernet Segment (which also has a presence in another DCI, i.e Multihomed) and DCIj is the DF for that specific Ethernet Segment.
- o The Type-3 SHOULD also be advertised if there is a "Single-Home" Ethernet Segment on an EVI.
- o Where a DCI is the first DF for an vES on an EVPN Instance, the IMET should be advertised, whereas on the Last DF to Non-DF transition, it should be withdrawn.

In the Figure 2 the same EVPN instance exists in DCI1, DCI2, DCI3, DCIj and PE1. However, only DCI1 and PE1 advertise the IMET route. So PE1 sends the flood traffic to DCI1 only.



An EVPN GW Network

Figure 2

With this approach, on a DF DCI1 failure, BUM traffic will be dropped until the IMET from the next elected DF [DCI2 through DCIj] is received at PE1. Note however; present behaviour is that BUM is also dropped based on route type 4 withdraw in the peering PEs. In comparison of this proposal with the existing methods, convergence delay will be MAX[Type 4, Type 3 Propagation delays] after the New DF is elected. This leads to our next solution extension, where convergence cannot be traded off over bandwidth optimization.

#### 4.3. Advertisement of the IMET route from the BDF

1. Multihomed PEs can easily compute the Backup DF, based on the DF election mode in operation.
2. Extending the previous solution, we are proposing that a PE should only advertise Type-3 for an EVI if and only if one of the conditions hold:
  - \* It has an Single Home Ethernet Segment, in the EVI
  - \* It is DF for at least one ES or vES, for that EVI
  - \* It is BDF for at least one ES or vES, for that EVI

This would mean that, in Fig. 2, in addition to the IMET routes that are being advertised from DCI1, DCI2 also advertises the IMET route since it is the BDF. It can be seen from the above example that with increasing number of multi-homed PEs sharing the same vESs, only two DCIs will advertise IMET on behalf of an EVI. Of course, if there are some single-homed hosts, there may be some additional IMET advertisements. But the real benefits are in the data plane since this results in no BUM traffic for DCIs that do not need it; but would have, nevertheless, got it, as per the existing EVPN procedures.

It is important to note that the solutions involving suppression of IMET should be limited to the following use case caveats;

1. BUM traffic for Ingress Replication (IR) cases
2. BDs with no igmp/mld/pim proxy
3. BDs with no OISM or IRBs
4. BDs with vES associated to overlay tunnels and no other ACs

With these caveats, the suppression of IMET at non DF or BDF EVPN GWs provide complete control over BUM traffic distribution per-vES (per-BD).

#### 5. Protocol Considerations

This idea conforms to existing EVPN drafts that deal with BUM handling [RFC7432], and [I-D.ietf-bess-evpn-igmp-mld-proxy]. Additionally, to take DF Type 4 as explained in : "I-D.draft-ietf-bess-evpn-per-mcast-flow-df-election" into consideration, along the other conditions specified in Sections 4 and 5, the PE should

advertise IMET if and only if there is at least one (S,G) for which it is DF. For all other DF Types, no additional considerations are required.

## 6. Operational Considerations

None

## 7. Security Considerations

This document raises no new security issues for EVPN.

## 8. Acknowledgements

The authors would like to thank Jorge Rabadan, John Drake and Eric Rosen for discussions related to this draft.

## 9. Contributors

Samir Thoria  
Cisco Systems  
US

Email: sthoria@cisco.com

Sameer Gulrajani  
Cisco Systems  
US

Email: sameerg@cisco.com

## 10. References

### 10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.



- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, R., Sajassi, N., Drake, A., Nagaraj, K., and S. Sathappan, "BGP MPLS-Based Ethernet VPN", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

## 10.2. Informative References

- [I-D.ietf-bess-evpn-igmp-mld-proxy]  
Sajassi, A., Thoria, S., Patel, K., Yeung, D., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-04 (work in progress), September 2019.
- [I-D.ietf-bess-evpn-per-mcast-flow-df-election]  
Sajassi, A., mishra, m., Thoria, S., Rabadan, J., and J. Drake, "Per multicast flow Designated Forwarder Election for EVPN", draft-ietf-bess-evpn-per-mcast-flow-df-election-01 (work in progress), March 2019.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.

## Authors' Addresses

Satya Ranjan Mohanty  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

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

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

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

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

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

Sandy Breeze  
Claranet  
21 Southampton Row  
London WC1B 5HA  
United Kingdom

Email: [sandy.breeze@eu.clara.net](mailto:sandy.breeze@eu.clara.net)

Jim Uttaro  
ATT  
200 S. Laurel Avenue  
Middletown, CA 07748  
USA

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

BESS Workgroup  
Internet Draft  
Intended status: Standards Track

J. Rabadan, Ed.  
K. Nagaraj  
Nokia

W. Lin  
Juniper

A. Sajassi  
Cisco

Expires: May 4, 2020

November 1, 2019

EVPN Multi-Homing Extensions for Split Horizon Filtering  
draft-nr-bess-evpn-mh-split-horizon-02

Abstract

Ethernet Virtual Private Network (EVPN) is commonly used along with Network Virtualization Overlay (NVO) tunnels. The EVPN multi-homing procedures may be different depending on the NVO tunnel type used in the EVPN Broadcast Domain. In particular, there are two multi-homing Split Horizon procedures to avoid looped frames on the multi-homed CE: ESI Label based and Local Bias. ESI Label based Split Horizon is used for MPLSoX tunnels, E.g., MPLSoUDP, whereas Local Bias is used for others, E.g., VXLAN tunnels. The current specifications do not allow the operator to decide which Split Horizon procedure to use for tunnel encapsulations that could support both. Examples of tunnels that may support both procedures are MPLSoGRE, MPLSoUDP, GENEVE or SRv6. This document extends the EVPN Multi-Homing procedures so that an operator can decide the Split Horizon procedure for a given NVO tunnel depending on their own requirements.

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 May 4, 2020.

#### Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

1. Introduction . . . . .	3
1.1 Conventions and Terminology . . . . .	5
2. BGP EVPN Extensions . . . . .	7
2.1 The Split Horizon Type (SHT) . . . . .	8
2.2 Use of the Split Horizon Type In A-D Per ES Routes . . . . .	8
2.3 ESI Label Value In A-D Per ES Routes . . . . .	9
2.4 Backwards Compatibility With [RFC8365] NVEs . . . . .	10
3. Procedures for NVEs Supporting Multiple Encapsulations . . . . .	11
7. IANA Considerations . . . . .	12
8. References . . . . .	12
8.1. Normative References . . . . .	12
8.2. Informative References . . . . .	13
9. Acknowledgments . . . . .	14
10. Contributors . . . . .	14
Authors' Addresses . . . . .	14

## 1. Introduction

Ethernet Virtual Private Network (EVPN) is commonly used along with Network Virtualization Overlay (NVO) tunnels and specified in [RFC8365]. The EVPN multi-homing procedures may be different depending on the NVO tunnel type used in the EVPN Broadcast Domain. In particular, there are two Multi-Homing Split Horizon procedures to avoid looped frames on the multi-homed CE: ESI Label based and Local Bias. ESI Label based Split Horizon is used for MPLSoX tunnels, E.g., MPLSoUDP [RFC7510], and its procedures described in [RFC7432]. Local Bias is used by non-MPLS NVO tunnels, E.g., VXLAN tunnels, and it is described in [RFC8365].

As a refresher:

- o ESI Label based Split-Horizon filtering [RFC7432]

If MPLS-based tunnels are used in EVPN, an MPLS label is used for Split Horizon filtering to support All-Active multi-homing where an ingress NVE adds a label corresponding to the source Ethernet Segment (aka an ESI label) when encapsulating the packet. The egress NVE checks the ESI label when attempting to forward a multi-destination frame out a local ES interface, and if the label corresponds to the same site identifier (ESI) associated with that ES interface, the packet is not forwarded. This prevents the occurrence of forwarding loops for BUM traffic.

The ESI Label Split Horizon filtering SHOULD also be used with Single-Active multi-homing to avoid transient loops for in-flight packets when the egress NVE takes over as DF for an Ethernet Segment.

- o Local Bias for non-MPLS NVO tunnels [RFC8365]

Since non-MPLS NVO tunnels, such as VXLAN and NVGRE encapsulations, do not include the ESI label, a different Split Horizon filtering procedure must be used for All-Active multi-homing. This mechanism is called Local Bias and relies on the NVO tunnel source IP address to decide whether to forward BUM traffic to a local ES interface at the egress NVE.

In a nutshell, every NVE tracks the IP address(es) associated with the other NVE(s) with which it has shared multi-homed ESs. When the egress NVE receives a BUM frame encapsulated in a VXLAN or NVGRE packet, it examines the source IP address in the tunnel header (which identifies the ingress NVE) and filters out the frame on all local interfaces connected to ESes that are shared with the ingress NVE.

Due to this behavior at the egress NVE, the ingress NVE's behavior is also changed to perform replication locally to all directly attached Ethernet segments (regardless of the DF election state) for all BUM ingress from the access ACs. Because of this "local" replication at the ingress NVE, this approach is referred to as Local Bias.

Local Bias cannot be used for Single-Active multi-homing, since the ingress NVE brings operationally down the ACs for which it is non-DF (hence local replication to non-DF ACs cannot be done). This means transient in-flight BUM packets may be looped back to the originating site by new elected DF egress NVEs.

[RFC8365] states that Local Bias is used only for non-MPLS NVO tunnels, and ESI Label based Split Horizon for MPLS NVO tunnels. However, MPLS NVO tunnels, such as MPLSoGRE or MPLSoUDP, can potentially support both procedures, since they can carry ESI Labels and they also use a tunnel IP header where the source IP address identifies the ingress NVE.

Similarly, some non-MPLS NVO tunnels may potentially follow either procedure too. Some examples are GENEVE or SRv6:

- o In a GENEVE tunnel the source IP address identifies the ingress NVE therefore local bias is possible. Also, [EVPN-GENEVE] defines an Ethernet option TLV (Type Length Value) to encode an ESI label value.
- o In an SRv6 tunnel, the source IP address also identifies the ingress NVE, however, by default and as described in [SRv6-Services] the ingress PE will add information in the SRv6 packet so that the egress PE can identify the source ES of the BUM packet. That information is the ESI filtering argument of the service SID received on an A-D per ES route from the egress PE.

Table 1 shows different tunnel encapsulations and their supported and default Split Horizon method. In the case of GENEVE, the default Split Horizon Type (SHT) depends on whether the Ethernet Option with Source ID TLV is negotiated. In the case of SRv6, the default SHT is listed as ESI label filtering in Table 1, since the behavior is equivalent to that of ESI Label filtering

Tunnel Encapsulation	Default Split Horizon Type (SHT)	Supports Local Bias	Supports ESI Label
VXLAN	Local Bias	Yes	No
NVGRE	Local Bias	Yes	No
MPLS	ESI Label filtering	No	Yes
MPLSoGRE	ESI Label filtering	Yes	Yes
MPLSoUDP	ESI Label filtering	Yes	Yes
GENEVE	Local Bias (no ESI Lb) ESI Label (if ESI Lb)	Yes	Yes
SRv6	ESI Label filtering	Yes	Yes

Table 1 - Tunnel Encapsulations and Split Horizon Types

The ESI Label method works for All-Active and Single-Active, while Local Bias only works for All-Active. In addition, the ESI Label method works across different networks, whereas Local Bias is limited to networks with no next hop change between the NVEs in the same Ethernet Segment. However, some operators prefer the Local Bias method, since it simplifies the encapsulation, consumes less resources on the NVEs and the ingress NVE always forwards locally to other interfaces.

This document extends the EVPN Multi-Homing procedures so that an operator can decide the Split Horizon procedure for a given NVO tunnel depending on their own specific requirements. The choice of Local Bias or ESI Label Split Horizon is now allowed for NVO tunnels that support both methods. Non-MPLS NVO tunnels that do not support both methods, E.g., VXLAN or NVGRE, will follow [RFC8365] procedures.

### 1.1 Conventions and Terminology

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

14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

- o BUM: Broadcast, Unknown unicast and Multicast traffic.
- o ES and ESI: Ethernet Segment and Ethernet Segment Identifier.
- o A-D per ES route: refers to the EVPN Ethernet Auto-Discovery per Ethernet Segment route defined in [RFC7432].
- o AC: Attachment Circuit.
- o NVE: Network Virtualization Edge device.
- o EVI and EVI-RT: EVPN Instance and EVI Route Target. A group of NVEs attached to the same EVI will share the same EVI-RT.
- o MPLS and non-MPLS NVO tunnels: refer to Multi-Protocol Label Switching (or the absence of it) Network Virtualization Overlay tunnels. Network Virtualization Overlay tunnels use an IP encapsulation for overlay frames, where the source IP address identifies the ingress NVE and the destination IP address the egress NVE.
- o MPLSoUDP: Multi-Protocol Label Switching over User Datagram Protocol, [RFC7510]
- o MPLSoGRE: Multi-Protocol Label Switching over Generic Network Encapsulation, [RFC4023].
- o MPLSoX: refers to MPLS over any IP encapsulation. Examples are MPLSoUDP or MPLSoGRE.
- o GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].
- o VXLAN: Virtual eXtensible Local Area Network, [RFC7348].
- o NVGRE: Network Virtualization Using Generic Routing Encapsulation, [RFC7637].
- o VNI: Virtual Network Identifier. A 24-bit identifier used by Network Virtualization Overlay (NVO) over IP encapsulations. Examples are VXLAN (Virtual Extended Local Area Network) or GENEVE (Generic Network Virtualization Encapsulation).
- o Broadcast Domain (BD): an emulated ethernet, such that two systems on the same BD will receive each other's link-local broadcasts. In this document, BD also refers to the instantiation of a Broadcast



Domain on an EVPN PE. An EVPN PE can be attached to one or multiple BDs of the same tenant.

- o Designated Forwarder (DF): as defined in [RFC7432], an ethernet segment may be multi-homed (attached to more than one PE). An ethernet segment may also contain multiple BDs, of one or more EVIs. For each such EVI, one of the PEs attached to the segment becomes that EVI's DF for that segment. Since a BD may belong to only one EVI, we can speak unambiguously of the BD's DF for a given segment.
- o SHT: Split Horizon Type, it refers to the Split Horizon method that a PE intends to use and advertises in an A-D per ES route.

This document also assumes familiarity with the terminology of [RFC7432] and [RFC8365].

## 2. BGP EVPN Extensions

EVPN extensions are needed so that NVEs can advertise their preference for the Split Horizon method to be used in the Ethernet Segment. Figure 1 shows the ESI Label extended community that is always advertised along with the EVPN A-D per ES route. All the NVEs attached to an Ethernet Segment advertise an A-D per ES route for the ES, including this extended community that conveys the information for the multi-homing mode (All-active or Single-Active), as well as the ESI Label to be used (if needed).

```

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(1 octet) | Reserved=0    |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Reserved=0     |               | ESI Label      |               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Figure 1 - ESI Label extended community

[RFC7432] defines the low-order bit of the Flags octet (bit 0) as the "Single-Active" bit:

- o A value of 0 means that the multi-homed Ethernet Segment is operating in All-Active mode.
- o A value of 1 means that the multi-homed Ethernet Segment is operating in Single-Active mode.

## 2.1 The Split Horizon Type (SHT)

[RFC8365] does not add any explicit indication about the Split Horizon method in the A-D per ES route. In this document the [RFC8365] Split Horizon procedure is the default behavior and assumes that Local Bias is used only for non-MPLS NVO tunnels, and ESI Label based Split Horizon for MPLS NVO tunnels. This document defines the two high-order bits in the Flags octet (bits 6 and 7) as the "Split Horizon Type" (SHT) field, where:

SHT bit 7 6

-----

0 0	--> Default SHT. Backwards compatible with [RFC8365]
0 1	--> Local Bias
1 0	--> ESI Label based filtering
1 1	--> reserved for future use

- o SHT = 00 is backwards compatible with [RFC8365] and indicates:
  - The advertising NVE intends to use the default or native SHT. The default SHT is shown in Table 1 for each NVO encapsulation.
  - An egress NVE that follows the [RFC8365] behavior and does not support this specification will use an SHT value of 00.
- o SHT = 01 indicates that the advertising NVE intends to use Local Bias procedures in the Ethernet Segment for which the AD per-ES route is advertised.
- o SHT = 10 indicates that the advertising NVE intends to use the ESI Label based Split Horizon method procedures in the Ethernet Segment for which the AD per-ES route is advertised.

## 2.2 Use of the Split Horizon Type In A-D Per ES Routes

The following must be observed:

- An SHT value of 01 or 10 MUST NOT be used with encapsulations that support only one SHT in Table 1, and MAY be used by encapsulations that support the two SHTs in Table 1.
- An SHT value different than 00 expresses the intend to use a specific Split Horizon method, but does not reflect the actual operational SHT used by the advertising NVE, unless all the NVEs attached to the ES advertise the same SHT.
- In case of inconsistency in the SHT value advertised by the NVEs attached to the same ES for a given EVI, all the NVEs MUST revert

to the [RFC8365] behavior, and use the default SHT in Table 1, irrespective of the advertised SHT.

- An SHT different from 00 MUST NOT be set if the Single-Active bit is set. A received A-D per ES route where Single-Active and SHT bits are different from zero MUST be treat-as-withdraw [RFC7606].
- The SHT MUST have the same value in each Ethernet A-D per ES route that an NVE advertises for a given ES and a given encapsulation (see Section 3 for NVEs supporting multiple encapsulations).

As an example, egress NVEs that support MPLS NVO tunnels, E.g., MPLSoGRE or MPLSoUDP, will advertise A-D per ES route(s) for the ES along with the [RFC5512] BGP Encapsulation extended community indicating the encapsulation (MPLSoGRE or MPLSoUDP) and MAY use the SHT = 01 or 10 to indicate the intend to use Local Bias or ESI Label, respectively.

An egress NVE MUST NOT use an SHT value different from 00 when advertising an A-D per ES route with encapsulation VXLAN, NVGRE, MPLS or no [RFC5512] BGP tunnel encapsulation extended community. We assume that, in all these cases, there is no Split Horizon method choice, and therefore the SHT value must be 00. A received route with one of the above encapsulation options and SHT value different from 00 SHOULD be treat-as-withdraw.

An egress NVE advertising A-D per ES route(s) for an ES with encapsulation GENEVE MAY use an SHT value of 01 or 10. A value of 01 indicates the intend to use Local Bias, irrespective of the presence of an Ethernet option TLV with a non-zero Source-ID [EVPN-GENEVE]. A value of 10 indicates the intend to use ESI Label based Split Horizon. A value of 00 indicates the default behavior in Table 1, that is, use Local Bias if no ESI-Label exists in the Ethernet option TLV or no Ethernet option TLV whatsoever. Otherwise the ESI Label Split Horizon method is used.

The above procedures assume a single encapsulation supported in the egress NVE. Section 3 describes additional procedures for NVEs supporting multiple encapsulations.

### 2.3 ESI Label Value In A-D Per ES Routes

This document also modifies the value that is advertised in the ESI Label field of the ESI Label extended community as follows:

- o The A-D per ES route(s) for an ES MAY have an ESI Label value of zero if the SHT value is 01. Section 2.2 specifies the cases where

the SHT can be 01. An ESI Label value of zero avoids the allocation of Labels in the cases where they are not used (Local Bias).

- o The A-D per ES route(s) for an ES MAY have an ESI Label value of zero for VXLAN or NVGRE encapsulations.

## 2.4 Backwards Compatibility With [RFC8365] NVEs

As discussed in Section 2.2 this specification is backwards compatible with the Split Horizon filtering behavior in [RFC8365] and a non-upgraded NVE can be attached to the same ES as other NVEs supporting this specification.

An NVE has an administrative SHT value for an ES (the one that is advertised along with the A-D per ES route) and an operational SHT value (the one that is actually used irrespective of what the NVE advertised). The administrative SHT matches the operational SHT if all the NVEs attached to the ES have the same administrative SHT.

This document assumes that an [RFC7432] or [RFC8365] compatible implementation (that does not support this document) ignores the value or all the bits in the ESI Label extended community except for the Single-Active bit. Based on this assumption, a non-upgraded NVE will ignore an SHT different from 00. As soon as an upgraded NVE receives at least one A-D per ES route for the ES with SHT value of 00, it MUST revert its operational SHT to the default Split Horizon method, as in Table 1, and irrespective of its administrative SHT.

As an example, consider an NVE attached to Ethernet Segment N that receives two A-D per ES routes for N from different NVEs, NVE1 and NVE2. If the route from NVE1 has SHT = 00 and the one from NVE2 an SHT = 01, the NVE MUST use the default Split Horizon method in Table 1 as operational SHT, irrespective of its administrative SHT.

All the NVEs attached to an ES with operational SHT value of 10 MUST advertise a valid non-zero ESI Label. If the operational SHT value is 01, the ESI Label MAY be zero. If the operational SHT value is 00, the ESI Label MAY be zero only if the default encapsulation supports Local Bias only and the NVEs do not check the presence of a valid non-zero ESI Label.

If an NVE changes its operational SHT value from 01 to 00 (as a result of a new non-upgraded NVE present in the ES) and it previously advertised a zero ESI Label, it MUST send an update with a non-zero valid ESI Label, unless all the non-upgraded NVEs in the ES support Local Bias only.

### 3. Procedures for NVEs Supporting Multiple Encapsulations

As specified by [RFC8365], an egress NVE that supports multiple data plane encapsulations (I.e., VXLAN, NVGRE, MPLS, MPLSoUDP, GENEVE) needs to indicate all the supported encapsulations using BGP Encapsulation extended communities defined in [RFC5512] with all EVPN routes. This section clarifies the multi-homing Split Horizon behavior for NVEs advertising and receiving multiple BGP Encapsulation extended communities along with the A-D per ES routes. This section uses a notation of {x,y} to indicate the encapsulations advertised in [RFC5512] BGP Encapsulation extended communities, with x and y being different encapsulation values.

It is important to remember that an NVE MAY advertise multiple A-D per ES routes for the same ES (and not only one), each route conveying a number of EVI Route Targets (EVI-RTs). We refer to the total number of EVI-RTs in a given ES as EVI-RT-set for that ES. Any of the EVIs represented in the EVI-RT-set will have its EVI-RT included in one (and only one) A-D per ES route for the ES. When multiple A-D per ES routes are advertised for the same ES, each route MUST have a different Route Distinguisher.

As per [RFC8365], an NVE that advertises multiple encapsulations in the A-D per ES route(s) for an ES, MUST advertise encapsulations that use the same Split Horizon filtering method in the same route. For example:

- o An A-D per ES route for ES-x may be advertised with {VXLAN,NVGRE} encapsulations.
- o An A-D per ES route for ES-y may be advertised with {MPLS,MPLSoUDP,MPLSoGRE} encapsulations (or a subset).
- o But an A-D per ES route for ES-z MUST NOT be advertised with {MPLS,VXLAN} encapsulations.

This document extends this behavior as follows:

- (a) An A-D per ES route for ES-x may be advertised with multiple encapsulations where some support a single Split Horizon method. In this case, the SHT value MUST be 00. As an example, {VXLAN,NVGRE}, {VXLAN,GENEVE} or {MPLS,MPLSoGRE,MPLSoUDP} can be advertised in an A-D per ES route. In all those cases SHT MUST be 00.
- (b) An A-D per ES route for ES-y may be advertised with multiple encapsulations where all of them support both Split Horizon methods. In this case the SHT value MAY be 01 if the desired

method is Local Bias, or 10 if ESI Label based. For example, {MPLSoGRE,MPLSoUDP,GENEVE} (or a subset) may be advertised in an A-D per ES route with SHT value of 01. The ESI Label value in this case MAY be zero.

- (c) If ES-z with EVI-RT-set composed of (EVI-RT1,EVI-RT2,EVI-RT3..EVI-RTn) supports multiple encapsulations that require a different Split Horizon method, a different A-D per ES route (or group of routes) per Split Horizon method MUST be advertised. For example, consider n EVIs in ES-z and:

- the EVIs corresponding to (EVI-RT1..EVI-RTi) support VXLAN,
- the ones for (EVI-RTi+1..EVI-RTm) (with i<m) support MPLSoUDP with Local Bias,
- and the ones for (EVI-RTm+1..EVI-RTn) (with m<n) support GENEVE with ESI Label based Split Horizon.

In this case, three groups of A-D per ES routes MUST be advertised for ES-z:

- A-D per ES route group 1, including (EVI-RT1..EVI-RTi), with encapsulation {VXLAN}, SHT = 00. The ESI Label MAY be zero.
- A-D per ES route group 2, including (EVI-RTi+1..EVI-RTm), with encapsulation {MPLSoUDP}, SHT = 01. The ESI Label MAY be zero.
- A-D per ES route group 3, including (EVI-RTm+1..EVI-RTn), with encapsulation {GENEVE}, SHT = 10. The ESI Label MUST have a valid value, different from zero, and the Ethernet option [EVPN-GENEVE] MUST be advertised.

As per [RFC8365], it is the responsibility of the operator of a given EVI to ensure that all of the NVEs in that EVI support a common encapsulation. If this condition is violated, it could result in service disruption or failure.

## 7. IANA Considerations

IANA is requested to allocate the SHT bits (6 and 7) in the Flags Octet of the EVPN ESI Label extended community. This field is called "Split Horizon Type" bits.

## 8. References

### 8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.

## 8.2. Informative References

[RFC8584] Rabadan-Mohanty et al., "Framework for EVPN Designated Forwarder Election Extensibility", <<https://rfc-editor.org/rfc/rfc8584.txt>>, April 2019.

[EVPN-GENEVE] Boutros, S., Sajassi, A., Drake, J., and J. Rabadan, "EVPN control plane for Geneve", Work in Progress, draft-ietf-bess-evpn-geneve-00, August 2019.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <<https://www.rfc-editor.org/info/rfc7348>>.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <<https://www.rfc-editor.org/info/rfc5512>>.

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005, <<https://www.rfc-editor.org/info/rfc4023>>.

[RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation", RFC 7637, DOI 10.17487/RFC7637, September 2015, <<https://www.rfc->

[editor.org/info/rfc7637](https://www.rfc-editor.org/info/rfc7637)>.

[RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", RFC 7510, DOI 10.17487/RFC7510, April 2015, <<https://www.rfc-editor.org/info/rfc7510>>.

[GENEVE] Gross, J., Ed., Ganga, I., Ed., and T. Sridhar, Ed., "Geneve: Generic Network Virtualization Encapsulation", Work in Progress, draft-ietf-nvo3-geneve-14, September 2019.

[TUNNEL-ENCAP] Rosen, E., Ed., Patel, K., and G. Vesde, "The BGP Tunnel Encapsulation Attribute", Work in Progress draft-ietf-idr-tunnel-encaps-14, September 2019.

[RFC7606] Chen, E., Scudder, J., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, August 2015, <<http://www.rfc-editor.org/info/rfc7606>>.

[SRv6-Services] Dawra, G. et al., "SRv6 BGP based Overlay services", Work in Progress, draft-ietf-bess-srv6-services-00, October 2019.

## 9. Acknowledgments

## 10. Contributors

### Authors' Addresses

Jorge Rabadan (Editor)  
Nokia  
777 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

Kiran Nagaraj  
Nokia  
701 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: [kiran.nagaraj@nokia.com](mailto:kiran.nagaraj@nokia.com)

Wen Lin



Juniper Networks  
Email: wlin@juniper.net

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

SPRING WG  
Internet-Draft  
Intended status: Standards Track  
Expires: April 26, 2020

Shaofu. Peng  
Zheng. Zhang  
Greg. Mirsky  
ZTE Corporation  
October 24, 2019

SRv6 and MPLS interworking for VPN service  
draft-pzm-bess-spring-interdomain-vpn-00

Abstract

This document describes a method to achieve an inter-domain connection for a VPN (Virtual Private Network) service.

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 <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on April 26, 2020.

Copyright Notice

Copyright (c) 2019 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 (<https://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. Specification . . . . .	2
2.1. SRv6 to SR-MPLS domain signaling . . . . .	2
2.2. SR-MPLS to SRv6 domain signaling . . . . .	3
3. IANA Considerations . . . . .	4
4. Security Considerations . . . . .	4
5. References . . . . .	4
5.1. Normative References . . . . .	4
5.2. Informative References . . . . .	5
Authors' Addresses . . . . .	5

## 1. Introduction

[I-D.agrawal-spring-srv6-mpls-interworking] describes SRv6 and MPLS/SR-MPLS interworking and co-existence procedures. The document leverages the function defined in [I-D.ietf-spring-srv6-network-programming] to give guidance to the forwarding in routers.

[RFC4364] describes a method by which a Service Provider may use an IP backbone to provide IP Virtual Private Networks (VPNs) for its customers. When SRv6 and SR-MPLS are co-existed in the backbone, controller or a control plane, for example, using BGP, should be used to instantiate the VPN service as described in [I-D.agrawal-spring-srv6-mpls-interworking].

In case of option B inter-domain interconnection [RFC4364], only ASBR needs to do the stitching work between two ASes. Thus PEs in SRv6 and SR-MPLS domains do not have to support both SRv6 and SR-MPLS functions. This document discusses the use of BGP for achieving VPN service through option B defined in [RFC4364] across a backbone that includes SRv6 and SR-MPLS domains.

## 2. Specification

## 2.1. SRv6 to SR-MPLS domain signaling

[I-D.ietf-bess-srv6-services] defines the new TLVs for the BGP Prefix-SID Attribute that can be used to signaling of SRv6 SID for L3 and L2 services. In this document, we use L3 case as the example, the procedures for L2 are the same as in L3 scenario.

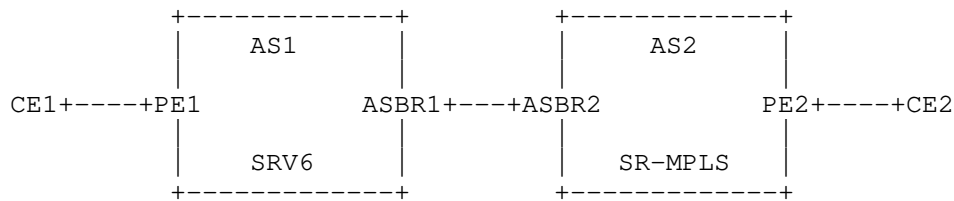


Figure 1

For example, CE1 and CE2 are connected through a backbone that includes AS1 and AS2. AS1 supports SRv6 only, and AS2 supports SR-MPLS only. ASBR1 supports both SRv6 and SR-MPLS capabilities, but ASBR2 supports SR-MPLS capability only.

For a prefix advertised by CE1 to PE1, PE1 assigns SID with End.DT4 (or End.DT6) defined in [I-D.ietf-spring-srv6-network-programming] section 4 (e.g., End.DT4 is used while the prefix is IPv4 prefix, End.DT6 is used while the prefix is IPv6 prefix), and advertises it to ASBR1. Because ASBR2 supports SR-MPLS function only, the SRv6 SID advertised by ASBR1 cannot be executed by ASBR2 because ASBR2 cannot recognize it.

ASBR1 uses specific execution function that is different from the function used in a single SRv6 domain or a single SR-MPLS domain. In this situation, ASBR1 assigns an MPLS label for the prefix received from PE1 and advertises it to ASBR2. The MPLS label has local significance that indicates this packet is associated with an SRv6 SID list which leads the packet from ASBR1 to PE1. The advertisement is the same as the format in [I-D.ietf-idr-bgp-prefix-sid].

When a data flow packet which has the destination to CE1 is received by ASBR1, ASBR1 recognizes the MPLS label, removes the label and adds an SRH to the packet, then forwards it to PE1.

## 2.2. SR-MPLS to SRv6 domain signaling

In the same example, PE2 advertises a prefix received from CE2 with assigned SID to ASBR2 according to [I-D.ietf-idr-bgp-prefix-sid], ASBR2 assigns SID for this prefix and advertises it to ASBR1. When ASBR1 advertises this prefix to PE1, ASBR1 should assign an SRv6 SID for it. The SID indicates the new execution function (e.g., END.RM, it indicates that MPLS should replace the SRH) for exchanging the packet header from SRH to MPLS list. The new function format is like the definition in [I-D.ietf-spring-srv6-network-programming] section 4.

When a data flow packet, which has the destination to CE2, is received by ASBR1, ASBR1 recognizes the SRv6 SID, removes the SRH and adds a or a list of MPLS label in the packet, and forwards it to PE2.

### 3. IANA Considerations

There is no IANA consideration.

### 4. Security Considerations

This document introduces no new security consideration beyond those already specified in [RFC4364], [I-D.ietf-idr-bgp-prefix-sid], [I-D.ietf-spring-srv6-network-programming], [I-D.ietf-bess-srv6-services] and [I-D.agrawal-spring-srv6-mpls-interworking].

### 5. References

#### 5.1. Normative References

[I-D.ietf-bess-srv6-services]

Dawra, G., Filsfils, C., Brissette, P., Agrawal, S., Leddy, J., daniel.voyer@bell.ca, d., daniel.bernier@bell.ca, d., Steinberg, D., Raszuk, R., Decraene, B., Matsushima, S., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay services", draft-ietf-bess-srv6-services-00 (work in progress), October 2019.

[I-D.ietf-idr-bgp-prefix-sid]

Previdi, S., Filsfils, C., Lindem, A., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix SID extensions for BGP", draft-ietf-idr-bgp-prefix-sid-27 (work in progress), June 2018.

[I-D.ietf-spring-srv6-network-programming]

Filsfils, C., Camarillo, P., Leddy, J., daniel.voyer@bell.ca, d., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-ietf-spring-srv6-network-programming-04 (work in progress), October 2019.

[RFC4364]

Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.

## 5.2. Informative References

[I-D.agrawal-spring-srv6-mpls-interworking]  
Agrawal, S., Ali, Z., Filsfils, C., daniel.voyer@bell.ca,  
d., and Z. Li, "SRv6 and MPLS interworking", draft-  
agrawal-spring-srv6-mpls-interworking-01 (work in  
progress), July 2019.

## Authors' Addresses

Shaofu Peng  
ZTE Corporation

Email: peng.shaofu@zte.com.cn

Zheng Zhang  
ZTE Corporation

Email: zzhang\_ietf@hotmail.com

Greg Mirsky  
ZTE Corporation

Email: gregimirsky@gmail.com

BESS  
Internet-Draft  
Intended status: Standards Track  
Expires: May 4, 2020

W. Lin, Ed.  
S. Sivaraj  
V. Garg  
Juniper Networks, Inc.  
J. Rabadan  
Nokia  
November 1, 2019

Extended Procedures for EVPN Optimized Ingress Replication  
draft-wsv-bess-extended-evpn-optimized-ir-02

Abstract

[EVPN-AR] specifies an optimized ingress replication solution for more efficient multicast and broadcast delivery in a Network Virtualization Overlay (NVO) network for EVPN.

This document extends the optimized ingress replication procedures specified in [EVPN-AR] to overcome the limitation that an AR-REPLICATOR may have. An AR-REPLICATOR may be unable to retain the source IP address or include the expected ESI label that is required for EVPN split horizon filtering when replicating the packet on behalf of its multihomed AR-LEAF. Under this circumstance, the extended procedures specified in this document allows the support of EVPN multihoming on the AR-LEAFs as well as optimized ingress replication for the rest of the EVPN overlay network.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 <https://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 May 4, 2020.

#### Copyright Notice

Copyright (c) 2019 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Terminology . . . . .	3
2. Introduction . . . . .	3
2.1. Background . . . . .	3
2.1.1. EVPN Multihoming and Split Horizon Filtering Rule . .	3
2.2. Optimized-IR and the Need to Maintain the Original Source IP address or Include the ESI Label . . . . .	4
3. Solution . . . . .	5
3.1. AR-REPLICATOR Announcing Multihoming Assistant Capability for Optimized-IR . . . . .	5
3.2. Multihomed AR-LEAF and Extended-MH AR-REPLICATOR . . . .	6
3.3. The Benefit of the Extended Optimized-IR Procedure . . . .	7
3.4. Support for Mixed AR-REPLICATORS . . . . .	7
4. Extended Optimized-IR Procedure for Supporting Extended-MH AR-REPLICATOR . . . . .	7
4.1. AR-LEAF Procedure . . . . .	8
4.1.1. Control Plane Procedure for AR-LEAF . . . . .	8
4.1.2. Forwarding Procedure for AR-LEAF . . . . .	9
4.2. AR-REPLICATOR Procedure . . . . .	9
4.2.1. Control Plane Procedure for AR-REPLICATOR . . . . .	9
4.2.2. Forwarding Procedure for AR-REPLICATOR . . . . .	10
4.3. RNVE Procedure . . . . .	10
5. AR-LEAF's Peer multihomed NVE in the Extended Optimized-IR Procedure . . . . .	11
6. Multicast Flags Extended Community . . . . .	11
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	12



9. Acknowledgements . . . . .	12
10. Normative References . . . . .	12
Authors' Addresses . . . . .	13

## 1. Terminology

### AR-IP Tunnel

An overlay tunnel with a destination IP address of AR-IP that an AR-REPLICATOR advertises in its REPLICATE-AR route.

This document heavily uses the terminology specified in [EVPN-AR]. It also uses the terminology specified in [RFC7432] and [RFC8365].

## 2. Introduction

### 2.1. Background

#### 2.1.1. EVPN Multihoming and Split Horizon Filtering Rule

This section gives a brief overview of the existing split horizon filtering rules used for EVPN multihoming.

[RFC7432] defines the split-horizon filtering rule based on ESI label for EVPN multihoming with MPLS encapsulation, and this filtering rule also applies for EVPN with IP-based encapsulation for MPLS, such as MPLS over GRE or MPLS over UDP. [RFC8365] defines the split horizon filtering rule based on "Local-Bias" for EVPN multihoming with VXLAN encapsulation.

When EVPN is used in an NVO network, a Tenant System (TS) may connect to a set of Network Virtualization Edge (NVE) devices through a multihomed Ethernet segment (ES). The split-horizon filtering rule for EVPN all-active multihoming ensures that a Broadcast, Unknown unicast or Multicast (BUM) packet received from an ES that is a part of a multihomed ES is not looped back to the multihomed TS through an egress NVE connected to the same multihomed ES. For EVPN with VXLAN encapsulation, the split-horizon filtering rule is based on the egress NVE examining the source IP address of the BUM packet received from an overlay tunnel. The egress PE identifies the ingress NVE through the source IP address. The egress NVE does not forward the BUM packet received from an overlay tunnel to the multihomed Ethernet segment that it has in common with the ingress NVE.

For EVPN with MPLS over IP tunnel, the split-horizon filtering rule is based on the ESI label. For ingress replication, an ESI label is downstream assigned per multihomed ES. The ingress NVE MUST include the ESI label, assigned by the egress PE, when it forwards a BUM

packet to the egress NVE if the BUM traffic is from the AC that is part of the multihomed ES associated with that ESI label. The egress NVE does not forward the BUM packet it received from an overlay tunnel to the multihomed ES if the ESI label is allocated by the egress NVE for that multihomed ES.

## 2.2. Optimized-IR and the Need to Maintain the Original Source IP address or Include the ESI Label

[EVPN-AR] specifies an optimized ingress replication procedures for the delivery of Multicast and Broadcast (BM) traffic within a bridge domain. It defines the control plane and forwarding plane procedures for AR-REPLICATOR, AR-LEAF and RNVE. To support EVPN AR-LEAF multihoming, [EVPN-AR] recommends that split horizon filtering rule based on "Local-Bias" procedures is used for EVPN NVO network using either 24-bit VNI or MPLS label.

To support EVPN all-active multihoming based on "Local-Bias" procedures, when an AR-REPLICATOR performs assisted replication on behalf of a multihomed AR-LEAF, the AR-REPLICATOR shall use the source IP address of the ingress AR-LEAF for packet received on the AR-IP tunnel. This ensures that other remote NVEs, when receiving a packet from its AR-REPLICATOR, can perform the regular split horizon filtering based on the source IP address.

To support EVPN all-active multihoming with MPLSoGRE or MPLSoUDP, sometimes it is desirable to continue using the existing split horizon filtering rule based on [RFC7432] procedures. In this case, when performing assisted replication on behalf of a multihomed AR-LEAF, an AR-REPLICATOR shall include the ESI label advertised by a remote NVE for that multihomed ES.

Due to either implementation complexity or hardware limitation, an AR-REPLICATOR may be unable to retain the source IP address or include the ESI label when replicating the packet to the remote NVEs on behalf of a multihomed AR-LEAF. Under this circumstance, when receiving the packet, a remote NVE is unable to use the existing split horizon filtering rules to prevent the looping of BM traffic required for all-active multihoming.

For example, with VXLAN encapsulation, consider a case where TS1 is multihomed to AR-LEAF1 and AR-LEAF2 through a multihomed ES. When AR-LEAF1 receives an IP multicast packet from TS1, AR-LEAF1 sends the packet to its AR-REPLICATOR with the source IP address set to AR-LEAF1's IR-IP and the destination IP address set to the AR-IP of the AR-REPLICATOR. Since the AR-REPLICATOR is unable to retain the source IP address for the packet it received on the AR-IP tunnel, the AR-REPLICATOR uses one of its own IP addresses as the source IP

address when it replicates the packet to other NVEs. When AR-LEAF2 receives the packet from the AR-REPLICATOR, it checks for the source IP address. AR-LEAF2 is unable to detect that this packet was originally sent by AR-LEAF1. If AR-LEAF2 is the DF for the multihomed ES connected to TS1, AR-LEAF2 forwards the packet to TS1. This causes the same IP multicast packet to be looped back to TS1.

The same problem can also happen to EVPN with MPLS over IP network if an AR-REPLICATOR cannot include the ESI label to the remote NVE for the multihomed ES when the split horizon filtering rule based on [RFC7432] is used.

### 3. Solution

This document extends the procedures defined in the [EVPN-AR] to support EVPN multihoming on AR-LEAFs when an NVE acts as an AR-REPLICATOR is incapable of retaining the source IP address or including an ESI label for its AR-LEAF either due to its hardware limitation or implementation complexity. The solution specified in this document is intended to work for EVPN over IP-based network with NVO tunnel using either 24-bit VNI or MPLS label. The solution relies on either [RFC7432] or "Local-Bias" split-horizon filtering rules to prevent the looping of BUM traffic. We refer to the procedures specified in this document as the extended Optimized-IR procedures. The extended Optimized-IR procedures also work with RNVE. The extended Optimized-IR procedures do not apply to EVPN with MPLS encapsulation.

#### 3.1. AR-REPLICATOR Announcing Multihoming Assistant Capability for Optimized-IR

An AR-REPLICATOR announces its AR-REPLICATOR role through the control plane. A REPLICATOR-AR route, as it is specified in the [EVPN-AR], is an Inclusive Multicast Ethernet Tag (IMET) route that an AR-REPLICATOR originates for its AR-IP and corresponding AR-replication tunnel.

If an AR-REPLICATOR cannot or chose not to retain the source IP address or include the expected ESI label for its multihomed AR-LEAFs, it MUST inform other NVEs in the control plane through the use of EVPN Multicast Flags Extended Community as follow: a) the AR-REPLICATOR MUST set the "Extended-MH-AR" flag, as it is specified in the section 6, in the multicast flags extended community, and b) it MUST attach this community to the REPLICATOR-AR route it originates. We call such an AR-REPLICATOR an Extended-MH AR-REPLICATOR.

An Extended-MH AR-REPLICATOR supports extended Optimized-IR procedures defined in this document for its multihomed AR-LEAFs. An

Extended-MH AR-REPLICATOR keeps track of its AR-LEAF's multihomed peer. An Extended-MH AR-REPLICATOR can perform assisted replication for an AF-LEAF to other NVEs that are not attached to the same multihomed ES as the AR-LEAF. An Extended-MH AR-REPLICATOR does not perform assisted replication for its AR-LEAF to other NVEs that have a multihomed ES in common with the AR-LEAF. The changes in the control plane and forwarding plan procedures for an Extended-MH AR-REPLICATOR is further explained in detail in section 5.2.

An AR-REPLICATOR originating a REPLICATOR-AR route without a multicast flags extended community or with the Extended-MH-AR flag unset is considered to be an MH-capable-assistant AR-REPLICATOR. An MH-capable-assistant AR-REPLICATOR can perform assisted replication for its single-homed AR-LEAF as well as multihomed AR-LEAF.

### 3.2. Multihomed AR-LEAF and Extended-MH AR-REPLICATOR

An AR-LEAF follows the control plane and forwarding plane procedures specified in [EVPN-AR]. In addition, if a multihomed AR-LEAF detects that one of its AR-REPLICATORS is Extended-MH AR-REPLICATOR based on the processing of its REPLICATOR-AR route, the multihomed AR-LEAF follows the extended Optimized-IR procedures specified in this document. With the extended Optimized-IR procedures, within the same BD, the multihomed AR-LEAF will use the regular ingress replication procedure to deliver a copy of a BUM packet received from its local AC to each of the remote NVEs that has a multihomed ES in common with it. In this way, the egress NVE can use the regular split horizon filtering rule defined in [RFC7432] or [RFC8365] to prevent the BUM traffic to be looped through the egress NVE to the source of origin. The extended procedures required for an AR-LEAF is further specified in detail in section 5.

For an AR-LEAF, please note that the additional forwarding procedures specified above apply to BM packets coming from any of its ACs in the same BD, whether that AC is a single homed ES or a part of a multihomed ES. It may also applies to Unknown unicast traffic. This is to further alleviate the burden of an Extended-MH AR-REPLICATOR as it may be unable to detect whether a packet received on its AR-IP tunnel was originally received from a single-homed or multihomed ES.

Consider an EVPN NVO network with a tenant domain consists of a set of  $m$  AR-LEAFs in BD X: AR-LEAF1, AR-LEAF2, AR-LEAF3, ..., AR-LEAF $m$ . TS1 is multihomed to AR-LEAF1 and AR-LEAF2 in BD X through a multihomed ES ES1. TS2 is multihomed to AR-LEAF1 and AR-LEAF3 in BD X through another multihomed ES ES2. Also, suppose that there are two Extended-MH AR-REPLICATORS in the same tenant domain: AR-REPLICATOR1 and AR-REPLICATOR2. AR-LEAF1 will detect that its AR-REPLICATORS are Extended-MH AR-REPLICATORS. AR-LEAF1 will also

detect that both AR-LEAF2 and AR-LEAF3 have a multihomed ES in common with it. AR-LEAF1 will use regular ingress replication to send the BUM traffic it receives from its access to both AR-LEAF2 and AR-LEAF3. AR-LEAF1 will rely on one of its AR-REPLICATORS to send the BM traffic to AR-LEAF4, AR-LEAF5, ..., and AR-LEAFm.

### 3.3. The Benefit of the Extended Optimized-IR Procedure

The extended Optimized-IR procedures specified in this document greatly reduces the implementation complexity of an AR-REPLICATOR or helps to overcome the limitation of an AR-REPLICATOR. It frees all AR-REPLICATORS from performing multihoming assisted replication while at the same time, it allows the support of EVPN multihoming on the AR-LEAFs with the existing multihoming procedures and split horizon filtering rules. For EVPN with MPLS over IP-based encapsulation, an NVE can continue to use the split horizon filtering rule based on the ESI label. Furthermore, it still allows the support of efficient Optimized-IR for the rest of an EVPN NVO network.

For example, in a typical NVO network, a TS is most likely multihomed to two or a small set of NVEs for redundancy. In an NVO network consisting of many NVEs, the AR-REPLICATOR is still responsible for replicating the BM packet to the most of NVEs for its AR-LEAF and thus it inherits the benefit of optimized ingress replication for the most of its NVO network.

### 3.4. Support for Mixed AR-REPLICATORS

When there are mixed MH-capable-assistant AR-REPLICATORS and Extended-MH AR-REPLICATORS in the same tenant domain, all AR capable NVEs MUST follow the extended Optimized-IR procedures as long as one of the AR-REPLICATORS is an Extended-MH AR-REPLICATOR.

When there are mixed AR-REPLICATORS, this document recommends that all MH-capable-assistant AR-REPLICATORS to be administratively provisioned to behave as Extended-MH AR-REPLICATORS. In this case, each AR-REPLICATOR originates its REPLICATOR-AR route with the Extended-MH-AR flag set in the multicast flags extended community.

The procedure for using mixed AR-REPLICATORS is beyond the scope of this document.

## 4. Extended Optimized-IR Procedure for Supporting Extended-MH AR-REPLICATOR

#### 4.1. AR-LEAF Procedure

This section covers the extended Optimized-IR procedures required for an AR-LEAF in further detail when at least one of the AR-REPLICATORS is an Extended-MH AR-REPLICATOR. It is assumed that an AR-LEAF follows the procedures defined in [EVPN-AR] unless it is specified otherwise.

##### 4.1.1. Control Plane Procedure for AR-LEAF

An AR-LEAF detects whether an AR-REPLICATOR is capable of performing multihoming assisted replication through the Extended-MH-AR flag in the multicast flags extended community carried in the REPLICATOR-AR route. An AR-REPLICATOR originating a REPLICATOR-AR route without a multicast flags extended community or with the Extended-MH-AR flag unset is considered to be multihoming assistant capable.

If an AR-LEAF does not have any locally attached segment that is a part of a multihomed ES, then there is no additional extended Optimized-IR procedure for an AR-LEAF to follow and we can go directly to section 4.2.

If selective assistant-replication is used for the EVI, selective AR-LEAFs that share the same multihomed ES MUST select the same primary AR-REPLICATOR and the same backup AR-REPLICATOR, if there is one. This can be achieved through either manual configuration on each multihomed selective AR-LEAF or by other methods that are beyond the scope of this document. Each selective AR-LEAF follows the procedures defined in the [EVPN-AR] to send its corresponding leaf-AD routes to its AR-REPLICATOR.

An AR-LEAF follows the normal procedures defined in [RFC7432] when it originates a type-4 ES route and type-1 Ethernet A-D routes for its locally attached segment that is a part of a multihomed ES.

In addition, an AR-LEAF builds a peer-multihomed-flood-list for each BD it attaches. Per normal EVPN procedures defined in [RFC7432], an AR-LEAF discovers the ESI of each multihomed ES that every remote NVE connects to. For a given BD, an AR-LEAF constructs a peer-multihomed-flood-list that consists of its peer multihomed NVEs in that BD that have at least one multihomed ES in common with it. An AR-LEAF may consider a common multihomed ES that it shares with a remote NVE in a BD specific scope or an EVI scope. Please section 5 for detail.

#### 4.1.2. Forwarding Procedure for AR-LEAF

Suppose that a multihomed AR-LEAF detects through the control plane procedure that at least one of its AR-REPLICATORS is an Extended-MH AR-REPLICATOR, then in addition to follow the forwarding procedures defined in [EVPN-AR], the AR-LEAF will use regular ingress replication to send the BUM packet, received from one of its ACs, to each NVE in that BD's peer-multihomed-flood-list.

In the case that there are no more AR-REPLICATORS in the tenant domain, the AR-LEAF reverts back to the regular IR behavior as it is defined in [RFC7432].

An AR-LEAF will follow the regular EVPN procedures when it receives a packet from an overlay tunnel and it will never send the packet back to the core.

#### 4.2. AR-REPLICATOR Procedure

This section describes the additional procedures for an AR-REPLICATOR when there is at least one AR-REPLICATOR in the same tenant domain that is an Extended-MH AR-REPLICATOR.

It is also assumed that an AR-REPLICATOR follows the procedures defined in [EVPN-AR] unless specified otherwise.

##### 4.2.1. Control Plane Procedure for AR-REPLICATOR

An NVE that performs an AR-REPLICATOR role follows the control plane procedures for AR-REPLICATOR defined in the [EVPN-AR].

In addition, if an AR-REPLICATOR is an Extended-MH AR-REPLICATOR or if it is administratively provisioned to behave as an Extended-MH AR-REPLICATOR, it SHALL attach a multicast flags extended community to its REPLICATOR-AR route with the Extended-MH-AR flag set.

An AR-REPLICATOR also discovers whether another AR-REPLICATOR is an Extended-MH AR-REPLICATOR based on the multicast flags extended community. If at least one AR-REPLICATOR is an Extended-MH AR replicator, then the rest of AR-REPLICATORS SHALL fall back to support the extended procedures specified in this document.

When there are mixed AR-REPLICATORS, this document recommends that all MH-capable-assistant AR-REPLICATORS SHOULD fall back to behave as Extended-MH AR-REPLICATORS through administrative provisioning.

An Extended-MH AR-REPLICATOR builds a multihomed list for each BD that its AR-LEAF attaches to. We refer to such a multihomed list as

an AR-LEAF's multihomed-list. Per normal EVPN procedures defined in [RFC7432], an AR-REPLICATOR imports the Ethernet A-D per EVI route, the alias route, originated by each remote NVE in the same tenant domain. For a given BD that an AR-LEAF belongs to, an AR-LEAF's multihomed-list consists of all the NVEs in that BD that have at least one multihomed ES in common with the said AR-LEAF. Please also refer to section 5 for the common multihomed ES an AR-LEAF shares with its remote NVE.

Consider an EVPN NVO network specified in the section 3.2. Both AR-LEAF1 and AR-LEAF2 originate its Ethernet A-D per EVI route for ES1 respectively. Both AR-LEAF1 and AR-LEAF3 originate its Ethernet A-D per EVI route for ES2 respectively. Per normal EVPN procedures, each AR-REPLICATOR imports and processes Ethernet A-D per EVI routes. Each AR-REPLICATOR builds an AR-LEAF1's multihomed-list for BD X that consists of AR-LEAF2 and AR-LEAF3. Each AR-REPLICATOR also builds AR-LEAF's multihomed-lists for other AR-LEAFs.

#### 4.2.2. Forwarding Procedure for AR-REPLICATOR

When an AR-REPLICATOR determines that it is an Extended-MH AR-REPLICATOR or determines that it SHALL fall back to become an Extended-MH AR-REPLICATOR, it MUST follow the forwarding procedures described in this section.

For a given BD, when an AR-REPLICATOR replicates the packet, received from its AR-IP tunnel, to other overlay tunnels on behalf of its ingress AR-LEAF, the AR-REPLICATOR MUST skip any NVE that is in that ingress AR-LEAF's multihomed-list built for that said BD.

When replicating the traffic to other AR-REPLICATORS or other AR-LEAFs over an overlay tunnel, an AR-REPLICATOR does not set the source IP address to its ingress AR-LEAF's IR-IP. It is assumed under the scope of this document that an AR-LEAF does not share any common multihoming ES with any AR-REPLICATOR.

When replicating the traffic to other RNVEs, an AR-REPLICATOR should set the source IP address to its own IR-IP. This is because an RNVE does not recognize the AR-IP.

#### 4.3. RNVE Procedure

There is no change to the RNVE control and forwarding procedures. RNVE follows the regular ingress replication procedure defined in [RFC7432].



## 5. AR-LEAF's Peer multihomed NVE in the Extended Optimized-IR Procedure

For the extended Optimized-IR procedures specified in this document, a multihomed AR-LEAF may keep track of the common multihomed ES it shares with other remote NVEs in a BD specific scope or in an EVI scope. Correspondingly, an Extended-MH AR-REPLICATOR MUST also use the same scheme to keep track of the common multihomed ES that its AR-LEAF shares with other remote NVEs. All multihomed AR-LEAFs and all AR-REPLICATORS within the same EVI MUST use the same scheme to keep track of the common multihomed ES that an AR-LEAF shares with other remote NVEs. This consistency can be enforced through a manual configuration.

A multihomed AR-LEAF maintains a peer-multihomed-flood-list for each BD it attaches. If the common multihomed ES is tracked in a per EVI scope, an AR-LEAF's peer-multihomed-flood-list for a given BD X contains all the NVEs in BD X that have at least one multihomed ES in common with it, regardless whether each common multihomed ES contains BD X or not. If the common multihomed ES is tracked in a BD specific scope, for a given BD X, each common multihomed ES must contain BD X. The same MUST be applied to the AR-LEAF's multihomed-list for BD X an AR-REPLICATOR maintains for its AR-LEAF.

When the Ethernet A-D per EVI route is advertised at the granularity of per ES, the common multihomed ES is tracked in a per EVI scope.

## 6. Multicast Flags Extended Community

The EVPN Multicast Flags Extended Community is defined in the [EVPN-IGMP-PROXY]. This transitive extended community can carry many flags in its Flags field. This document proposes one new flag in the Flags bit vector.

### o Extended-MH-AR

The Extended-MH-AR flag, M flag for short, takes the next available low-order bit from the Flags field.

The Extended-MH-AR flag is used by the AR-REPLICATOR. When this flag is set, the AR-REPLICATOR indicates to other NVEs that it will not retain the source IP address or include the ESI label for an ingress NVE when replicating the packet over an NVO tunnels on behalf of the ingress NVE. Such an AR-REPLICATOR supports the extended optimized-IR procedures defined in this document.

## 7. IANA Considerations

A request for a new flag named Extended-MH-AR flag in the Flags field of the multicast flags extended community will be submitted to IANA.

## 8. Security Considerations

This document inherits the same securities as they are defined in the [RFC7432], [RFC8365] and [EVPN-AR].

## 9. Acknowledgements

The authors would like to thank Eric Rosen and Jeffrey Zhang for their valuable comments and feedbacks. The authors would also like to thank Aldrin Isaac for his useful discussion, insight on this subject.

## 10. Normative References

- [EVPN-AR] Rabadan, J., Ed., "Optimized Ingress Replication solution for EVPN", internet-draft ietf-bess-evpn-optimized-ir-06.txt, October 2018.
- [EVPN-IGMP-PROXY] Sajassi, A., Ed., "IGMP and MLD Proxy for EVPN", internet-draft ietf-bess-evpn-igmp-ml-d-proxy-04.txt, June 2018.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.

Authors' Addresses

Wen Lin (editor)  
Juniper Networks, Inc.

EMail: wlin@juniper.net

Selvakumar Sivaraj  
Juniper Networks, Inc.

EMail: ssivaraj@juniper.net

Vishal Garg  
Juniper Networks, Inc.

EMail: vishalg@juniper.net

Jorge Rabadan  
Nokia

EMail: jorge.rabadan@nokia.com

BESS WG  
Internet-Draft  
Intended status: Standards Track  
Expires: April 25, 2020

Z. Zhang  
Y. Wang  
G. Mirsky  
ZTE Corporation  
October 23, 2019

Bidirectional Forwarding Detection (BFD) for EVPN Ethernet Segment  
Failover Use Case  
draft-zwm-bess-es-failover-01.txt

Abstract

This document introduces a method for fast switchover of Designated Forwarder for Ethernet Segment failover by using Bidirectional Forwarding Detection protocol.

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 <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on April 25, 2020.

Copyright Notice

Copyright (c) 2019 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 (<https://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.

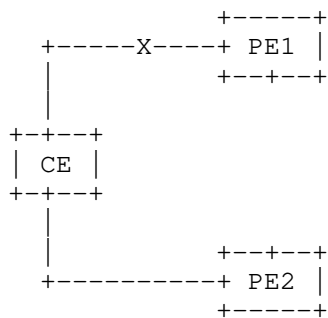
# Table of Contents

1.	Introduction . . . . .	2
2.	Conventions used in this document . . . . .	3
2.1.	Terminology . . . . .	3
2.2.	Requirements Language . . . . .	3
3.	Proposal . . . . .	3
4.	Specification . . . . .	4
4.1.	BDF changes . . . . .	5
5.	Security Considerations . . . . .	5
6.	IANA Considerations . . . . .	5
7.	Normative References . . . . .	5
	Authors' Addresses . . . . .	6

## 1. Introduction

[RFC7432] introduces Ethernet Virtual Private Network (EVPN) technology. Designated Forwarder (DF) election procedures for multi-homing Ethernet Segments has been described in it. When PE (provider edge) receives BUM (Broadcast, Unknown Unicast and Multicast) flows, only DF forwards the BUM flows to CE (customer edge). Non-DFs do not forward the BUM flows in order to avoid duplication. If the link between DF and CE fails, another PE will forward the BUM flows after it is elected as DF.

[RFC8584] defines the DF election framework, including that Backup Designated Forwarder (BDF) can be elected as the next best for the role. But before the BDF is elected as DF, the BUM flows are discarded after the link between DF and CE fails.



For example, CE is multihomed to PE1 and PE2. PE1 is elected as DF. All BUM flows are forwarded by PE1 when the link between PE1 and CE is operational. When the link between PE1 and CE fails, the BUM flows are discarded until PE2 is elected as DF.

This document will use terminology defined in [RFC7432] and [I-D.ietf-bess-evpn-lsp-ping].

## 2. Conventions used in this document

### 2.1. Terminology

BFD: Bidirectional Forwarding Detection

BDF: Backup Designated Forwarder

DF: Designated Forwarder

BUM: Broadcast, Unknown unicast, and Multicast

PE: Provider Edge

EVPN: Ethernet Virtual Private Network

CE: Customer Edge

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

### 2.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 3. Proposal

In order to avoid the BUM packet loss on BDF after the link between DF and CE fails, a data-plane detection function is needed for DF fast switchover. [RFC5884] provides mechanisms for using LSP Ping to bootstrap a BFD session. [I-D.ietf-bess-evpn-lsp-ping] introduces four new Target FEC Stack sub-TLVs that are included in the LSP-Ping Echo Request packet. This document uses the mechanisms defined in [RFC5884] and the EVPN Ethernet Auto-Discovery (AD) sub-TLV defined in [I-D.ietf-bess-evpn-lsp-ping] to provide DF fast switchover by data-plane failure detection.

An LSP-Ping Echo Request message which carries EVPN AD Sub-TLV associated with the DF-CE Ethernet Segment Identifier (ESI) is used to bootstrap the BFD session between BDF and DF. After the BFD

session is built, when the Ethernet Segment (ES) fault occurs on DF-CE link, BDF detects the fault by the state change BFD control packet sent by DF, or BDF detects the fault when the detection timer expires. Then BDF becomes DF and will forward the BUM flows to CE.

#### 4. Specification

[I-D.ietf-bess-evpn-lsp-ping] section 4.3 defines an Ethernet AD sub-TLV as a new Target FEC Stack sub-TLV. It is carried in the LSP-Ping Echo Request message. BDF generates an LSP-Ping Echo Request message which carries the associated ES AD sub-TLV. And BDF sends the message with a local discriminator assigned by BDF for this BFD session to DF. DF responds with the BFD control packet with 'Your discriminator' set to the discriminator value received in the Echo request message from the BDF. BDF can demultiplex the BFD session based on the received 'Your Discriminator' field.

After the BFD session is established, when the link between DF and CE fails, DF MUST send a BFD control packet with the value of State field set to AdminDown through the established BFD session to BDF. If DF is not operational, BDF also detects the failure when the BFD detection time expires. Then BDF becomes DF immediately and forwards the BUM flows to CE.

When the ES between 'old' DF and CE recovers, the BFD session MAY be reused or a new BFD session can be established for the ES failover monitor.

For the same example in last section, PE2 generates an LSP-Ping Echo Request message which carries the associated ES AD sub-TLV and sends the message with an assigned local discriminator to DF. PE1 responds with a BFD control packet with 'Your Discriminator' set to the received discriminator from PE2. PE2 can demultiplex the BFD session based on the received 'Your Discriminator' field.

When the link between PE1 and CE fails, PE1 sends a BFD control packet with the state set to AdminDown to PE2 through the BFD session. If the packet is lost, PE2 also can detect the fault by the session detection time expiration. PE2 becomes DF immediately, then the BUM packets can be forwarded to CE.

The value of bfd.DetectMult (detect multiplier) determines when a BFD system detects a failure. Once BDF detects the loss of the number, equal to the detect multiplier, of consecutive BFD messages for the session between DF and BDF are lost, the BDF will elect itself as DF. Then, BUM flows are duplicated because of the two DFs. To avoid this situation, the bfd.DetectMult MUST be set to more than 1 (common default value is 3).

#### 4.1. BDF changes

If a new router, which can become new BDF, joins the network, the 'old' BDF MUST send a number of consecutive BFD messages with the State set to AdminDown to DF, then DF will remove this BFD session. When DF receives a new session request from the new BDF, DF establishes a new BFD session with the new BDF.

#### 5. Security Considerations

This document does not introduce any new security considerations other than already discussed in [RFC7432] and [RFC5884].

#### 6. IANA Considerations

There is no IANA consideration.

#### 7. Normative References

- [I-D.ietf-bess-evpn-lsp-ping]  
Jain, P., Salam, S., Sajassi, A., Boutros, S., and G. Mirsky, "LSP-Ping Mechanisms for EVPN and PBB-EVPN", draft-ietf-bess-evpn-lsp-ping-00 (work in progress), May 2019.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <<https://www.rfc-editor.org/info/rfc5884>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.



[RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

#### Authors' Addresses

Zheng(Sandy) Zhang  
ZTE Corporation  
No. 50 Software Ave, Yuhuatai Distinct  
Nanjing  
China

Email: [zzhang\\_ietf@hotmail.com](mailto:zzhang_ietf@hotmail.com)

Yubao Wang  
ZTE Corporation  
No. 50 Software Ave, Yuhuatai Distinct  
Nanjing  
China

Email: [wang.yubao2@zte.com.cn](mailto:wang.yubao2@zte.com.cn)

Greg Mirsky  
ZTE Corporation

Email: [gregimirsky@gmail.com](mailto:gregimirsky@gmail.com)