

INTERNET-DRAFT
Intended Status: Proposed Standard

Patrice Brissette
Ali Sajassi
Luc Andre Burdet
Cisco Systems

Daniel Voyer
Bell Canada

Expires: April 23, 2019

October 20, 2018

EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols
draft-brissette-bess-evpn-l2gw-proto-03

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 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	3
1.1 Terminology	3
1.2 Acronyms	3
2. Solution	5
3. Requirements	6
4. Handling of Topology Change Notification (TCN)	7
5. ESI-label Extended Community Extension	8
6. EVPN MAC Flush Extcomm	8
7. EVPN Inter-subnet Forwarding	9
8. Conclusion	9
9. Security Considerations	10
10. Acknowledgements	10
11. IANA Considerations	10
12. References	10
12.1 Normative References	10
12.2 Informative References	10

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 Terminology

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

1.2 Acronyms

AC	: Attachment Circuit
BUM	: Broadcast, Unknown unicast, Multicast
DF	: Designated Forwarder
EVLAG	: EVPN LAG (equivalent to EVPN MC-LAG)
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

INTERNET DRAFT draft-brisette-bess-evpn-l2gw-proto October 20, 2018

REP-AG : Resilient Ethernet Protocol Access Gateway
TCN : Topology Change Notification

2. Solution

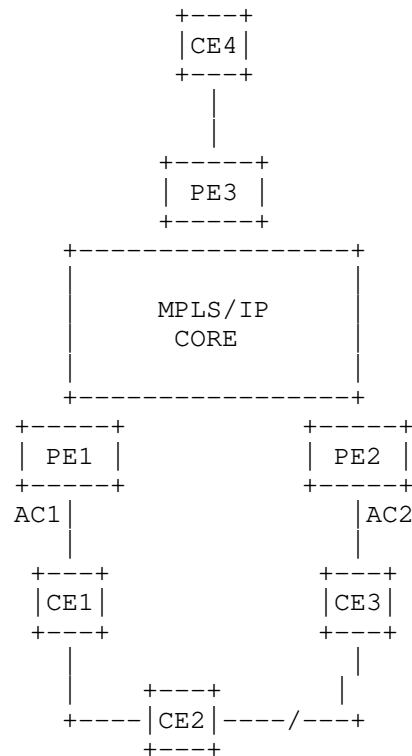


Figure 1 EVPN network with L2 access GW protocols

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. PE1 and PE2 are peering PEs (EVLAG capable) in a redundancy group sharing a same ESI. The L2GW protocol is used for loop avoidance. In above example, the loop is broken on the right side of CE2. 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 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.

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 guarantees drive that state.
- 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 per ES/EAD and per EVI/EAD 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 }

- 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 When MST-AG runs in the access, 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 extcomm is extended. The 1 octet flag field, as part of the ESI-label Extcomm, is updated as follow:

Each ESI Label extended community is encoded as an 8-octet value, as follows:

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=0x01									Flags(1 octet)									Reserved=0																				
Reserved=0									ESI Label																																						

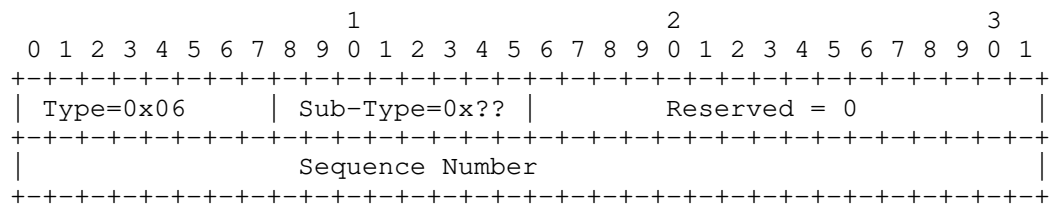
```
Low-order bit: [7:0]
[2:0]- 000 = all-active,
        001 = single-active,
        010 = single-flow-active,
        others = unassigned
[7:3]- Reserved
```

6. EVPN MAC Flush Extcomm

A new BGP Extended community, similar to MAC mobility BGP-extcomm, is required by the TCN procedure. It may get advertised along with Ethernet Auto-discovery routes (per EVI/EAD) 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 readvertised 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.



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.

8. Conclusion

EVPN 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

[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>>.

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 Brisette
Cisco Systems
EMail: pbrisset@cisco.com

Ali Sajassi
Cisco Systems
EMail: sajassi@cisco.com

Luc Andre Burdet
Cisco Systems
EMail: lburdet@cisco.com

INTERNET DRAFT draft-brisette-bess-evpn-l2gw-prot0 October 20, 2018

Daniel Voyer
Bell Canada
EMail: daniel.voyer@bell.ca

INTERNET-DRAFT
Intended Status: Proposed Standard

Patrice Brissette
Samir Thoria
Ali Sajassi
Cisco Systems

Expires: April 25, 2019

October 22, 2018

EVPN multi-homing port-active load-balancing
draft-brissette-bess-evpn-mh-pa-02

Abstract

The Multi-Chassis Link Aggregation Group (MC-LAG) technology enables the establishment of a logical port-channel 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 draft, 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) 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	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
5.	Port-active over Integrated Routing-Bridging Interface	6
6.	Convergence considerations	7
6.	Applicability	7
7.	Overall Advantages	8
8	Security Considerations	9
9	IANA Considerations	9
10.	Acknowledgements	9
11	References	9
11.1	Normative References	9
11.2	Informative References	9
	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. Main consideration for this mode of redundancy is the determinism of traffic forwarding through 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.

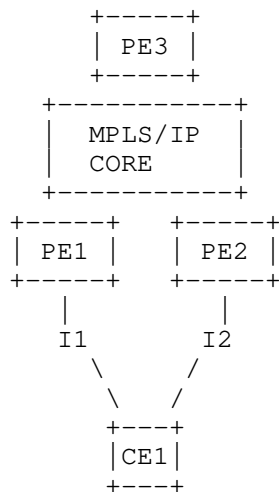


Figure 1. MC-LAG topology

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. When EVPN is used to provide MC-LAG functionality, we refer to it as EVLAG in this draft.

1.1 Terminology

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

2. 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. ICCP-based protocol has been used for that purpose. EVLAG simplifies greatly that solution. Along with the simplification comes few assumptions:

- Links in the Ethernet Bundle MUST operate in all-active load-balancing mode
- Same LACP parameters MUST be configured on peering PEs such as system id, port priority, etc.

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 EVLAG to support port-active load-balancing mode:

- 1- 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.
- 2- Ethernet-Segment MUST be configured in port-active load-balancing mode on peering PEs for specific interface

3- Peering PEs MAY exchange only Ethernet-Segment route (Route Type-4)

4- PEs in the redundancy group leverages DF election defined in [draft-ietf-bess-evpn-df-election-framework] 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 [draft-ietf-bess-evpn-df-election-framework] is per <ES, VLAN> 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.

4. Algorithm to elect per port-active PE

The default DF Election algorithm, or modulus-based algorithm as in [RFC7432], is used here also, at the granularity of <ES> only. For Modulo calculation, byte 10 of the ESI is used.

Highest Random Weight (HRW) algorithm defined in [draft-ietf-bess-evpn-df-election-framework] MAY also be used and signaled, and modified to operate at the granularity of <ES> rather than per <ES, VLAN>.

Let Active(ESI) denote the PE that will be the active PE for port with Ethernet segment identifier - ESI. The other PEs in the redundancy group will be standby PE(s) for the same port (ES). A_i is the address of the PE_i and $weight()$ is a pseudorandom function of ESI and A_i , $Wrand()$ function defined in [draft-ietf-bess-evpn-df-election-framework] is used as the $Weight()$ function.

$Active(ESI) = PE_i$: if $Weight(ESI, A_i) \geq Weight(ESI, A_j)$, for all j , $0 \leq i, j \leq \text{Number of PEs in the redundancy group}$. In case of a tie, choose the PE whose IP address is numerically the least.

5. Port-active over Integrated Routing-Bridging Interface

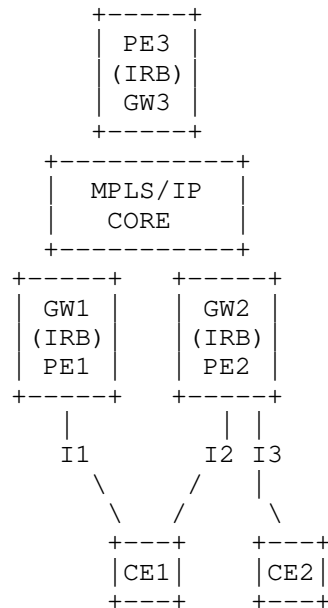


Figure 2. EVPN-IRB Port-active load-balancing

Figure 2 shows a simple network where EVPN-IRB is used for inter-subnet connectivity. IRB interfaces on PE1 and PE2 are configured in anycast gateway (same MAC, same IP). CE1 device is multi-homed to both PE1 and PE2. The Ethernet-segment load-balancing mode, of the connected CE1 to peering PEs, can be of any type e.g. all-active, single-active or port-active. CE2 device is connected to a single PE (PE2). It operates as single-homed device via an orphan port I3. Finally, port-active load-balancing is apply to IRB interface on peering PEs (PE1 and PE2). Manual Ethernet-Segment Identifier is assigned per IRB interface. ESI auto-generation is also possible based on the IRB anycast IP address.

DF election is performed between peering PE over IRB interface (per ESI/EVI). Designed forwarder (DF) IRB interface remains in up state. Non-designated forwarder (NDF) IRB interface may goes in down state. Furthermore, if all access interfaces connected to an IRB interface are down state (failure or admin) OR in blocked forward state(NDF), IRB interface is brought down. For example, interface I3 fails at the same time than interface I2 (in single-active load-balancing mode) is in blocked forwarding state.

In the example where IRB on PE2 is NDF, all L3 traffic coming from

PE3 is going via PE1. An IRB interface in down state doesn't attract traffic from core side. CE2 device reachability is done via an L2 subnet stretch between PE1 and PE2. Therefore L3 traffic coming from PE3 destined to CE2 goes via GW1 first, then via an L2 connection to PE2 and finally via interface I3 to CE2 device.

There are many reasons of configuring port-active load-balancing mode over IRB interface:

- Ease replacement of legacy technology such VRRP / HSRP
- Better scalability than legacy protocols
- Traffic predictability
- Optimal routing and entirely independent of load-balancing mode configured on any access interfaces

6. 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 / MLD cache may be synchronized. Moreover, associated VRF tables may also be synchronized. For L2 services, MAC table synchronization may be considered. Finally, using bundle-Ethernet interface, where LACP is running, is usually a smart thing since it provides the ability to set the "standby" port in "out-of-sync" state aka "warm-standby".

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 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 overlay encapsulation.

7. Overall Advantages

There are many advantages in EVLAG to support port-active load-balancing mode. Here is a non-exhaustive list:

- 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 RFC-7432, 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

There are no new IANA considerations in this document.

10. Acknowledgements

Authors would like to thank Luc Andre Burdet for valuable reviews and inputs.

11 References

11.1 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>>.
- [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>>.

11.2 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>>.
- [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

Patrice Brissette

Cisco Systems
EMail: pbrisset@cisco.com

Samir Thoria
Cisco Systems
EMail: sthoria@cisco.com

Ali Sajassi
Cisco Systems
EMail: sajassi@cisco.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2019

G. Dawra, Ed.
LinkedIn
C. Filsfils
D. Dukes
P. Brissette
S. Sethuram
P. Camarilo
Cisco Systems
J. Leddy
Comcast
D. Voyer
D. Bernier
Bell Canada
D. Steinberg
Steinberg Consulting
R. Raszuk
Bloomberg LP
B. Decraene
Orange
S. Matsushima
SoftBank
S. Zhuang
Huawei Technologies
March 11, 2019

SRv6 BGP based Overlay services
draft-dawra-bess-srv6-services-00

Abstract

This draft defines procedures and messages for SRv6-based BGP services including L3VPN, EVPN and Internet services. It builds on RFC4364 "BGP/MPLS IP Virtual Private Networks (VPNs)" and RFC7432 "BGP MPLS-Based Ethernet VPN" and provides a migration path from MPLS-based VPNs to SRv6 based VPNs.

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 RFC8174 [RFC8174].

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 September 12, 2019.

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. SRv6 Services TLVs	4
2.1. SRv6 Service Sub-TLVs	5
2.1.1. SRv6 SID Information Sub-TLV	6
2.1.2. SRv6 Service Data Sub-Sub-TLVs	7
3. BGP based L3 service over SRv6	7
3.1. IPv4 VPN Over SRv6 Core	8
3.2. IPv6 VPN Over SRv6 Core	9
3.3. Global IPv4 over SRv6 Core	9
3.4. Global IPv6 over SRv6 Core	9
4. BGP based Ethernet VPN (EVPN) over SRv6	10
4.1. Ethernet Auto-discovery route over SRv6 Core	11
4.1.1. Per-ES A-D route	11
4.1.2. Per-EVI A-D route	12

4.2.	MAC/IP Advertisement route over SRv6 Core	12
4.3.	Inclusive Multicast Ethernet Tag Route over SRv6 Core . .	13
4.4.	Ethernet Segment route over SRv6 Core	15
4.5.	IP prefix route over SRv6 Core	15
4.6.	EVPN multicast routes (Route Types 6, 7, 8) over SRv6 core	16
5.	Migration from MPLS based Segment Routing to SRv6 Segment Routing	16
6.	Implementation Status	17
7.	Error Handling	18
8.	IANA Considerations	19
8.1.	BGP Prefix-SID TLV Types registry	19
8.2.	SRv6 Service Sub-TLV Types registry	20
9.	Security Considerations	20
10.	Conclusions	20
11.	References	20
11.1.	Normative References	20
11.2.	Informative References	21
Appendix A.	Contributors	23
Authors' Addresses	23

1. Introduction

SRv6 refers to Segment Routing instantiated on the IPv6 dataplane [I-D.filsfils-spring-srv6-network-programming] [I-D.ietf-6man-segment-routing-header].

SRv6 based BGP services refers to the L3 and L2 overlay services with BGP as control plane and SRv6 as dataplane.

SRv6 SID refers to a SRv6 Segment Identifier as defined in [I-D.filsfils-spring-srv6-network-programming].

SRv6 Service SID refers to an SRv6 SID that MAY be associated with one of the service specific behavior on the advertising Provider Edge(PE) router, such as (but not limited to), in the case of L3VPN service, END.DT (Table lookup in a VRF) or END.DX (crossconnect to a nexthop) functions as defined in [I-D.filsfils-spring-srv6-network-programming].

To provide SRv6 service with best-effort connectivity, the egress PE signals an SRv6 Service SID with the BGP overlay service route. The ingress PE encapsulates the payload in an outer IPv6 header where the destination address is the SRv6 Service SID provided by the egress PE. The underlay between the PEs only need to support plain IPv6 forwarding [RFC2460].

To provide SRv6 service in conjunction with an underlay SLA from the ingress PE to the egress PE, the egress PE colors the overlay service route with a Color extended community[I-D.ietf-idr-segment-routing-te-policy]. The ingress PE encapsulates the payload packet in an outer IPv6 header with an SRH that contains the SR policy associated with the related SLA followed by the SRv6 Service SID associated with the route. The underlay nodes whose SRv6 SID's are part of the SRH must support SRv6 data plane.

BGP is used to advertise the reachability of prefixes of a particular service from an egress PE to ingress PE nodes.

This document describes how existing BGP messages between PEs may carry SRv6 Service SIDs as a means to interconnect PEs and form VPNs.

2. SRv6 Services TLVs

This document extends the BGP Prefix-SID attribute [I-D.ietf-idr-bgp-prefix-sid] to carry SRv6 SIDs and associated information.

The SRv6 Service TLVs are defined as two new TLVs of the BGP Prefix-SID Attribute to achieve signaling of SRv6 SIDs for L3 and L2 services.

- o SRv6 L3 Service TLV: This TLV encodes Service SID information for SRv6 based L3 services. It corresponds to the equivalent functionality provided by an MPLS Label when received with a Layer 3 service route. Some functions which may be encoded, but not limited to, are End.DX4, End.DT4, End.DX6, End.DT6, etc.
- o SRv6 L2 Service TLV: This TLV encodes Service SID information for SRv6 based L2 services. It corresponds to the equivalent functionality provided by an MPLS Label for EVPN Route-Types as defined in[RFC7432]. Some functions which may be encoded, but not limited to, are End.DX2, End.DX2V, End.DT2U, End.DT2M etc.

BGP Prefix-SID Attribute [I-D.ietf-idr-bgp-prefix-sid] is referred to as BGP SID Attribute in the rest of the document.

When an egress PE is capable of SRv6 data-plane, it SHOULD signal one or more SRv6 Service SIDs enclosed in SRv6 Service TLV(s) within the BGP SID Attribute attached to MP-BGP NLRI's defined in [RFC4760] [RFC4659] [RFC5549] [RFC7432] [RFC4364].

The following depicts the SRv6 Service TLVs encoded in the BGP SID attribute:

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|  TLV Type   |          TLV Length          |  RESERVED   |
+-----+-----+-----+-----+-----+-----+-----+-----+
//  SRv6 Service Sub-TLVs                                     //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

- o TLV Type (1 octet): This field is assigned values from the IANA registry "BGP Prefix-SID TLV Types". It is set to [TBD1] (to be assigned by IANA) for SRv6 L3 Service TLV. It is set to [TBD2] (to be assigned by IANA) for SRv6 L2 Service TLV.
- o TLV Length (2 octets): Specifies the total length of the TLV Value.
- o RESERVED (1 octet): This field is reserved; it SHOULD be set to 0 by the sender and MUST be ignored by the receiver.
- o SRv6 Service Sub-TLVs (variable): This field contains SRv6 Service related information and is encoded as an unordered list of Sub-TLVs whose format is described below.

2.1. SRv6 Service Sub-TLVs

The format of a single SRv6 Service Sub-TLV is depicted below:

```

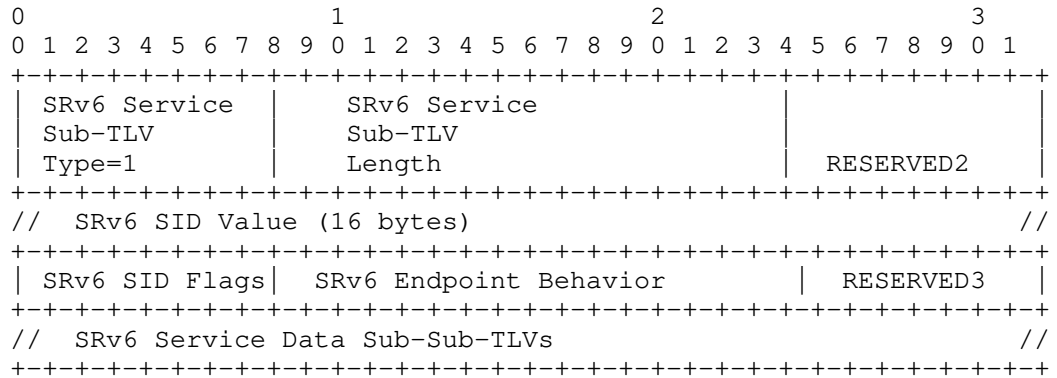
      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| SRv6 Service |          SRv6 Service          | SRv6 Service //
| Sub-TLV      |          Sub-TLV          | Sub-TLV      //
| Type         |          Length         | value        //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

- o SRv6 Service Sub-TLV Type (1 octet): Identifies the type of SRv6 service information. It is assigned values from the IANA Registry "SRv6 Service Sub-TLV Types".
- o SRv6 Service Sub-TLV Length (2 octets): Specifies the total length of the Sub-TLV Value field.
- o SRv6 Service Sub-TLV Value (variable): Contains data specific to the Sub-TLV Type. In addition to fixed length data, this may also optionally contain other properties of the SRv6 Service encoded as a set of SRv6 Service Data Sub-sub-TLVs whose format is described in another sub-section below.

2.1.1.1. SRv6 SID Information Sub-TLV

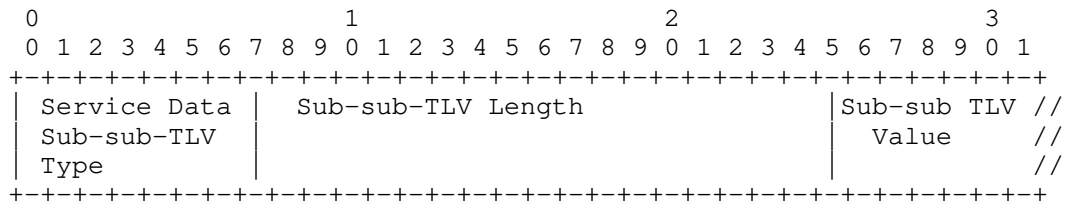
SRv6 Service Sub-TLV Type 1 is assigned for SRv6 SID Information Sub-TLV. This Sub-TLV contains a single SRv6 SID along with its properties. Its encoding is depicted below:



- o SRv6 Service Sub-TLV Type (1 octet): This field is set to 1 to represent SRv6 SID Information Sub-TLV.
- o SRv6 Service Sub-TLV Length (2 octets): This field contains the total length of the Value field of the Sub-TLV.
- o RESERVED2 (1 octet): SHOULD be set to 0 by the sender and MUST be ignored by the receiver.
- o SRv6 SID Value (16 octets): Encodes an SRv6 SID as defined in [I-D.filsfils-spring-srv6-network-programming]
- o SRv6 SID Flags (1 octet): Encodes SRv6 SID Flags - none are currently defined.
- o SRv6 Endpoint Behavior (2 octets): Encodes SRv6 Endpoint behavior defined in [I-D.filsfils-spring-srv6-network-programming]. This field MUST be set to the Reserved value 0xFFFF.
- o RESERVED3 (1 octet): SHOULD be set to 0 by the sender and MUST be ignored by the receiver.
- o SRv6 Service Data Sub-TLV Value (variable): This field contains optional properties of the SRv6 SID. It is encoded as a set of SRv6 Service Data Sub-Sub-TLVs. None are applicable at this time.

2.1.2. SRv6 Service Data Sub-Sub-TLVs

The format of the SRv6 Service Data Sub-Sub-TLV is depicted below:



- o SRv6 Service Data Sub-Sub-TLV Type (1 octet): Identifies the type of Sub-Sub-TLV. It is assigned values from the IANA Registry "SRv6 Service Data Sub-Sub-TLVs".
- o SRv6 Service Data Sub-Sub-TLV Length (2 octets): Specifies the total length of the Sub-Sub-TLV Value field.
- o SRv6 Service Data Sub-Sub-TLV Value (variable): Contains data specific to the Sub-Sub-TLV Type.

At this time, no Sub-Sub-TLV Types are defined.

3. BGP based L3 service over SRv6

BGP egress nodes (egress PEs) advertise a set of reachable prefixes. Standard BGP update propagation schemes [RFC4271], which may make use of route reflectors [RFC4456], are used to propagate these prefixes. BGP ingress nodes (ingress PEs) receive these advertisements and may add the prefix to the RIB in an appropriate VRF.

Egress PEs which supports SRv6 based L3 services advertises overlay service prefixes along with a Service SID enclosed in a SRv6 L3 Service TLV within the BGP SID attribute. This TLV serves two purposes - first, it indicates that the egress PE is reachable via an SRv6 underlay and the BGP ingress PE receiving this route MAY choose to encapsulate or insert an SRv6 SRH; second, it indicates the value of the SID to include in the SRH encapsulation.

The Service SID thus signaled only has local significance at the egress PE, where it may be allocated or configured on a per-CE or per-VRF basis. In practice, the SID may encode a cross-connect to a specific Address Family table (END.DT) or next-hop/interface (END.DX) as defined in the SRv6 Network Programming Document [I-D.filsfils-spring-srv6-network-programming].

The SRv6 Service SID MAY be routable within the AS of the egress PE and serves the dual purpose of providing reachability between ingress PE and egress PE while also encoding the endpoint behavior.

If the BGP speaker supports MPLS based L3VPN services simultaneously, it MAY also populate a valid Label value in the service route NLRI encoding, and allow the BGP ingress PE to decide which encapsulation to use. If the BGP speaker does not support MPLS based L3VPN services the Label value in any service route NLRI encoding MUST be set to Implicit NULL [RFC3032].

At an ingress PE, BGP installs the received prefix in the correct RIB table, recursing via an SR Policy leveraging the received SRv6 Service SID.

Assuming best-effort connectivity to the egress PE, the SR policy has a path with a SID list made up of a single SID - the SRv6 Service SID received with the related BGP route update.

However, when the received route is colored with an extended color community 'C' and Next-Hop 'N', and the ingress PE has a valid SRv6 Policy (C, N) associated with SID list <S1,S2, S3> [I-D.filsfils-spring-segment-routing-policy], then the effective SR Policy is <S1, S2, S3, SRv6-Service-SID>.

Multiple VPN routes MAY resolve recursively via the same SR Policy.

3.1. IPv4 VPN Over SRv6 Core

IPv4 VPN Over IPv6 Core is defined in [RFC5549]. The MP_REACH_NLRI is encoded as follows for an SRv6 Core:

- o AFI = 1
- o SAFI = 128
- o Length of Next Hop Network Address = 16 (or 32)
- o Network Address of Next Hop = IPv6 address of the egress PE
- o NLRI = IPv4-VPN routes
- o Label = Implicit NULL

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DX4 or End.DT4.

3.2. IPv6 VPN Over SRv6 Core

IPv6 VPN over IPv6 Core is defined in [RFC4659]. The MP_REACH_NLRI is encoded as follows for an SRv6 Core:

- o AFI = 2
- o SAFI = 128
- o Length of Next Hop Network Address = 24 (or 48)
- o Network Address of Next Hop = 8 octets of RD set to 0 followed by IPv6 address of the egress PE
- o NLRI = IPv6-VPN routes
- o Label = Implicit NULL

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DX6 or End.DT6.

3.3. Global IPv4 over SRv6 Core

IPv4 over IPv6 Core is defined in [RFC5549]. The MP_REACH_NLRI is encoded with:

- o AFI = 1
- o SAFI = 1
- o Length of Next Hop Network Address = 16 (or 32)
- o Network Address of Next Hop = IPv6 address of Next Hop
- o NLRI = IPv4 routes

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DX4/6 or End.DT4.

3.4. Global IPv6 over SRv6 Core

The MP_REACH_NLRI is encoded with:

- o AFI = 2

- o SAFI = 1
- o Length of Next Hop Network Address = 16 (or 32)
- o Network Address of Next Hop = IPv6 address of Next Hop
- o NLRI = IPv6 routes

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DX4/6 or End.DT6.

Also, by utilizing the SRv6 L3 Service TLV to encode the Global SID, a BGP free core is possible by encapsulating all BGP traffic from edge to edge over SRv6.

4. BGP based Ethernet VPN (EVPN) over SRv6

Ethernet VPN(EVPN), as defined in [RFC7432] provides an extendable method of building an EVPN overlay. It primarily focuses on MPLS based EVPNs but calls out the extensibility to IP based EVPN overlays. [RFC7432] defines 4 Route Types which carry prefixes and MPLS Label fields; the Label fields have specific use for MPLS encapsulation of EVPN traffic. Route Type 5 carrying MPLS label information (and thus encapsulation information) for EVPN is defined in [I-D.ietf-bess-evpn-prefix-advertisement]. Route Types 6, 7 and 8 are defined in [I-D.ietf-bess-evpn-igmp-mld-proxy].

- o Ethernet Auto-discovery Route (Route Type 1)
- o MAC/IP Advertisement Route (Route Type 2)
- o Inclusive Multicast Ethernet Tag Route (Route Type 3)
- o Ethernet Segment route (Route Type 4)
- o IP prefix route (Route Type 5)
- o Selective Multicast Ethernet Tag route (Route Type 6)
- o IGMP join sync route (Route Type 7)
- o IGMP leave sync route (Route Type 8)

To support SRv6 based EVPN overlays, one or more SRv6 Service SIDs are advertised with Route Type 1,2,3 and 5. The SRv6 Service SID(s) per Route Type are advertised in SRv6 L3/L2 Service TLVs within the

BGP SID attribute. Signaling of SRv6 Service SID(s) serves two purposes – first, it indicates that the BGP egress device is reachable via an SRv6 underlay and the BGP ingress device receiving this route MAY choose to encapsulate or insert an SRv6 SRH; second, it indicates the value of the SID(s) to include in the SRH encapsulation. If the BGP egress device does not support MPLS based EVPN services, the MPLS Label fields within EVPN Route Types MUST be set to Implicit NULL.

4.1. Ethernet Auto-discovery route over SRv6 Core

Ethernet Auto-Discovery (A-D) routes are Route Type 1 defined in [RFC7432] and may be used to achieve split horizon filtering, fast convergence and aliasing. EVPN Route Type 1 is also used in EVPN-VPWS as well as in EVPN flexible cross-connect; mainly used to advertise point-to-point services ID.

Multi-homed PEs MAY advertise an Ethernet Auto-Discovery route per Ethernet segment along with the ESI Label extended community defined in [RFC7432]. The extended community label MUST be set to Implicit NULL. PEs may identify other PEs connected to the same Ethernet segment after the EVPN Route Type 4 ES route exchange. All the multi-homed and remote PEs that are part of same EVI may import the Auto-Discovery route.

EVPN Route Type 1 is encoded as follows for SRv6 Core:

```

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

```

4.1.1. Per-ES A-D route

- o BGP next-hop: IPv6 address of an egress PE
- o Ethernet Tag ID: set to 0xFFFF
- o MPLS Label: always set to zero value
- o Extended Community: Per ES AD, ESI label extended community

A Service SID enclosed in a SRv6 L2 Service TLV within the BGP SID attribute is advertised along with the A-D route. The behavior of the Service SID thus signaled is entirely up to the originator of the advertisement. This is typically used to signal Arg.FE2 SID argument for applicable End.DT2M SIDs.

4.1.2. Per-EVI A-D route

- o BGP next-hop: IPv6 address of an egress PE
- o Ethernet Tag ID: non-zero for VLAN aware bridging, EVPN VPWS and FXC
- o MPLS Label: Implicit NULL

A Service SID enclosed in a SRv6 L2 Service TLV within the BGP SID attribute is advertised along with the A-D route. The behavior of the Service SID thus signaled is entirely up to the originator of the advertisement. In practice, the behavior would likely be END.DX2, END.DX2V or END.DT2U.

4.2. MAC/IP Advertisement route over SRv6 Core

EVPN Route Type 2 is used to advertise unicast traffic MAC+IP address reachability through MP-BGP to all other PEs in a given EVPN instance.

EVPN Route Type 2 is encoded as follows for SRv6 Core:

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

- o BGP next-hop: IPv6 address of an egress PE
- o MPLS Label1: Implicit NULL
- o MPLS Label2: Implicit NULL

Service SIDs enclosed in SRv6 L2 Service TLV and optionally in SRv6 L3 Service TLV within the BGP SID attribute is advertised along with the MAC/IP Advertisement route.

Described below are different types of Route Type 2 advertisements.

- o MAC/IP Advertisement route with MAC Only
 - * BGP next-hop: IPv6 address of egress PE
 - * MPLS Label1: Implicit NULL
 - * MPLS Label2: Implicit NULL
- o A Service SID enclosed in a SRv6 L2 Service TLV within the BGP SID attribute is advertised along with the route. The behavior of the Service SID thus signaled is entirely up to the originator of the advertisement. In practice, the behavior would likely be END.DX2 or END.DT2U.
- o MAC/IP Advertisement route with MAC+IP
 - * BGP next-hop: IPv6 address of egress PE
 - * MPLS Label1: Implicit NULL
 - * MPLS Label2: Implicit NULL
- o An L2 Service SID enclosed in a SRv6 L2 Service TLV within the BGP SID attribute is advertised along with the route. In addition, an L3 Service SID enclosed in a SRv6 L3 Service TLV within the BGP SID attribute MAY also be advertised along with the route. The behavior of the Service SID(s) thus signaled is entirely up to the originator of the advertisement. In practice, the behavior would likely be END.DX2 or END.DT2U for the L2 Service SID, and END.DT6/4 or END.DX6/4 for the L3 Service SID.

4.3. Inclusive Multicast Ethernet Tag Route over SRv6 Core

EVPN Route Type 3 is used to advertise multicast traffic reachability information through MP-BGP to all other PEs in a given EVPN instance.

EVPN Route Type 3 is encoded as follows for SRv6 core:

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

- o BGP next-hop: IPv6 address of egress PE

PMSI Tunnel Attribute [RFC6514] MAY contain MPLS Implicit NULL label and Tunnel Type would be similar to that defined in EVPN Route Type 6 i.e. Ingress replication route.

The format of PMSI Tunnel Attribute attribute is encoded as follows for SRv6 Core:

Flag (1 octet)
Tunnel Type (1 octet)
MPLS label (3 octet)
Tunnel Identifier (variable)

- o Flag: zero value defined per [RFC7432]
- o Tunnel Type: defined per [RFC6514]
- o MPLS label: Implicit NULL
- o Tunnel Identifier: IP address of egress PE

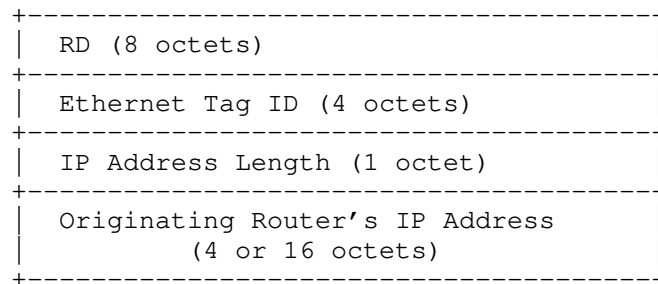
A Service SID enclosed in a SRv6 L2 Service TLV within the BGP SID attribute is advertised along with the route. The behavior of the Service SID thus signaled, is entirely up to the originator of the advertisement. In practice, the behavior of the SRv6 SID is as follows:

- o END.DX2 or END.DT2M function

- o The ESI Filtering argument (Arg.FE2) of the Service SID carried along with EVPN Route Type 1 route MAY be merged together with the applicable End.DT2M SID of Type 3 route advertised by remote PE by doing a bitwise logical-OR operation to create a single SID on the ingress PE for Split-horizon and other filtering mechanisms. Details of filtering mechanisms are described in [RFC7432].

4.4. Ethernet Segment route over SRv6 Core

An Ethernet Segment route i.e. EVPN Route Type 4 is encoded as follows for SRv6 core:



- o BGP next-hop: IPv6 address of egress PE

SRv6 Service TLVs within BGP SID attribute are not advertised along with this route. The processing of the route has not changed - it remains as described in [RFC7432].

4.5. IP prefix route over SRv6 Core

EVPN Route Type 5 is used to advertise IP address reachability through MP-BGP to all other PEs in a given EVPN instance. IP address may include host IP prefix or any specific subnet.

EVPN Route Type 5 is encoded as follows for SRv6 core:

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

- o BGP next-hop: IPv6 address of egress PE
- o MPLS Label: Implicit NULL

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DT6/4 or End.DX6/4.

4.6. EVPN multicast routes (Route Types 6, 7, 8) over SRv6 core

These routes do not require the advertisement of SRv6 Service TLVs along with them. Similar to EVPN Route Type 4, the BGP Nexthop is equal to the IPv6 address of egress PE. More details may be added in future revisions of this document.

5. Migration from MPLS based Segment Routing to SRv6 Segment Routing

Migration from IPv4 to IPv6 is independent of SRv6 BGP endpoints, and the selection of which route to use (received via the IPv4 or the IPv6 session) is a local configurable decision of the ingress PE, and is outside the scope of this document.

Migration from MPLS based underlay to an SRv6 underlay with BGP speakers is achieved with a few simple rules at each BGP speaker.

At Egress PE:

If BGP offers an SRv6 service, then:

BGP allocates an SRv6 Service SID for the L3 service and includes it in the BGP SRv6 L3 Service TLV while advertising the overlay prefixes.

If BGP offers an MPLS service, then:

BGP allocates an MPLS Label for the L3 service and encode it as part of the NLRI as normal for MPLS based address-families; else, the MPLS label value for the L3 service is set to Implicit NULL.

At Ingress PE:

Selection of either MPLS encapsulation or SRv6 encapsulation is defined by local BGP policy.

If BGP supports SRv6 based services and receives overlay routes with BGP SID attribute containing SRv6 L3 Service TLV(s) encoding SRv6 Service SID(s), then:

BGP programs the destination prefix in RIB recursive via the related SR Policy.

If BGP supports MPLS service, and the MPLS Label value is not Implicit NULL, then:

the MPLS label is used as the overlay service label and inserted with the prefix into RIB via the BGP Nexthop.

6. Implementation Status

The SRv6 Service is available for SRv6 on various Cisco hardware and other software platforms. An end-to-end integration of SRv6 L3VPN, SRv6 Traffic-Engineering and Service Chaining. All of that with data-plane interoperability across <<http://www.segment-routing.net>> different implementations:

- o Three Cisco Hardware-forwarding platforms: ASR 1K, ASR 9k and NCS 5500
- o Two Cisco network operating systems: IOS XE and IOS XR
- o Huawei Hardware-forwarding platforms: ATN, CX, ME, NE5000E, NE9000, NG-OLT
- o Huawei network operating systems: VRPv8
- o Barefoot Networks Tofino on OCP Wedge-100BF
- o Linux Kernel officially upstreamed in 4.10
- o Fd.io

7. Error Handling

In case of any errors encountered while processing SRv6 Service TLVs, the details of the error SHOULD be logged for further analysis.

If multiple instances of SRv6 L3 Service TLV is encountered, all but the first instance MUST be ignored.

If multiple instances of SRv6 L2 Service TLV is encountered, all but the first instance MUST be ignored.

An SRv6 Service TLV is considered malformed in the following cases:

- o the TLV Length is less than 1
- o the TLV Length is inconsistent with the length of BGP SID attribute
- o atleast one of the constituent Sub-TLVs is malformed

An SRv6 Service Sub-TLV is considered malformed in the following cases:

- o the Sub-TLV Length is inconsistent with the length of the enclosing SRv6 Service TLV

An SRv6 SID Information Sub-TLV is considered malformed in the following cases:

- * the Sub-TLV Length is less than 21
- * the Sub-TLV Length is inconsistent with the length of the enclosing SRv6 Service TLV
- * atleast one of the constituent Sub-Sub-TLVs is malformed

An SRv6 Service Data Sub-sub-TLV is considered malformed in the following cases:

- o the Sub-Sub-TLV Length is inconsistent with the length of the enclosing SRv6 service Sub-TLV

Any TLV or Sub-TLV or Sub-Sub-TLV is not considered malformed because its Type is unrecognized.

Any TLV or Sub-TLV or Sub-Sub-TLV is not considered malformed because of failing any semantic validation of its Value field.

The BGP SID attribute is considered malformed if it contains atleast one constituent SRv6 Service TLV that is malformed. In such cases, the attribute MUST be discarded [RFC7606] and not propagated further. Note that if a path whose BGP SID attribute is discarded in this manner is selected as the best path to be installed in the RIB, traffic forwarding for the corresponding prefix may be affected. Implementations MAY choose to make such paths less preferable or even ineligible during the selection of best path for the corresponding prefix.

A BGP speaker receiving a path containing BGP SID attribute with one or more SRv6 Service TLVs observes the following rules when advertising the received path to other peers:

- o if the nexthop is unchanged during advertisement, the SRv6 Service TLVs, including any unrecognized Types of Sub-TLV and Sub-Sub-TLV, SHOULD be propagated further. In addition, all Reserved fields in the TLV or Sub-TLV or Sub-Sub-TLV MUST be propagated unchanged.
- o if the nexthop is changed during advertisement, any unrecognized Sub-TLVs and Sub-Sub-TLVs MUST NOT be propagated.
- o if the nexthop is changed during advertisement, the TLVs, Sub-TLVs and Sub-Sub-TLVs SHOULD be re-originated if appropriate, and not merely propagated unchanged. The interpretation of the meaning of re-origination versus propagation is a matter of local implementation.

A received VPN NLRI [RFC4364][RFC4659][RFC7432] that has neither a valid MPLS label nor a valid SRv6 Service TLV MUST be considered unreachable i.e. apply the -treat as withdraw- action specified in [RFC7606].

8. IANA Considerations

8.1. BGP Prefix-SID TLV Types registry

This document defines two new TLV Types of the BGP Prefix-SID attribute. IANA is requested to assign Type values in the registry "BGP Prefix-SID TLV Types" as follows:

Value	Type	Reference

[TBD1]	SRv6 L3 Service TLV	<this document>
[TBD2]	SRv6 L2 Service TLV	<this document>

IANA is also requested to reserve the following Type value. This was used in some implementations of previous versions of this draft.

Value	Type	Reference

4	Reserved	<this document>

8.2. SRv6 Service Sub-TLV Types registry

IANA is requested to create and maintain a new registry called "SRv6 Service Sub-TLV Types". The allocation policy for this registry is:

0 : Reserved
1-127 : IETF Review
128-254 : First Come First Served
255 : Reserved

The following Sub-TLV Types are defined in this document:

Value	Type	Reference

1	SRv6 SID Information Sub-TLV	<this document>

9. Security Considerations

This document introduces no new security considerations beyond those already specified in [RFC4271] and [RFC8277].

10. Conclusions

This document proposes extensions to the BGP to allow advertising certain attributes and functionalities related to SRv6.

11. References

11.1. Normative References

[I-D.filsfils-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., Hegde, S.,
daniel.voyer@bell.ca, d., Lin, S., bogdanov@google.com,
b., Krol, P., Horneffer, M., Steinberg, D., Decraene, B.,
Litkowski, S., Mattes, P., Ali, Z., Talaulikar, K., Liste,
J., Clad, F., and K. Raza, "Segment Routing Policy
Architecture", draft-filsfils-spring-segment-routing-
policy-06 (work in progress), May 2018.

- [I-D.filsfils-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J.,
daniel.voyer@bell.ca, d., Matsushima, S., and Z. Li, "SRv6
Network Programming", draft-filsfils-spring-srv6-network-
programming-07 (work in progress), February 2019.
- [I-D.ietf-6man-segment-routing-header]
Filsfils, C., Previdi, S., Leddy, J., Matsushima, S., and
d. daniel.voyer@bell.ca, "IPv6 Segment Routing Header
(SRH)", draft-ietf-6man-segment-routing-header-16 (work in
progress), February 2019.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6
(IPv6) Specification", RFC 2460, DOI 10.17487/RFC2460,
December 1998, <<https://www.rfc-editor.org/info/rfc2460>>.
- [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>>.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP
Encodings and Procedures for Multicast in MPLS/BGP IP
VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012,
<<https://www.rfc-editor.org/info/rfc6514>>.
- [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>>.
- [RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address
Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017,
<<https://www.rfc-editor.org/info/rfc8277>>.

11.2. 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-02 (work in progress), June 2018.

- [I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.
- [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-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Jain, D., Mattes, P., Rosen, E., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-05 (work in progress), November 2018.
- [I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Ginsberg, L., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-22 (work in progress), December 2018.
- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <<https://www.rfc-editor.org/info/rfc3032>>.
- [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>>.
- [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>>.

- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [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>>.

Appendix A. Contributors

Bart Peirens
Proximus
Belgium

Email: bart.peirens@proximus.com

Authors' Addresses

Gaurav Dawra (editor)
LinkedIn
USA

Email: gdawra.ietf@gmail.com

Clarence Filsfils
Cisco Systems
Belgium

Email: cfilsfil@cisco.com

Darren Dukes
Cisco Systems
Canada

Email: ddukes@cisco.com

Patrice Brissette
Cisco Systems
Canada

Email: pbrisset@cisco.com

Shyam Sethuram
Cisco Systems
USA

Email: shsethur@cisco.com

Pablo Camarilo
Cisco Systems
Spain

Email: pcamaril@cisco.com

Jonn Leddy
Comcast
USA

Email: john_leddy@cable.comcast.com

Daniel Voyer
Bell Canada
Canada

Email: daniel.voyer@bell.ca

Daniel Bernier
Bell Canada
Canada

Email: daniel.bernier@bell.ca

Dirk Steinberg
Steinberg Consulting
Germany

Email: dws@steinberg.net

Robert Raszuk
Bloomberg LP
USA

Email: robert@raszuk.net

Bruno Decraene
Orange
France

Email: bruno.decraene@orange.com

Satoru Matsushima
SoftBank
1-9-1, Higashi-Shimbashi, Minato-Ku
Japan 105-7322

Email: satoru.matsushima@g.softbank.co.jp

Shunwan Zhuang
Huawei Technologies
China

Email: zhuangshunwan@huawei.com

BESS Working Group
Internet Draft
Intended Status: Proposed Standard
Expires: April 25, 2019

P. Brissette Ed.
Cisco System
H. Shah Ed.
Ciena Corporation
I. Chen Ed.
Jabil
I. Hussain Ed.
Infinera Corporation
K. Tiruveedhula Ed.
Juniper Networks
J. Rabadan Ed.
Nokia

October 22, 2018

Yang Data Model for EVPN
draft-ietf-bess-evpn-yang-06

Abstract

This document describes a YANG data model for Ethernet VPN services. The model is agnostic of the underlay. It apply to MPLS as well as to VxLAN encapsulation. The model is also agnostic of the services including E-LAN, E-LINE and E-TREE services. This document mainly focuses on EVPN and Ethernet-Segment instance framework.

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.

Convention

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].

Table of Contents

1. Introduction	2
2. Specification of Requirements	3
3. EVPN YANG Model	4
3.1. Overview	4
3.2 Ethernet-Segment Model	4
3.3 EVPN Model	5
4. YANG Module	9
4.1 Ethernet Segment Yang Module	9
4.2 EVPN Yang Module	14
5. Security Considerations	25
6. IANA Considerations	26
7. References	26
7.1. Normative Reference	26
7.2. Informative References	26
Authors' Addresses	27

1. Introduction

The Network Configuration Protocol (NETCONF) [RFC6241] is a network management protocol that defines mechanisms to manage network devices. YANG [RFC6020] is a modular language that represents data structures in an XML or JSON tree format, and is used as a data modeling language for the NETCONF.

This document introduces a YANG data model for Ethernet VPN services (EVPN) [RFC7432], Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN) [RFC7623] as well as other WG draft such as EVPN-VPWS, etc. The EVPN services runs over MPLS and VxLAN underlay.

The Yang data model in this document defines Ethernet VPN based services. The model leverages the definitions used in other IETF Yang draft such as L2VPN Yang.

The goal is to propose a data object model consisting of building blocks that can be assembled in different order to realize different EVPN-based services. The definition work is undertaken initially by a smaller working group with members representing various vendors and service providers. The EVPN basic framework consist of two modules: EVPN and Ethernet-Segment. These models are completely orthogonal. They usually work in pair but user can definitely use one or the other for its own need.

The data model is defined for following constructs that are used for managing the services:

- o Configuration
- o Operational State
- o Notifications

The document is organized to first define the data model for the configuration, operational state and notifications of EVPN and Ethernet-Segment.

The EVPN data object model defined in this document uses the instance centric approach whereby EVPN service attributes are specified for a given EVPN instance.

The Ethernet-Segment data object model defined in this document refer to a specific interface. That interface can be a physical interface, a bundle interface or virtual interface. The latter includes attachment-circuit and pseudowire. The purpose of creating a separate module is due to the fact that it can be used without having the need to have EVPN configured as layer 2/3 service. For example, an access node can be dual-homed to two service nodes servicing a VPLS or an IPVPN core. The access connectivity can be represented by an Ethernet-Segment where EVPN BGP DF election is performed over both service nodes.

2. Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL

NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. EVPN YANG Model

3.1. Overview

Two top level module, Ethernet-Segment and EVPN, are defined. The Ethernet-Segment contains a list of interface to which any Ethernet-Segment attributes are configured/applied.

The EVPN module has two main containers: common and instance. The first one has common attributes to all VPNs where as the latter has attributes specific to an EVI (EVPN instance). This document state the scope of the EVPN object models definition. The following documents are within the scope. This is not an exhaustive list but a representation of documents that are covered for this work:

- o Reqs for EVPN:[RFC7209]
- o EVPN: [RFC7432]
- o PBB-EVPN: [RFC7623]
- o EVPN-VPWS: [RFC8214]
- o EVPN-ETREE: [RFC8317]
- o EVPN Overlay [RFC8365]

The integration with L2VPN instance Yang model is being done as part of the L2VPN Yang model.

Following documents will be covered at that time:

- o (PBB-)EVPN Seamless Integration with (PBB-)VPLS:
draft-ietf-bess-evpn-vpls-seamless-integ
- o EVPN Virtual Ethernet Segment:
draft-sajassi-bess-evpn-virtual-eth-segment
- o IP Prefix Advertisement in EVPN:
draft-ietf-bess-evpn-prefix-advertisement
- o VXLAN DCI Using EVPN:
draft-boutros-l2vpn-vxlan-evpn
- o Interconnect Solution for EVPN Overlay networks:
draft-ietf-bess-dci-evpn-overlay
- o Integrated Routing and Bridging in EVPN:
draft-ietf-bess-evpn-inter-subnet-forwarding

3.2 Ethernet-Segment Model

The Ethernet-Segment data model has a list of ES where each refer to an interface. All attributes are optional due to auto-sensing default mode where all values are auto-derive from the network connectivity.

module: ietf-ethernet-segment

```

+--rw ethernet-segments
  +--rw ethernet-segment* [name]
    +--rw name string
    +--ro service-type? string
    +--ro status? status-type
    +--rw (ac-or-pw)?
      | +--:(ac)
      | | +--rw ac* if:interface-ref
      | +--:(pw)
      | | +--rw pw* pw:pseudowire-ref
    +--ro interface-status? status-type
    +--rw ethernet-segment-identifier? uint32
    +--rw (active-mode)
      | +--:(single-active)
      | | +--rw single-active-mode? empty
      | +--:(all-active)
      | | +--rw all-active-mode? empty
    +--rw pbb-parameters {ethernet-segment-pbb-params}?
      | +--rw backbone-src-mac? yang:mac-address
    +--rw bgp-parameters
      | +--rw common
      | | +--rw rd-rt* [route-distinguisher]
      | | | {ethernet-segment-bgp-params}?
      | | | +--rw route-distinguisher
      | | | | rt-types:route-distinguisher
      | | +--rw vpn-target* [route-target]
      | | | +--rw route-target
      | | | | rt-types:route-target
      | | | +--rw route-target-type
      | | | | rt-types:route-target-type
    +--rw df-election
      | +--rw df-election-method? df-election-method-type
      | +--rw preference? uint16
      | +--rw revertive? boolean
      | +--rw election-wait-time? uint32
    +--rw ead-evi-route? boolean
    +--ro esi-label? string
    +--ro member*
      | +--ro ip-address? inet:ip-address
    +--ro df*
      | +--ro service-identifier? uint32
      | +--ro vlan? uint32
      | +--ro ip-address? inet:ip-address

```

3.3 EVPN Model

The evpn-instances container contains a list of evpn-instance. Each entry of the evpn-instance represents a different Ethernet VPN and it

is represented by a EVI. Again, mainly all attributes are optional for the same reason as for the Ethernet-Segment module.

```

module: ietf-evpn
  +--rw evpn
    +--rw common
      +--rw (replication-type)?
        +--:(ingress-replication)
          | +--rw ingress-replication?   boolean
        +--:(p2mp-replication)
          | +--rw p2mp-replication?      boolean
    +--rw evpn-instances
      +--rw evpn-instance* [name]
        +--rw name                               string
        +--rw evi?                               uint32
        +--rw pbb-parameters {evpn-pbb-params}?
          | +--rw source-bmac?   yang:hex-string
        +--rw bgp-parameters
          +--rw common
            +--rw rd-rt* [route-distinguisher]
              {evpn-bgp-params}?
            +--rw route-distinguisher
              | rt-types:route-distinguisher
            +--rw vpn-target* [route-target]
              +--rw route-target
                rt-types:route-target
            +--rw route-target-type
              rt-types:route-target-type
          +--rw arp-proxy?                       boolean
          +--rw arp-suppression?                 boolean
          +--rw nd-proxy?                       boolean
          +--rw nd-suppression?                 boolean
          +--rw underlay-multicast?             boolean
          +--rw flood-unknown-unicast-supression? boolean
          +--rw vpws-vlan-aware?               boolean
        +--ro routes
          +--ro ethernet-auto-discovery-route*
            +--ro rd-rt* [route-distinguisher]
              +--ro route-distinguisher
                rt-types:route-distinguisher
              +--ro vpn-target* [route-target]
                +--ro route-target      rt-types:route-target
          +--ro ethernet-segment-identifier?   uint32
          +--ro ethernet-tag?                  uint32
          +--ro path*
            +--ro next-hop?   inet:ip-address
            +--ro label?     rt-types:mpls-label
            +--ro detail

```

```

    +--ro attributes
    |   +--ro extended-community*   string
    |   +--ro bestpath?             empty
+--ro mac-ip-advertisement-route*
+--ro rd-rt* [route-distinguisher]
|   +--ro route-distinguisher
|       rt-types:route-distinguisher
+--ro vpn-target* [route-target]
|   +--ro route-target
|       rt-types:route-target
+--ro ethernet-segment-identifier?   uint32
+--ro ethernet-tag?                   uint32
+--ro mac-address?                    yang:hex-string
+--ro mac-address-length?             uint8
+--ro ip-prefix?                      inet:ip-prefix
+--ro path*
|   +--ro next-hop?   inet:ip-address
|   +--ro label?      rt-types:mpls-label
|   +--ro label2?     rt-types:mpls-label
|   +--ro detail
|   +--ro attributes
|   |   +--ro extended-community*   string
|   |   +--ro bestpath?             empty
+--ro inclusive-multicast-ethernet-tag-route*
+--ro rd-rt* [route-distinguisher]
|   +--ro route-distinguisher
|       rt-types:route-distinguisher
+--ro vpn-target* [route-target]
|   +--ro route-target
|       rt-types:route-target
+--ro ethernet-segment-identifier?   uint32
+--ro originator-ip-prefix?          inet:ip-prefix
+--ro path*
|   +--ro next-hop?   inet:ip-address
|   +--ro label?      rt-types:mpls-label
|   +--ro detail
|   +--ro attributes
|   |   +--ro extended-community*   string
|   |   +--ro bestpath?             empty
+--ro ethernet-segment-route*
+--ro rd-rt* [route-distinguisher]
|   +--ro route-distinguisher
|       rt-types:route-distinguisher
+--ro vpn-target* [route-target]
|   +--ro route-target
|       rt-types:route-target
+--ro ethernet-segment-identifier?   uint32
+--ro originator-ip-prefix?          inet:ip-prefix

```

```

    +--ro path*
      +--ro next-hop?  inet:ip-address
      +--ro detail
        +--ro attributes
          | +--ro extended-community*  string
          +--ro bestpath?  empty
    +--ro ip-prefix-route*
      +--ro rd-rt* [route-distinguisher]
        +--ro route-distinguisher
          rt-types:route-distinguisher
        +--ro vpn-target* [route-target]
          +--ro route-target  rt-types:route-target
      +--ro ethernet-segment-identifier?  uint32
      +--ro ip-prefix?  inet:ip-prefix
      +--ro path*
        +--ro next-hop?  inet:ip-address
        +--ro label?  rt-types:mpls-label
        +--ro detail
          +--ro attributes
            | +--ro extended-community*  string
            +--ro bestpath?  empty
    +--ro statistics
      +--ro tx-count?  uint32
      +--ro rx-count?  uint32
      +--ro detail
        +--ro broadcast-tx-count?  uint32
        +--ro broadcast-rx-count?  uint32
        +--ro multicast-tx-count?  uint32
        +--ro multicast-rx-count?  uint32
        +--ro unknown-unicast-tx-count?  uint32
        +--ro unknown-unicast-rx-count?  uint32
augment /pw:pseudowires/pw:pseudowire/pw:pw-type:
  +--:(evpn-pw)
    +--rw evpn-pw
      +--rw remote-id?  uint32
      +--rw local-id?  uint32
augment
/nl:network-instances/nl:network-instance/nl:nl-type/l2vpn:l2vpn:
  +--rw evpn-instance?  evpn-instance-ref
augment
/nl:network-instances/nl:network-instance/nl:nl-type/l2vpn:l2vpn:
  +--rw vpls-contraints

notifications:
  +---n evpn-state-change-notification
    +--ro evpn-instance?  evpn-instance-ref
    +--ro state?  identityref

```

4. YANG Module

The EVPN configuration container is logically divided into following high level configuration areas:

4.1 Ethernet Segment Yang Module

```
<CODE BEGINS> file "ietf-ethernet-segment@2018-02-20.yang"
module iETF-ethernet-segment {
  namespace "urn:ietf:params:xml:ns:yang:ietf-ethernet-segment";
  prefix "es";

  import iETF-yang-types {
    prefix "yang";
  }

  import iETF-inet-types {
    prefix "inet";
  }

  import iETF-routing-types {
    prefix "rt-types";
  }

  import iETF-interfaces {
    prefix "if";
  }

  import iETF-pseudowires {
    prefix "pw";
  }

  organization "ietf";
  contact "ietf";
  description "ethernet segment";

  revision "2018-02-20" {
    description " - Change the type of attachment circuit to " +
               " if:interface-ref " +
               "";
    reference "";
  }

  revision "2017-10-21" {
    description " - Updated ethernet segment's AC/PW members to " +
               " accommodate more than one AC or more than one " +
               " PW " +
               " - Added the new preference based DF election " +
               "
```

```
        "    method " +
        " - Referenced pseudowires in the new " +
        "    ietf-pseudowires.yang model " +
        " - Moved model to NMDA style specified in " +
        "    draft-dsdt-nmda-guidelines-01.txt " +
        " ";
    reference " ";
}

revision "2017-03-08" {
    description " - Updated to use BGP parameters from " +
        "    ietf-routing-types.yang instead of from " +
        "    ietf-evpn.yang " +
        " - Updated ethernet segment's AC/PW members to " +
        "    accommodate more than one AC or more than one " +
        "    PW " +
        " - Added the new preference based DF election " +
        "    method " +
        " ";
    reference " ";
}

revision "2016-07-08" {
    description " - Added the configuration option to enable or " +
        "    disable per-EVI/EAD route " +
        " - Added PBB parameter backbone-src-mac " +
        " - Added operational state branch, initially " +
        "    to match the configuration branch" +
        " ";
    reference " ";
}

revision "2016-06-23" {
    description "WG document adoption";
    reference " ";
}

revision "2015-10-15" {
    description "Initial revision";
    reference " ";
}

/* Features */

feature ethernet-segment-bgp-params {
    description "Ethernet segment's BGP parameters";
}
```



```
feature ethernet-segment-pbb-params {
  description "Ethernet segment's PBB parameters";
}

/* Typedefs */
typedef status-type {
  type enumeration {
    enum up {
      description "Status is up";
    }
    enum down {
      description "Status is down";
    }
  }
  description "status type";
}

typedef df-election-method-type {
  type enumeration {
    enum default {
      value 0;
      description "The default DF election method";
    }
    enum highest-random-weight {
      value 1;
      description "The highest random weight (HRW) method";
      reference "draft-mohanty-bess-evpn-df-election";
    }
    enum preference {
      value 2;
      description "The preference based method";
      reference "draft-rabadan-bess-evpn-pref-df";
    }
  }
  description "The DF election method type";
}

/* EVPN Ethernet Segment YANG Model */

container ethernet-segments {
  description "ethernet-segment";
  list ethernet-segment {
    key "name";
    leaf name {
      type string;
      description "Name of the ethernet segment";
    }
    leaf service-type {
```

```
    type string;
    config false;
    description "service-type";
  }
  leaf status {
    type status-type;
    config false;
    description "Ethernet segment status";
  }
  choice ac-or-pw {
    description "ac-or-pw";
    case ac {
      leaf-list ac {
        type if:interface-ref;
        description "Name of attachment circuit";
      }
    }
    case pw {
      leaf-list pw {
        type pw:pseudowire-ref;
        description "Reference to a pseudowire";
      }
    }
  }
  leaf interface-status {
    type status-type;
    config false;
    description "interface status";
  }
  leaf ethernet-segment-identifier {
    type uint32;
    description "Ethernet segment identifier (esi)";
  }
  choice active-mode {
    mandatory true;
    description "Choice of active mode";
    case single-active {
      leaf single-active-mode {
        type empty;
        description "single-active-mode";
      }
    }
    case all-active {
      leaf all-active-mode {
        type empty;
        description "all-active-mode";
      }
    }
  }
}
```

```
}
container pbb-parameters {
  if-feature ethernet-segment-pbb-params;
  description "PBB configuration";
  leaf backbone-src-mac {
    type yang:mac-address;
    description "backbone-src-mac, only if this is a PBB";
  }
}
container bgp-parameters {
  description "BGP parameters";
  container common {
    description "BGP parameters common to all pseudowires";
    list rd-rt {
      if-feature ethernet-segment-bgp-params;
      key "route-distinguisher";
      leaf route-distinguisher {
        type rt-types:route-distinguisher;
        description "Route distinguisher";
      }
      uses rt-types:vpn-route-targets;
      description "A list of route distinguishers and " +
        "corresponding VPN route targets";
    }
  }
}
container df-election {
  description "df-election";
  leaf df-election-method {
    type df-election-method-type;
    description "The DF election method";
  }
  leaf preference {
    when "../df-election-method = 'preference'" {
      description "The preference value is only applicable " +
        "to the preference based method";
    }
    type uint16;
    description "The DF preference";
  }
  leaf revertive {
    when "../df-election-method = 'preference'" {
      description "The revertive value is only applicable " +
        "to the preference method";
    }
    type boolean;
    default true;
    description "The 'preempt' or 'revertive' behavior";
  }
}
```

```
    }
    leaf election-wait-time {
        type uint32;
        description "election-wait-time";
    }
}
leaf ead-evi-route {
    type boolean;
    default false;
    description "Enable (true) or disable (false) ead-evi-route";
}
leaf esi-label {
    type string;
    config false;
    description "esi-label";
}
list member {
    config false;
    leaf ip-address {
        type inet:ip-address;
        description "ip-address";
    }
    description "member of the ethernet segment";
}
list df {
    config false;
    leaf service-identifier {
        type uint32;
        description "service-identifier";
    }
    leaf vlan {
        type uint32;
        description "vlan";
    }
    leaf ip-address {
        type inet:ip-address;
        description "ip-address";
    }
    description "df of an evpn instance's vlan";
}
description "An ethernet segment";
}
}
}
<CODE ENDS>
```

4.2 EVPN Yang Module

```
<CODE BEGINS> file "ietf-evpn@2018-02-20.yang"
module ietf-evpn {
  namespace "urn:ietf:params:xml:ns:yang:ietf-evpn";
  prefix "evpn";

  import ietf-inet-types {
    prefix "inet";
  }

  import ietf-yang-types {
    prefix "yang";
  }

  import ietf-routing-types {
    prefix "rt-types";
  }

  import ietf-network-instance {
    prefix "ni";
  }

  import ietf-l2vpn {
    prefix "l2vpn";
  }

  import ietf-pseudowires {
    prefix "pw";
  }

  organization "ietf";
  contact "ietf";
  description "evpn";

  revision "2018-02-20" {
    description " - Incorporated ietf-network-instance model" +
      " - on which ietf-l2vpn is now based " +
      "";
    reference "";
  }

  revision "2017-10-21" {
    description " - Modified the operational state augment " +
      " - Renamed evpn-instances-state to evpn-instances" +
      " - Added vpws-vlan-aware to an EVPN instance " +
      " - Added a new augment to L2VPN to add EPVN " +
      " - pseudowire for the case of EVPN VPWS " +
      " - Added state change notification " +
      "";
  }
}
```

```
    reference    "";
  }

  revision "2017-03-13" {
    description " - Added an augment to base L2VPN model to " +
               "   reference an EVPN instance " +
               " - Reused ietf-routing-types.yang " +
               "   vpn-route-targets grouping instead of " +
               "   defining it in this module " +
               "";
    reference    "";
  }

  revision "2016-07-08" {
    description " - Added operational state" +
               " - Added a configuration knob to enable/disable " +
               "   underlay-multicast " +
               " - Added a configuration knob to enable/disable " +
               "   flooding of unknow unicast " +
               " - Added several configuration knobs " +
               "   to manage ARP and ND" +
               "";
    reference    "";
  }

  revision "2016-06-23" {
    description "WG document adoption";
    reference    "";
  }

  revision "2015-10-15" {
    description "Initial revision";
    reference    "";
  }

  feature evpn-bgp-params {
    description "EVPN's BGP parameters";
  }

  feature evpn-pbb-params {
    description "EVPN's PBB parameters";
  }

  /* Identities */

  identity evpn-notification-state {
    description "The base identity on which EVPN notification " +
               "states are based";
```

```
}

identity MAC-duplication-detected {
  base "evpn-notification-state";
  description "MAC duplication is detected";
}

identity mass-withdraw-received {
  base "evpn-notification-state";
  description "Mass withdraw received";
}

identity static-MAC-move-detected {
  base "evpn-notification-state";
  description "Static MAC move is detected";
}

/* Typedefs */

typedef evpn-instance-ref {
  type leafref {
    path "/evpn/evpn-instances/evpn-instance/name";
  }
  description "A leafref type to an EVPN instance";
}

/* Groupings */

grouping route-rd-rt-grp {
  description "A grouping for a route's route distinguishers " +
    "and route targets";
  list rd-rt {
    key "route-distinguisher";
    leaf route-distinguisher {
      type rt-types:route-distinguisher;
      description "Route distinguisher";
    }
    list vpn-target {
      key "route-target";
      leaf route-target {
        type rt-types:route-target;
        description "BGP route target";
      }
    }
    description "A list of route targets";
  }
  description "A list of route distinguishers and " +
    "corresponding VPN route targets";
}
```

```
    }

    grouping next-hop-label-grp {
      description "next-hop-label-grp";
      leaf next-hop {
        type inet:ip-address;
        description "next-hop";
      }
      leaf label {
        type rt-types:mpls-label;
        description "label";
      }
    }

    grouping next-hop-label2-grp {
      description "next-hop-label2-grp";
      leaf label2 {
        type rt-types:mpls-label;
        description "label2";
      }
    }

    grouping path-detail-grp {
      description "path-detail-grp";
      container detail {
        config false;
        description "path details";
        container attributes {
          leaf-list extended-community {
            type string;
            description "extended-community";
          }
          description "attributes";
        }
        leaf bestpath {
          type empty;
          description "Indicate this path is the best path";
        }
      }
    }

    /* EVPN YANG Model */

    container evpn {
      description "evpn";
      container common {
        description "common evpn attributes";
        choice replication-type {
```



```
    description "A choice of replication type";
    case ingress-replication {
      leaf ingress-replication {
        type boolean;
        description "ingress-replication";
      }
    }
    case p2mp-replication {
      leaf p2mp-replication {
        type boolean;
        description "p2mp-replication";
      }
    }
  }
}
container evpn-instances {
  description "evpn-instances";
  list evpn-instance {
    key "name";
    description "An EVPN instance";
    leaf name {
      type string;
      description "Name of EVPN instance";
    }
    leaf evi {
      type uint32;
      description "evi";
    }
    container pbb-parameters {
      if-feature "evpn-pbb-params";
      description "PBB parameters";
      leaf source-bmac {
        type yang:hex-string;
        description "source-bmac";
      }
    }
  }
  container bgp-parameters {
    description "BGP parameters";
    container common {
      description "BGP parameters common to all pseudowires";
      list rd-rt {
        if-feature evpn-bgp-params;
        key "route-distinguisher";
        leaf route-distinguisher {
          type rt-types:route-distinguisher;
          description "Route distinguisher";
        }
        uses rt-types:vpn-route-targets;
      }
    }
  }
}
```

```
        description "A list of route distinguishers and " +
                    "corresponding VPN route targets";
    }
}
leaf arp-proxy {
    type boolean;
    default false;
    description "Enable (TRUE) or disable (FALSE) ARP proxy";
}
leaf arp-suppression {
    type boolean;
    default false;
    description "Enable (TRUE) or disable (FALSE) " +
                "ARP suppression";
}
leaf nd-proxy {
    type boolean;
    default false;
    description "Enable (TRUE) or disable (FALSE) ND proxy";
}
leaf nd-suppression {
    type boolean;
    default false;
    description "Enable (TRUE) or disable (FALSE) " +
                "ND suppression";
}
leaf underlay-multicast {
    type boolean;
    default false;
    description "Enable (TRUE) or disable (FALSE) " +
                "underlay multicast";
}
leaf flood-unknown-unicast-supression {
    type boolean;
    default false;
    description "Enable (TRUE) or disable (FALSE) " +
                "flood unknown unicast suppression";
}
leaf vpws-vlan-aware {
    type boolean;
    default false;
    description "Enable (TRUE) or disable (FALSE) " +
                "VPWS VLAN aware";
}
container routes {
    config false;
    description "routes";
}
```

```
list ethernet-auto-discovery-route {
  uses route-rd-rt-grp;
  leaf ethernet-segment-identifier {
    type uint32;
    description "Ethernet segment identifier (esi)";
  }
  leaf ethernet-tag {
    type uint32;
    description "An ethernet tag (etag) indentifying a " +
      "broadcast domain";
  }
  list path {
    uses next-hop-label-grp;
    uses path-detail-grp;
    description "path";
  }
  description "ethernet-auto-discovery-route";
}
list mac-ip-advertisement-route {
  uses route-rd-rt-grp;
  leaf ethernet-segment-identifier {
    type uint32;
    description "Ethernet segment identifier (esi)";
  }
  leaf ethernet-tag {
    type uint32;
    description "An ethernet tag (etag) indentifying a " +
      "broadcast domain";
  }
  leaf mac-address {
    type yang:hex-string;
    description "Route mac address";
  }
  leaf mac-address-length {
    type uint8 {
      range "0..48";
    }
    description "mac address length";
  }
  leaf ip-prefix {
    type inet:ip-prefix;
    description "ip-prefix";
  }
  list path {
    uses next-hop-label-grp;
    uses next-hop-label2-grp;
    uses path-detail-grp;
    description "path";
  }
}
```

```
    }
    description "mac-ip-advertisement-route";
  }
  list inclusive-multicast-ethernet-tag-route {
    uses route-rd-rt-grp;
    leaf ethernet-segment-identifier {
      type uint32;
      description "Ethernet segment identifier (esi)";
    }
    leaf originator-ip-prefix {
      type inet:ip-prefix;
      description "originator-ip-prefix";
    }
    list path {
      uses next-hop-label-grp;
      uses path-detail-grp;
      description "path";
    }
    description "inclusive-multicast-ethernet-tag-route";
  }
  list ethernet-segment-route {
    uses route-rd-rt-grp;
    leaf ethernet-segment-identifier {
      type uint32;
      description "Ethernet segment identifier (esi)";
    }
    leaf originator-ip-prefix {
      type inet:ip-prefix;
      description "originator ip-prefix";
    }
    list path {
      leaf next-hop {
        type inet:ip-address;
        description "next-hop";
      }
      uses path-detail-grp;
      description "path";
    }
    description "ethernet-segment-route";
  }
  list ip-prefix-route {
    uses route-rd-rt-grp;
    leaf ethernet-segment-identifier {
      type uint32;
      description "Ethernet segment identifier (esi)";
    }
    leaf ip-prefix {
      type inet:ip-prefix;
    }
  }
}
```

```
        description "ip-prefix";
    }
    list path {
        uses next-hop-label-grp;
        uses path-detail-grp;
        description "path";
    }
    description "ip-prefix route";
}
}
container statistics {
    config false;
    description "Statistics";
    leaf tx-count {
        type uint32;
        description "transmission count";
    }
    leaf rx-count {
        type uint32;
        description "receive count";
    }
}
container detail {
    description "Detailed statistics";
    leaf broadcast-tx-count {
        type uint32;
        description "broadcast transmission count";
    }
    leaf broadcast-rx-count {
        type uint32;
        description "broadcast receive count";
    }
    leaf multicast-tx-count {
        type uint32;
        description "multicast transmission count";
    }
    leaf multicast-rx-count {
        type uint32;
        description "multicast receive count";
    }
    leaf unknown-unicast-tx-count {
        type uint32;
        description "unknown unicast transmission count";
    }
    leaf unknown-unicast-rx-count {
        type uint32;
        description "unknown-unicast receive count";
    }
}
```

```

    }
  }
}

/* augments */

augment "/pw:pseudowires/pw:pseudowire/pw:pw-type" {
  description "Augment for an L2VPN instance to add EVPN VPWS " +
    "pseudowire";
  case evpn-pw {
    container evpn-pw {
      description "EVPN pseudowire";
      leaf remote-id {
        type uint32;
        description "Remote pseudowire ID";
      }
      leaf local-id {
        type uint32;
        description "Local pseudowire ID";
      }
    }
  }
}

augment "/ni:network-instances/ni:network-instance/ni:ni-type" +
  "/l2vpn:l2vpn" {
  description "Augment for an L2VPN instance and EVPN association";
  leaf evpn-instance {
    type evpn-instance-ref;
    description "Reference to an EVPN instance";
  }
}

augment "/ni:network-instances/ni:network-instance/ni:ni-type" +
  "/l2vpn:l2vpn" {
  when "l2vpn:type = 'l2vpn:vpls-instance-type'" {
    description "Constraints only for VPLS pseudowires";
  }
  description "Augment for VPLS instance";
  container vpls-contstraints {
    must "not(boolean(/pw:pseudowires/pw:pseudowire" +
      "      [pw:name = current()/../l2vpn:endpoint" +
      "      /l2vpn:pw/l2vpn:name]" +
      "      /evpn-pw/remote-id)) and " +
      "not(boolean(/pw:pseudowires/pw:pseudowire" +
      "      [pw:name = current()/../l2vpn:endpoint" +
      "      /l2vpn:pw/l2vpn:name]" +

```

```

        "          /evpn-pw/local-id)) and " +
        "not(boolean(/pw:pseudowires/pw:pseudowire" +
        "          [pw:name = current()/../l2vpn:endpoint" +
        "          /l2vpn:primary-pw/l2vpn:name]" +
        "          /evpn-pw/remote-id)) and " +
        "not(boolean(/pw:pseudowires/pw:pseudowire" +
        "          [pw:name = current()/../l2vpn:endpoint" +
        "          /l2vpn:primary-pw/l2vpn:name]" +
        "          /evpn-pw/local-id)) and " +
        "not(boolean(/pw:pseudowires/pw:pseudowire" +
        "          [pw:name = current()/../l2vpn:endpoint" +
        "          /l2vpn:backup-pw/l2vpn:name]" +
        "          /evpn-pw/remote-id)) and " +
        "not(boolean(/pw:pseudowires/pw:pseudowire" +
        "          [pw:name = current()/../l2vpn:endpoint" +
        "          /l2vpn:backup-pw/l2vpn:name]" +
        "          /evpn-pw/local-id))" {
        description "A VPLS pseudowire must not be EVPN PW";
    }
    description "VPLS constraints";
}
}

/* Notifications */

notification evpn-state-change-notification {
    description "EVPN state change notification";
    leaf evpn-instance {
        type evpn-instance-ref;
        description "Related EVPN instance";
    }
    leaf state {
        type identityref {
            base evpn-notification-state;
        }
        description "State change notification";
    }
}
}
<CODE ENDS>

```

5. Security Considerations

The configuration, state, action and notification data defined in this document are designed to be accessed via the NETCONF protocol [RFC6241]. The lowest NETCONF layer is the secure transport layer and the mandatory-to-implement secure transport is SSH [RFC6242]. The NETCONF access control model [RFC6536] provides means to restrict

access for particular NETCONF users to a pre-configured subset of all available NETCONF protocol operations and content.

The security concerns listed above are, however, no different than faced by other routing protocols. Hence, this draft does not change any underlying security issues inherent in [I-D.ietf-netmod-routing-cfg]

6. IANA Considerations

None.

7. References

7.1. Normative Reference

- [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>>.

7.2. Informative References

- [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <<https://www.rfc-editor.org/info/rfc6241>>.
- [RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, DOI 10.17487/RFC6020, October 2010, <<https://www.rfc-editor.org/info/rfc6020>>.
- [RFC6242] Wasserman, M., "Using the NETCONF Protocol over Secure Shell (SSH)", RFC 6242, DOI 10.17487/RFC6242, June 2011, <<https://www.rfc-editor.org/info/rfc6242>>.
- [RFC6536] Bierman, A. and M. Bjorklund, "Network Configuration Protocol (NETCONF) Access Control Model", RFC 6536, DOI 10.17487/RFC6536, March 2012, <<https://www.rfc-editor.org/info/rfc6536>>.
- [RFC7209] Sajassi, A., Aggarwal, R., Uttaro, J., Bitar, N., Henderickx, W., and A. Isaac, "Requirements for Ethernet VPN (EVPN)", RFC 7209, DOI 10.17487/RFC7209, May 2014, <<https://www.rfc-editor.org/info/rfc7209>>.

- [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>>.
- [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>>.

Authors' Addresses

Patrice Brissette
Cisco Systems, Inc.
EMail: pbrisset@cisco.com

Himanshu Shah
Ciena Corporation
EMail: hshah@ciena.com

Helen Chen
Jabil
EMail: Ing-Wher_Chen@jabil.com

Iftekar Hussain
Infinera Corporation
EMail: ihussain@infinera.com

Kishore Tiruveedhula
Juniper Networks
EMail: kishoret@juniper.net

Jorge Rabadan
Nokia
EMail: jorge.rabadan@nokia.com

Ali Sajassi
Cisco Systems, Inc.
EMail: sajassi@cisco.com

Zhenbin Li

Huawei Technologies
EMail: lizhenbin@huawei.com

BESS Working Group
Internet Draft
Intended status: Standards Track
Expires: June 10, 2019

Yisong Liu
M. McBride
Huawei Technologies
December 10, 2018

Multicast DF Election for EVPN Based on bandwidth or quantity
draft-liu-bess-evpn-mcast-bw-quantity-df-election-00

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 June 10, 2019.

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.

Abstract

Ethernet Virtual Private Network (EVPN, RFC7432) is becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. When multi-homing from a CE to multiple PEs, including links in an EVPN instance on a given Ethernet Segment, in an all-active redundancy mode, [RFC7432] describes a basic mechanism to elect a Designated Forwarder (DF), and [I-D.ietf-bess-evpn-df-election-framework] improves basic DF election by a HRW algorithm. [I-D.ietf-bess-evpn-per-mcast-flow-df-election] enhances the HRW algorithm for the multicast flows to perform DF election at the granularity of (ESI, VLAN, Mcast flow). This document specifies a new algorithm, based on multicast bandwidth utilization and multicast state quantity, in order for the multicast flows to elect a DF.

Table of Contents

1. Introduction	2
1.1. Requirements Language	3
1.2. Terminology	3
2. Solution	4
2.1. DF Election Based on Bandwidth	4
2.2. DF Election Based on State Quantity	5
2.3. Inconsistent Timing between Multi-homed PEs	5
2.4. Increase or Decrease of Multi-homed PEs	6
2.4.1. Decrease of Multi-homed PEs	6
2.4.2. Increase of Multi-homed PEs	6
3. BGP Encoding	7
3.1. DF Election Extended Community	7
3.2. Multicast DF Extended Community	7
4. Security Considerations	8
5. IANA Considerations	8
6. References	8
6.1. Normative References	8
6.2. Informative References	9
7. Acknowledgments	9

1. Introduction

Ethernet Virtual Private Network (EVPN [RFC7432]) solutions are becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. When multi-homing from a CE to multiple PEs, with links in an EVPN instance on a given Ethernet Segment (ES), in an all-active redundancy mode, [RFC7432] defines the role of Designated Forwarder (DF) as the node that is responsible to forward multicast flows.

Per [RFC7432], the basic method of DF election is specified. The same ES is sorted in ascending order according to the IP address of the EVPN peer. The PE set is generated, and then the number of PEs is modulo according to the VLAN. The modulo value is equal to the position of the PE in the PE set. The election is the primary DF of the corresponding VLAN, and the other PEs are elected as standby.

[I-D.ietf-bess-evpn-df-election-framework] defines extended community attributes for DF elections, which can be extended to use different DF election algorithms and would be used for PEs in a redundancy group to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs in a redundancy group about its DF election algorithm by signaling a DF election extended community along with the ES route. The document also improves the basic DF election by a HRW algorithm.

[I-D.ietf-bess-evpn-per-mcast-flow-df-election] proposes a method for DF election by enhancing the HRW algorithm, adding the source and group address of the multicast flow as hash factors, and extending the types 4 and 5 of the extended community of the DF election for (S, G) and (*, G) types for different multicast flows. The source and group address is introduced as new elements to HRW algorithm, and the PE with the largest weight is selected as the DF of the multicast flow.

However, the relationship between the bandwidth of the multicast flows and the link capacity of different PEs, to the same CE device, is not considered in any of the current DF election algorithms. This may result in severe bandwidth utilization of different links due to different bandwidth usage of multicast flows. This document specifies a new algorithm for multicast flow DF election based on multicast bandwidth or multicast state quantity and extends the existing extended community defined in [I-D.ietf-bess-evpn-df-election-framework].

1.1. 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.

1.2. Terminology

CE: Customer Edge equipment

PE: Provider Edge device

EVPN: Ethernet Virtual Private Network

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'.

IGMP: Internet Group Management Protocol

MLD: Multicast Listener Discovery

PIM: Protocol Independent Multicast

2. Solution

In the DF election calculation, the bandwidth weight of each multi-homed link of the PE is added, and the bandwidth occupation of the multicast flows is calculated and divided into two scenarios:

- * The specific bandwidth value of the multicast flow exists, and the ratio of the current multicast flow bandwidth value to the link bandwidth weight is calculated according to the bandwidth weight of each multi-homed link, and the link with the smallest ratio is elected as the new multicast flow DF.
- * The specific bandwidth value of the multicast flow does not exist, and the ratio of the current multicast flow state quantity to the link bandwidth weight is calculated according to the bandwidth weight of each multi-homed link, and the link with the smallest ratio is elected as the new multicast flow DF.

In particular, if there are multiple PEs with the same calculated ratio, the DF is elected according to the method of maximum bandwidth weight of the link or maximum IP address of the EVPN peer.

Since [I-D.ietf-idr-link-bandwidth] defines the link bandwidth extended community, it can be reused to transfer the link bandwidth value of the local ES to other multi-homed PEs, so that each PE can calculate the bandwidth weight ratio of each link of the ES in advance.

2.1. DF Election Based on Bandwidth

Each PE obtains the link bandwidth values of the other multi-homed PEs in the same EVPN instance on a given ES according to the extended community of the Link bandwidth, and calculates the link bandwidth weight ratio, for example $W1:W2:...:Wn$ for N multi-homed PEs.

When the CE sends an IGMP or PIM join to one of the PEs, like PE1, PE1 advertises the PE2, PE3, ... and PEn by the EVPN IGMP/PIM Join Synch route defined in [I-D.ietf-bess-evpn-igmp-mld-proxy] and [I-D.skr-bess-evpn-pim-proxy]. If PE2, PE3, ... or PEn receives an IGMP or PIM join, the procedure will be the same.

Each PE calculates the ratio of the current multicast flows bandwidth to the link bandwidth weight. The one PE in PE1, PE2, ... and PEn, which has the smallest ratio, is elected as the DF of the new multicast flow. When the smallest ratios of more than one PE are the same, the PE with the maximum bandwidth weight of the link or the maximum EVPN peer IP address is elected as the DF.

2.2. DF Election Based on State Quantity

The procedure is almost the same as described in section 2.1. The only difference is that each PE calculates the ratio of the current number of multicast states instead of the bandwidth to the link bandwidth weight because of lacking specific bandwidth value of the multicast flows.

2.3. Inconsistent Timing between Multi-homed PEs

As a result of the same multicast join, only one of the multi-homed PEs can receive the multicast join message and advertise the EVPN Join Synch route (Type 7). The other PEs need to install the new multicast join state according to the received Synch route.

The inconsistent processing timing of the same multicast group joining process between PEs may cause electing different DFs. For example:

- * Multicast group G1, G2, and G3 join packets are sent from the CE to PE1, PE2 and PE3.

- * PE1 calculates the DF of G1, while PE2 calculates the DF of G2, and PE3 calculates the DF of G3, and at this moment each PE has not received the EVPN Join Synch route.

- * PE1, PE2 and PE3 select the link on the same ES to the CE using the algorithm as described in section 2.1 or 2.2, and the same DF may be elected for G1, G2, and G3.

- * After receiving the EVPN Join Synch route sent by PE2, PE1 may calculate the DF of G2 as PE3, which is inconsistent with the calculation result of PE2.

The DF calculation results of the PEs are inconsistent, which may result in multiple flows or traffic interruptions of the same

multicast flow state. Therefore, EVPN Join Synch routes need to carry elected DF information in the route advertisement as the extended community called Multicast DF Extended Community, which can make the DF information for a given multicast flow state between PEs consistent. The actual effect is that the PE that receives the multicast join packet completes the calculation of the DF election and notifies other PEs on the same ES.

2.4. Increase or Decrease of Multi-homed PEs

2.4.1. Decrease of Multi-homed PEs

When one of the multi-homed PEs on the same ES fails or is shut down for maintenance reasons, because the other PEs have received the synch routes of all the multicast flows, the multicast flows destined to the failed PE need to be in a specific order (for example, the group and source address ascending order) to reassign the DF. The DF election calculation based on the multicast flows bandwidth, or the number of multicast states, is completed by one of the specified multi-homing PEs, and the specified calculated PE can be selected according to the link bandwidth weight value or the IP address of the EVPN peer. The specified PE needs to advertise each DF election result of the multicast flow that belongs to the original faulty PE to the other multi-homed PEs that belong to the same ES by the EVPN Join Synch route carrying the Multicast DF Extended Community.

If a new multicast join is received in the above calculation process, the DF election calculation of the new multicast flow is still completed by the PE receiving the multicast join packet. Similarly, the PE needs to advertise the DF information to other multi-homed PEs belonging to the same ES by the EVPN Join Synch route carrying the Multicast DF Extended Community.

2.4.2. Increase of Multi-homed PEs

One multi-homing PE of the same ES is added, and no active adjustment can be performed. The DF of the subsequent new multicast flow is elected according to the algorithm of this document. The new multicast flow must be preferentially assigned to the new PE, and finally the multicast flows on the PEs of the same ES are approximately equalized.

If active adjustment is required, consider calculating the ratio using the algorithm as described in section 2.1 and 2.2. Each time the multicast entries in the PE, whose ratio of the existing multi-homed PE is the largest, are migrated to the new PE. The multicast entries are migrated in descending order of multicast flow bandwidth or in ascending order of the group and source address until the

ratio of the new PE is greater than the existing smallest ratio of other multi-homed PEs.

The calculation of the active adjustment is still performed by one specific PE among the multi-homed PEs. The specified calculated PE can be selected according to the link bandwidth weight value or the IP address of the EVPN peer.

After the new PE is started, in the synchronization process of all the multicast entries of other multi-homed PEs, the existing multicast join packet may be received on the new PE. To avoid having the existing multicast join appear as a new multicast join, and recalculating the DF and notifying the other PEs belonging to the same ES, it is necessary to start a timer to suppress the synchronization process from the new PE to other existing PE's. The timer range should also be configured.

3. BGP Encoding

3.1. DF Election Extended Community

[I-D.ietf-bess-evpn-df-election-framework] defines an extended community, which would be used for multi-homed PEs to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs its DF election capability by signaling a DF election extended community along with Ethernet-Segment Route (Type-4). The current document extends the existing extended community defined in [I-D.ietf-bess-evpn-df-election-framework]. This document defines a new DF type.

- o DF type (1 octet) - Encodes the DF Election algorithm values (between 0 and 255) that the advertising PE desires to use for the ES.

- * Type TBD: Based on bandwidth of multicast flow DF election(detailed in this document)

- * Type TBD+1: Based on quantity of multicast flow state DF election(detailed in this document)

3.2. Multicast DF Extended Community

This document defines a new extended community in EVPN Type 7 route to notify other multi-homed PEs the elected DF of a given multicast flow. The new extended community is called Multicast DF Extended Community and it belongs to the transitive extended community. The type is to be assigned. It is used to carry DF information of a given (S,G) or (*,G) multicast flow selection. The role of this extended community has been described in sections 2.3 and 2.4.

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=TBD																Sub-Type=TBD																Reserved																DF Length															
DF IP Address(Variable)																																																															

4. Security Considerations

TBD

For general EVPN Security Considerations, see [RFC7432].

5. IANA Considerations

TBD

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC7432] A. Sajassi, Ed., R. Aggarwal, N. Bitar, A. Isaac, J. Uttaro, J. Drake, and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017
- [I-D.ietf-bess-evpn-df-election-framework] J. Rabadan Ed., S. Mohanty, Ed., A. Sajassi, J. Drake, K. Nagaraj and S. Sathappan, " Framework for EVPN Designated Forwarder Election Extensibility ", December 2018, work-in-progress, draft-ietf-bess-evpn-df-election-framework-06.
- [I-D.ietf-bess-evpn-per-mcast-flow-df-election] Ali Sajassi, Mankamana Mishra, Samir Thoria, Jorge Rabadan and John Drake, " Per multicast flow Designated Forwarder Election for EVPN ", September 2018, work-in-progress, draft-ietf-bess-evpn-per-mcast-flow-df-election-00.

[I-D.ietf-idr-link-bandwidth] P. Mohapatra and R. Fernando, " BGP Link Bandwidth Extended Community ", March 2018, expired, draft-ietf-idr-link-bandwidth-07.

[I-D.ietf-bess-evpn-igmp-mld-proxy] Ali Sajassi, Samir Thoria, Keyur Patel, Derek Yeung, John Drake and Wen Lin, "IGMP and MLD Proxy for EVPN", June 2018, work-in-progress, draft-ietf-bess-evpn-igmp-mld-proxy-02.

[I-D.skr-bess-evpn-pim-proxy] J. Rabadan, Ed., J. Kotalwar, S. Sathappan, Z. Zhang and A. Sajassi, "PIM Proxy in EVPN Networks", October 2017, expired, draft-skr-evpn-bess-pim-proxy-01.

6.2. Informative References

TBD

7. Acknowledgments

The authors would like to thank the following for their valuable contributions of this document:

TBD

Authors' Addresses

Yisong Liu
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: liuyisong@huawei.com

Mike McBride
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95055
USA

Email: Michael.mcbride@huawei.com

INTERNET-DRAFT

N. Malhotra, Ed.
K. Patel
Arrcus

Intended Status: Proposed Standard

J. Rabadan
Nokia

Expires: Sept 12, 2019

Mar 11, 2019

LSoE-based PE-CE Control Plane for EVPN
draft-malhotra-bess-evpn-lsoe-00

Abstract

In an EVPN network, EVPN PEs 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 PEs 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 LSoE 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	10
3.3	Overlay IPv4 Prefix Encapsulation PDU	12
3.4	Overlay IPv6 Prefix Encapsulation PDU	13
4.	CE MAC/IP Learning on a PE AC	14
4.1	PE <-> CE LSoE Session Establishment	14
4.2	CE MAC/IP Learning	14
5.	PE Any-cast GW MAC/IP Learning on CE	15
6.	Remote CE MAC/IP Learning on CE	15
7.	PE <-> CE Control Plane with EVPN All-active Multi-Homing	16
7.1	All-active Multi-Homing Mode	16
7.2	Source MAC	17
7.3	CE MAC/IP Learning with EVPN All-active Multi-Homing	17
7.4	LAG Member Link Failure	18
7.4.1	Session Re-establishment	18
7.4.2	TLV Retention	18
7.4	LAG Failure	18
7.5	Example PE <-> CE Control Plane Flow with All-active	

Multi-Homing	19
8. Software Neighbor Tables	21
9. MAC/IP Learning Conflict Resolution	21
10. PE-CE Overlay Prefix Learning	22
11. Asymmetric EVPN-IRB	22
12. Centralized Gateway EVPN-IRB	22
13. Use Cases	22
13.1 Simplified EVPN Operations	22
13.1.1 EVPN All-active Multi-Homing	23
13.1.2 Convergence on CE Host Moves	24
13.1.2.1 Silent Hosts	24
13.1.2.2 Probing	25
13.1.3 ARP Gleaning Latency	26
13.2 Applicability to non-EVPN Use Cases	26
14. Summary	26
15. References	28
15.1 Normative References	28
15.2 Informative References	28
16. Acknowledgements	29
Contributors	29
Authors' Addresses	29

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 in-turn results in simplification of certain EVPN operations such as aliasing, MAC and IP syncing across multi-homing PEs, and probing on MAC/IP moves. In addition, it helps achieve a deterministic convergence behavior across various boot-up, mobility, and failure scenarios.

A PE may also use local policy configuration for learning prefixes behind a CE that does not run a dynamic routing protocol. A PE-CE control plane can provide an operationally simpler alternative to local configuration for such use cases, where CE and PE devices are not under the same configuration management entity.

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 [LSOE] 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.

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 LSoE: Link State over Ethernet Protocol defined in [LSOE]
- 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.

2. PE <-> CE Control Plane Overview

The Link State over Ethernet (LSoE) protocol is defined in [LSoE] 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 LSoE 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 LSoE protocol is proposed for CE MAC learning as an alternative to data-plane based source MAC learning.
- o PE-CE control plane based on LSoE 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 LSoE 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 LSoE 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 LSoE protocol document for the exact LSoE protocol specification.

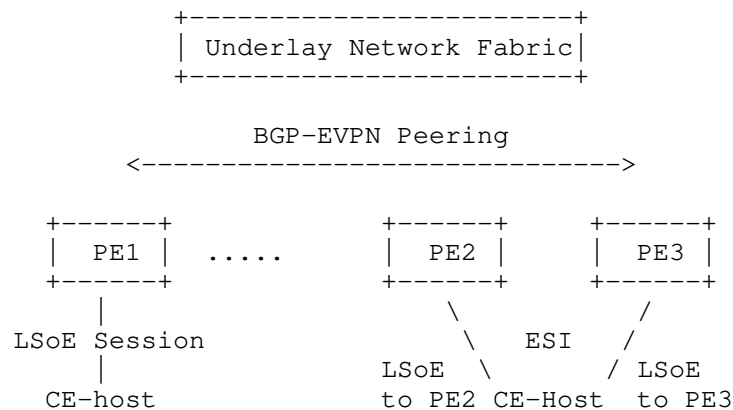


Figure 1

An LSoE session is established on layer-2 access interfaces between the EVPN PE and each connected CE host device. A session end-point is identified by a peer device MAC address on a layer-2 interface. LSoE HELLO messages are used for end-point discovery and OPEN messages are exchanged between two end-points to establish an LSoE peering. Once LSoE 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 LSoE 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.

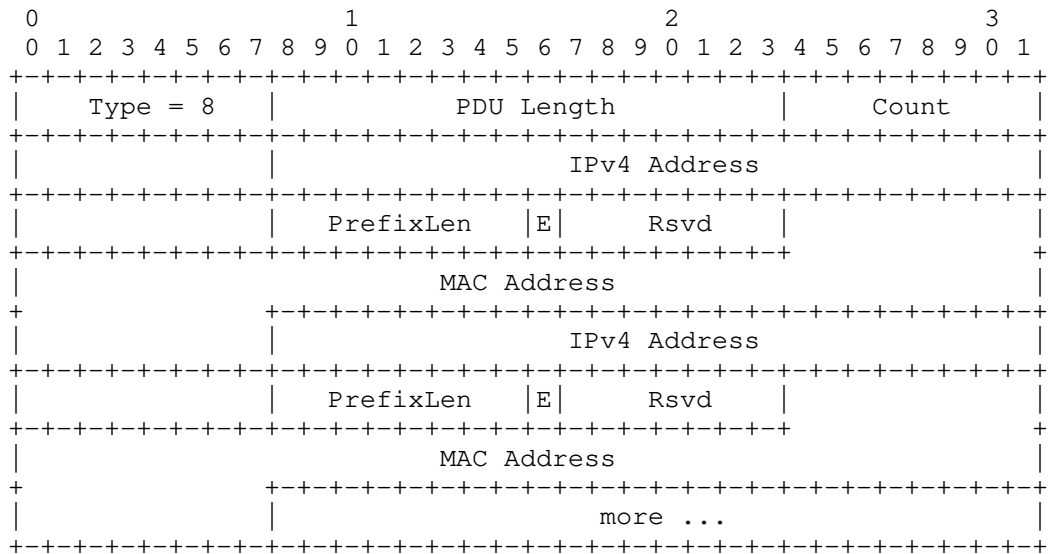


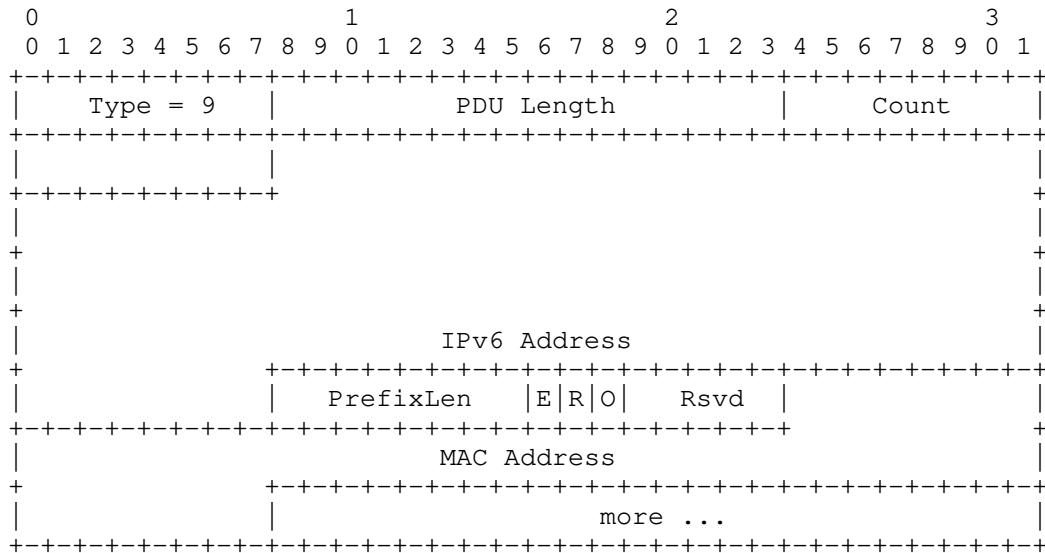
Figure 2

- o A new LSoE PDU type (8) 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.

This PDU is 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, as well as to send local CE IPv4 address and MAC binding to the PE. Procedures related to use of this PDU are

The encapsulation list in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv6 and MAC bindings:



- o A new LSoE PDU type (9) 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]

This PDU is 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 within a subnet, as well as to send local CE IPv6 address and MAC bindings to the PE. Procedures related to usage of this PDU are discussed in subsequent sections.

The encapsulation list contained in this PDU MUST follow full replace semantics as in the LSoE 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]:

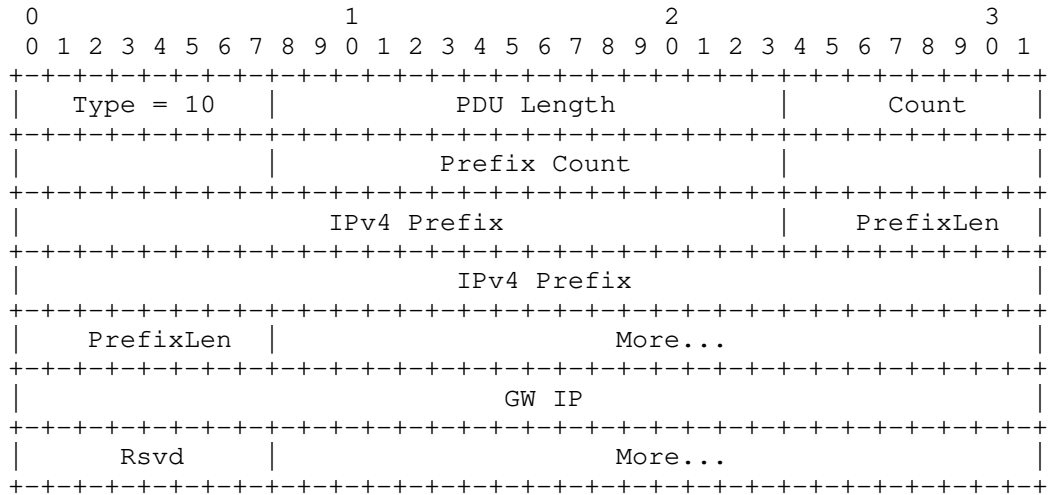


Figure 4

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 LSoE 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 LSoE 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]:

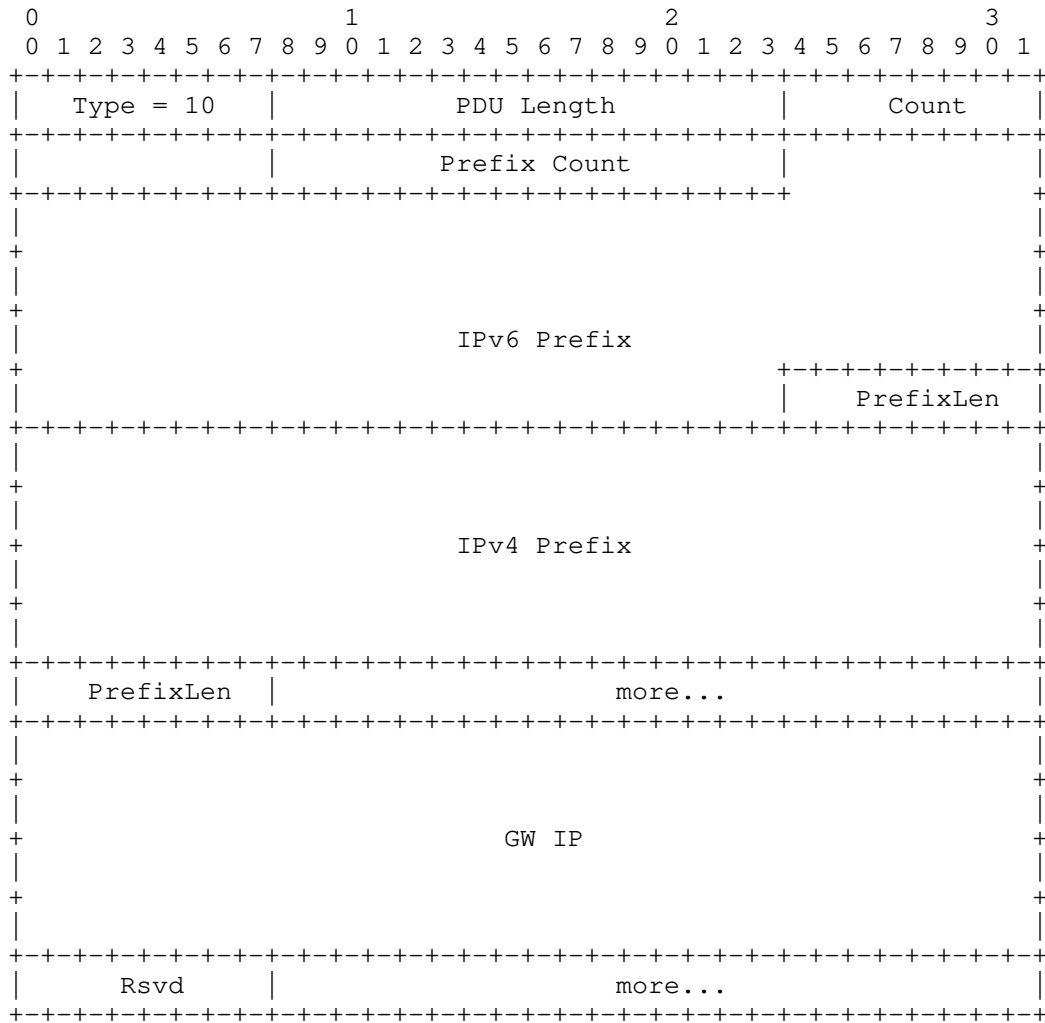


Figure 5

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 LSoE PDU type (11) 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 LSoE 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 LSoE 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 LSoE 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 LSoE 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 LSoE 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 LSoE 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 LSoE based learning is enabled in the fabric, LSoE 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 LSoE 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 <-> CE Control Plane with EVPN All-active Multi-Homing

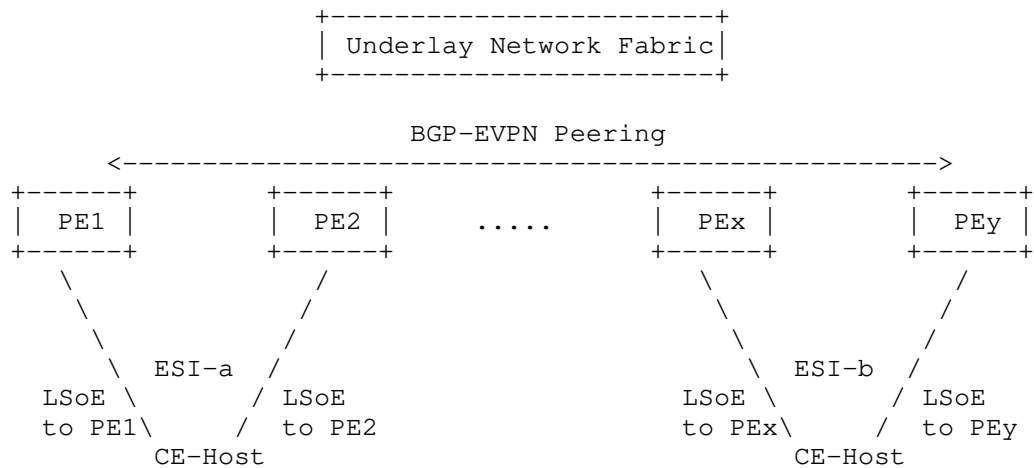


Figure 6

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 LSoE based learning to work with EVPN all-active multi-homing, a separate LSoE peering MUST be established between the CE host and each PE device. For this reason, while an EVPN PE MAY form an LSoE peering to a CE host on its local LAG AC, the CE host MUST form an LSoE 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 LSoE 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 LSoE HELLO messages MUST be replicated on ALL LAG member ports.
- o An LSoE OPEN message sent in response to a HELLO MUST be sent on the LAG member port on which the HELLO was received.
- o An LSoE session MUST be bound to the local LAG member port on

which the OPEN message was received.

- o LSoE encapsulation PDUs MUST be sent on the local LAG member port on which the session was bound.
- o LSoE 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 MAC. This however is harmless, as per the [LSOE] 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

LSoE relies on the source MAC address in the Ethernet frame to establish a peering. When running LSoE on a LAG port (in all-active multi-homing mode or regular mode), LSoE 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 LSoE on a local LAG interfaces in All-active Multi-Homing mode defined above.
- o EVPN PE MAY run LSoE 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 LSoE sessions to a CE.

With the above rules in place,

- o An LSoE session on the CE is bound to a local LAG member-port.
- o An LSoE session on the PE is bound to a local LAG AC port.
- o A single LSoE session is established at the PE to a CE on the local LAG AC.
- o 'N' LSoE 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 LSoE 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 LSoE session to a PE is bound to a LAG member interface. If the link that the LSoE session is bound to fails, LSoE 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 LSoE 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

LSoE session at the CE is torn down immediately following the session interface failure. While the LAG interface at the PE is still operationally UP, LSoE 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 LSoE 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. LSoE 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 <-> CE Control Plane Flow with All-active Multi-Homing

An example LSoE over all-active multi-homing session flow is discussed below for clarity.

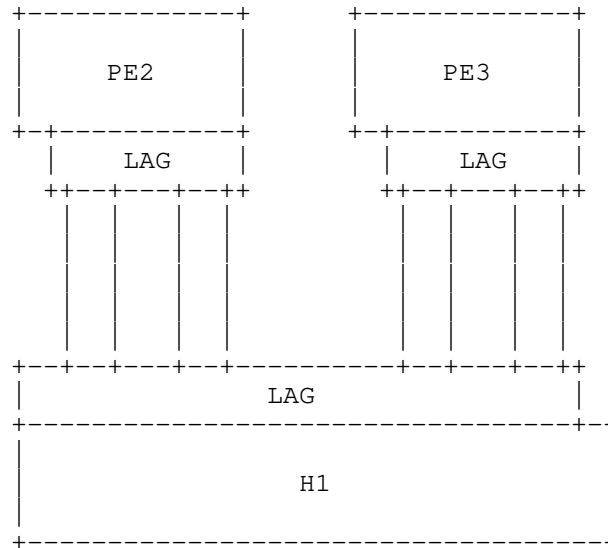


Figure 7

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 LSoE on MLAG1 in All-active Multi-Homing mode

PE2 running LSoE on LAG2 in regular mode

PE3 running LSoE on LAG3 in regular mode

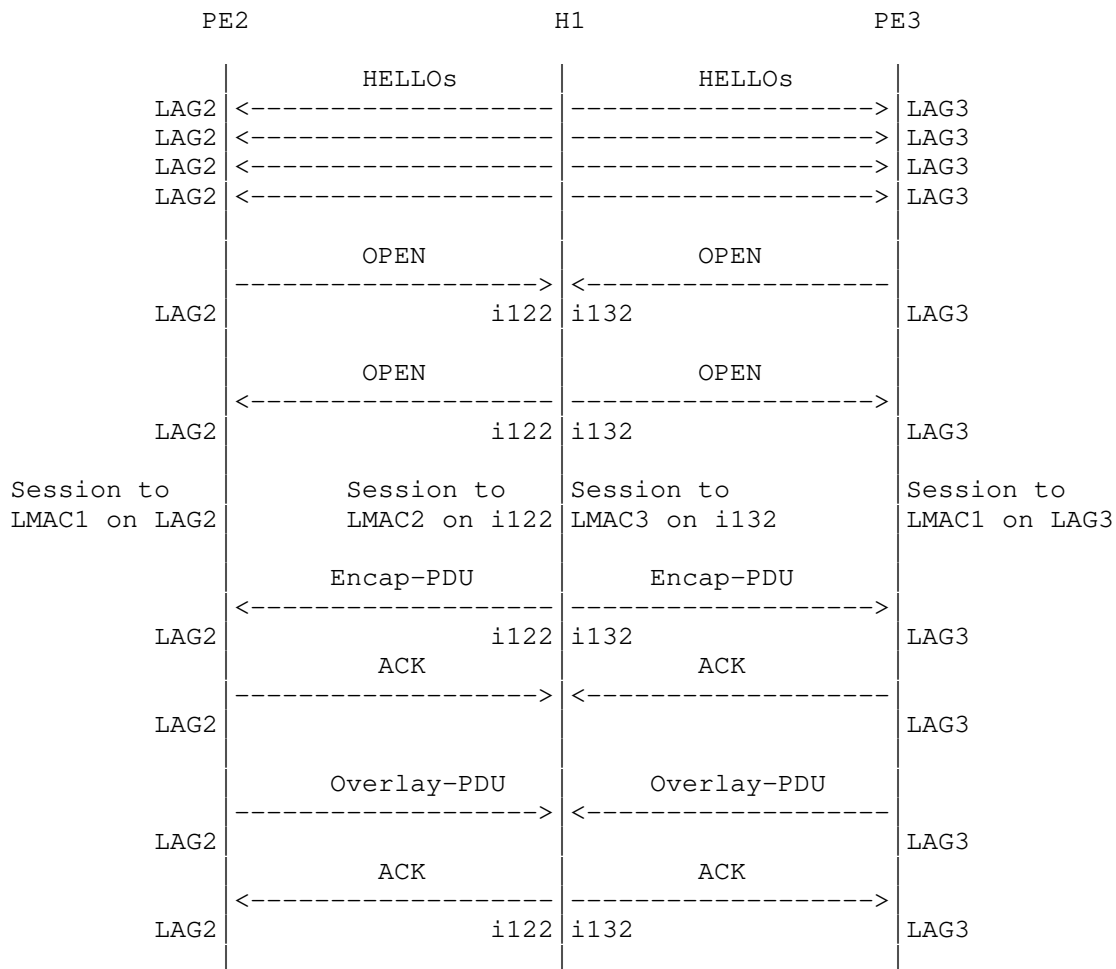


Figure 8

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 LSoE 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 LSoE learned neighbor entries may help reduce potential conflict with ARP or ND learned neighbor entries.
- o Pre-installing LSoE 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 LSoE 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 LSoE and ARP/ND. Even if LSoE learned entries were pre-installed in neighbor table, a race condition is still possible leading to a potential conflict between ARP/ND learned and LSoE learned neighbor IP adjacency. In such scenarios, LSoE 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 LSoE based learning is enabled. However, if it is not, data plane learned entries MUST be reconciled with LSoE learned entries in software and, in case of a conflict, LSoE learned entries preferred if LSoE based learning is enabled.

10. 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 LSoE 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 LSoE 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

11. 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.

12. 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.

13. Use Cases

13.1 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.

13.1.1.1 EVPN All-active Multi-Homing

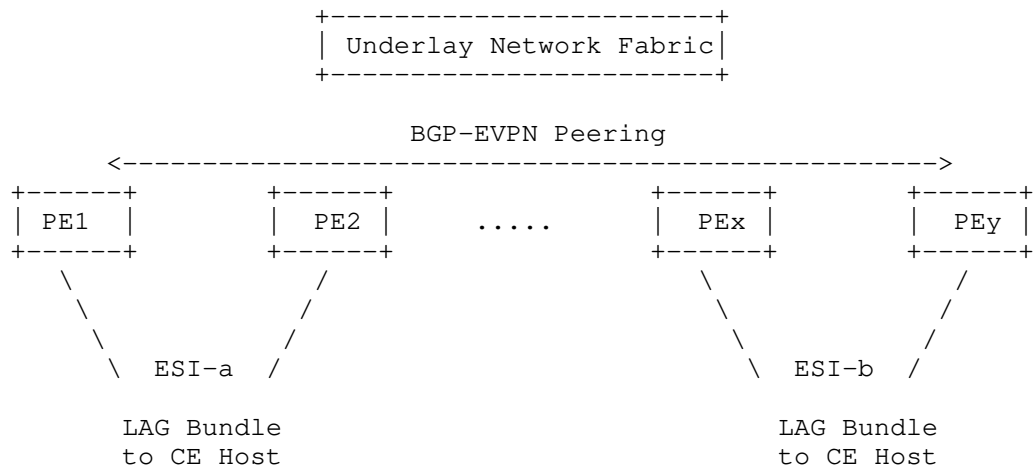


Figure 9

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.

13.1.2 Convergence on CE Host Moves

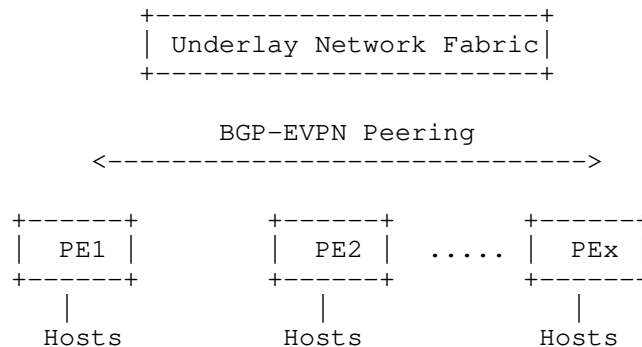


Figure 10

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.

13.1.2.1 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.

13.1.2.2 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.

13.1.3 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.

13.2 Applicability to non-EVPN Use Cases

While the LSoE 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.

14. 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.

15. References

15.1 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>>.
- [LSOE] Bush, R., Austein R., Patel, K., "Link State Over Ethernet", Feb 2019, <<https://tools.ietf.org/html/draft-ietf-lsvr-lsoe-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", Jan 2019, <<https://tools.ietf.org/html/draft-malhotra-bess-evpn-irb-extended-mobility-04>>.
- [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

16. Acknowledgements

Authors would like to thank Randy Bush and Rob Austein for detailed review and feedback to ensure consistency with base LSOE protocol specification, as well as for helping build detailed LSOE 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)
Arrcus
2077 Gateway Place, Suite #400
San Jose, CA 95119, USA

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 Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2019

S. Mohanty
M. Misra
A. Lindem
A. Sajassi
Cisco Systems, Inc.
March 11, 2019

Weighted HRW and its applications
draft-mohanty-bess-weighted-hrw-00

Abstract

Rendezvous Hashing also known as Highest Random Weight (HRW) has been used in many load balancing applications where the central problem is how to map an object to a server such that the mapping is uniform and also minimally affected by the change in the server set. Recently, it has found use in DF election algorithms in the EVPN context and load balancing using DMZ. This draft deals with the problem of achieving load balancing with minimal disruption when the servers have different weights. It provides an algorithm to do so and also describes a few use-case scenarios where this algorithmic technique can apply.

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 September 12, 2019.

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	2
2. Introduction	2
3. HRW Introduction	3
4. HRW with weights	4
5. HRW and Consistent Hashing	5
6. Weighted HRW and its application to the EVPN DF Election . .	5
7. Weighted HRW and its application to Resilient Hashing	7
8. Weighted HRW and its application to Multicast DR Election . .	7
9. Protocol Considerations	8
10. Operational Considerations	8
11. Security Considerations	8
12. Acknowledgements	8
13. References	8
13.1. Normative References	8
13.2. Informative References	9
Authors' Addresses	10

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. Introduction

Given an object *O*, a set of servers and a set of clients, a fundamental problem is how do the set of clients, independently and unanimously agree in a distributed framework, which server to assign *O*? This is the distributed hash table problem. The assignment should be "minimally disruptive" which means that there should be a minimal remapping of objects whenever a server is down or a new server comes up or the object set changes. This is a very common problem in practice in the Internet load balancing and web caching as described in the 'Akamai' paper [CHASH], database [DYNAMODB] and networking context.

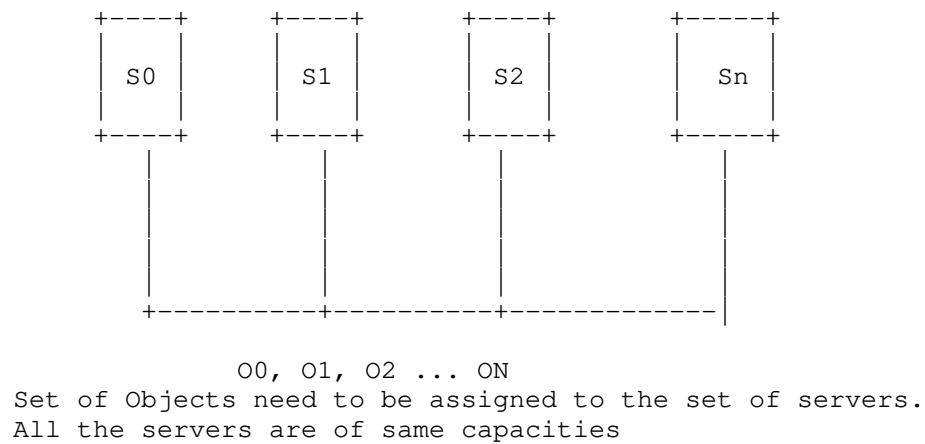


Figure 1 The object to server assignment problem

Figure 1

In the Fig 1, we show a set of servers, S_0, \dots, S_n and object pool O_0, \dots, O_n and the requirement is to assign O_i to S_j such that the servers are uniformly loaded. In addition, when any server goes down or a new one is introduced, there should be minimal reassignments.

There are two standard techniques to address this problem.

1. Consistent Hashing
2. Rendezvous Hashing
3. HRW Introduction

Highest Random Weight (HRW) as defined in [HRW1999] is originally proposed in the context of Internet Caching and proxy Server load balancing. Given an object name and a set of servers, HRW maps a request to a server using the object-id (O_i) and server-id (S_j) rather than the state of the server states. HRW computes a hash, $\text{Hash}(O_i, S_j)$ from the server-id and the object-id; this hash value can be considered as a score, and forms an ordered list of the servers based on the hash value (i.e. score) in decreasing order. The server for which the score is the highest, serves as the primary responsible for that particular object, and the server with the next highest score serves as the backup server. HRW always maps a given object name to the same server within a given cluster; consequently it can

be used at client sites to achieve global consensus on object-server mappings. When that server goes down, the backup server becomes the responsible designate.

Choosing an appropriate hash function that is statistically oblivious to the key distribution and imparts a good uniform distribution of the hash output is an important aspect of the algorithm. The original HRW [HRW1999] provides pseudorandom functions based on Unix utilities `rand` and `srand` and easily constructed XOR functions that perform considerably well. Any good uniform hash function like the Jenkins hash for instance will also work. HRW already finds use in multicast and ECMP [RFC2991], [RFC2992].

4. HRW with weights

The issue when the servers are not of the same capacity is also quite a common problem. However this problem has not gained as much attention as it should. In such a case, an obvious approach is to take the normalized weight factor into account, $f_i = w_i / \sum(w_i)$ and multiply the `Hash(Oi, Sj)` with that value i.e. the value $f_i * \text{Hash}(O_i, S_j)$. The Cache Array Routing Protocol [CARP] used this method. However there is a problem with this approach, since any change in weight of any of the servers, will result in a change in the normalized weights for everyone. This will necessitate re-computing all the weighted hash values all over again. Therefore this approach does not have the minimal disruption property of the HRW. We address this issue of the weighted HRW with minimal disruption in this draft.

Instead of re-normalizing the weights, or, in other words relatively scaling them, the approach taken by [WHRW] is to adjust the score before weighing them. When a server is added, removed or modified (its weight changes), only the score for that server changes. That server may win or lose some objects. Other servers remain affected. There is no needless transfer of objects between servers whose weight did not change. [WHRW] uses a clever way to accomplish this by defining the score function as:

1. $\text{Score}(O_i, S_j) = -w_i / \log(\text{Hash}(O_i, S_j) / H_{\max})$; where H_{\max} is the maximum hash value.

The author provides a mathematical proof as to why this choice of the Score function works with very mild assumptions on the probability distribution of the hash function.

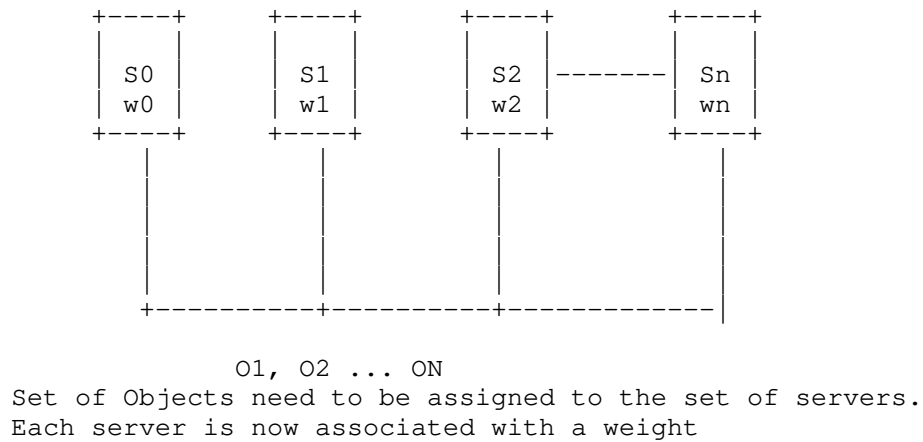


Figure 1 The object to server assignment problem

Figure 2

5. HRW and Consistent Hashing

HRW is not the only algorithm that addresses the object to server mapping problem with goals of fair load distribution, redundancy and fast access. There is another family of algorithms that also addresses this problem; these fall under the umbrella of the Consistent Hashing Algorithms [CHASH]. These will not be considered here.

6. Weighted HRW and its application to the EVPN DF Election

The notion and need for the Designated Forwarder is described in [RFC7432]. Consider a CE that is a host or a router that is multi-homed directly to more than one PE in an EVPN instance on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

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

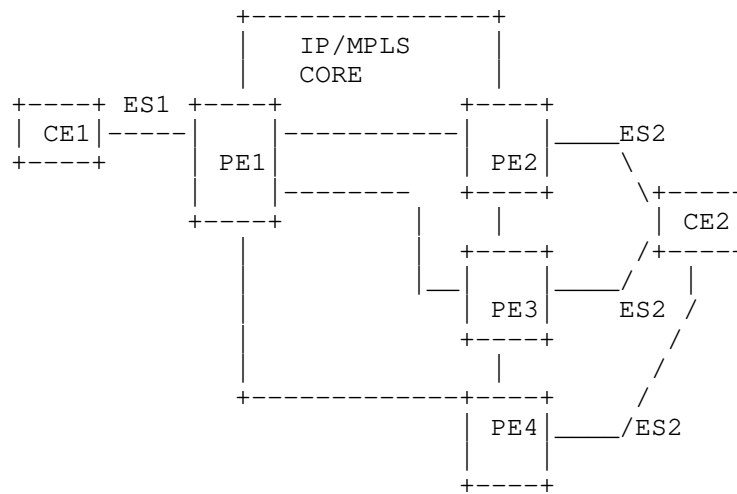


Figure 3 Multi-homing Network of EVPN

Figure 3

Figure 3 illustrates a case where there are two Ethernet Segments, ES1 and ES2. PE1 is attached to CE1 via Ethernet Segment ES1 whereas PE2, PE3 and PE4 are attached to CE2 via ES2 i.e. PE2, PE3 and PE4 form a redundancy group. Since CE2 is multi-homed to different PEs on the same Ethernet Segment, it is necessary for PE2, PE3 and PE4 to agree on a DF to satisfy the above mentioned requirements.

The use of HRW in the EVPN DF Election is described in [I-D.ietf-bess-evpn-df-election-framework]. In that draft it is explained how the HRW DF Election performs better than the modulo DF Election algorithm in [RFC7432]. However, it is implicitly assumed there that all the PEs are of the same capacity (weights equal).

DMZ link bandwidth for load balancing flows across multiple EBGP egress points is described in [I-D.ietf-idr-link-bandwidth]. It has been extended to the case of cumulative DMZ load balancing [I-D.mohanty-bess-ebgp-dmz] in the case of an all EBGP network in the data center. [I-D.ietf-bess-evpn-unequal-lb] describes the use of the DMZ in the EVPN DF Election. The argument is made that ideally one should be able to change the link bandwidth in one or more of the multi-homed PEs rather than have to change in all of the multi-homed PEs simultaneously. The draft describes the bandwidth increments to be taken into consideration and proposes an iterative way to assign

the score function. The description in Section 4.3.2 of [I-D.ietf-bess-evpn-unequal-lb] is a non-optimal solution and somewhat empirical. It does not obey the minimal disruption property of the HRW.

In contrast to the procedures for weighted HRW in 4.3.2 of [I-D.ietf-bess-evpn-unequal-lb], we can achieve an optimal solution for weighted HRW in [I-D.ietf-bess-evpn-unequal-lb] using the score function as described in Section 4 above and obviating the need to take bandwidth increments. It is an order of magnitude faster and efficient and minimally disruptive.

7. Weighted HRW and its application to Resilient Hashing

With the exponential increase in the number of physical links used in data centers, there is also the potential for an increase in the number of failed physical links. In systems that employ static hashing for load balancing flows across members of port channels or Equal Cost Multipath (ECMP) groups, each flow is hashed to a link. When a link fails, all flows including those that were previously mapped to the non-failed links are rehashed across the remaining working links. This causes packet reordering of flows that were in fact not mapped to the link that failed. A similar rehashing with packet re-ordering also happens when a link is added to the port channel or Equal Cost Multipath (ECMP) group. With the ever increasing number of physical links used in the data centers there the possibility for increasing number of failed links only increases. Hence the resilient hashing is very important.

However when the links are not of the same speed, Resilient hashing for ECMP does not apply per-se. However, one can use the method explained in Section 4 to achieve resilient hashing even in the Unequal Cost Multipath (UCMP) case or when member links are of different bandwidths.

8. Weighted HRW and its application to Multicast DR Election

[I-D.mankamana-pim-bdr] propose a mechanism to elect backup DR on a shared LAN. A backup DR on LAN would be useful for faster convergence. When the access bandwidth is different for the PIM routers and we want to do a load balancing among the PIM routers for DR/backup DR functionality with regards to the various (S,G) flow, technique similar to Section 4 can be applied. The details of the problem is out of the scope of the current draft and is being worked on separately at this time.

9. Protocol Considerations

A request needs to be registered with IANA registry for the weighted HRW EVPN DF Election Algorithm in the DF Alg field in the DF Election Extended Community in draft [I-D.ietf-bess-evpn-df-election-framework].

10. Operational Considerations

TBD.

11. Security Considerations

This document raises no new security issues for EVPN.

12. Acknowledgements

The authors would like to thank Shyam Sethuram and Peter Psenak for useful discussions related to this draft.

13. References

13.1. Normative References

[HRW1999] Thaler, D. and C. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998.

[I-D.ietf-bess-evpn-df-election-framework]
Rabadan, J., satyamoh@cisco.com, s., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for EVPN Designated Forwarder Election Extensibility", draft-ietf-bess-evpn-df-election-framework-09 (work in progress), January 2019.

[I-D.ietf-bess-evpn-unequal-lb]
Malhotra, N., Sajassi, A., Rabadan, J., Drake, J., Lingala, A., and S. Thoria, "Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing", draft-ietf-bess-evpn-unequal-lb-00 (work in progress), September 2018.

[I-D.ietf-idr-extcomm-iana]
Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", draft-ietf-idr-extcomm-iana-02 (work in progress), December 2013.

- [I-D.ietf-idr-link-bandwidth]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-07 (work in progress), March 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>>.
- [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>>.
- [WHRW] Resch, J., "New hashing Algorithms for Data Storage", Storage Developer Conference 18, November 2015.

13.2. Informative References

- [CARP] Valloppillil, V. and K. Ross, "Cache Array Routing Protocol v1.1", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998.
- [CHASH] Karger, D., Lehman, E., Leighton, T., Panigrahy, R., Levine, M., and D. Lewin, "Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web", ACM Symposium on Theory of Computing ACM Press New York, May 1997.
- [CLRS2009] Cormen, T., Leiserson, C., Rivest, R., and C. Stein, "Introduction to Algorithms (3rd ed.)", MIT Press and McGraw-Hill ISBN 0-262-03384-4., February 2009.
- [DYNAMODB] Decennia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Voshall, P., and W. Vogels, "Dynamo: Amazon's Highly Available Key-value Store", SOSR 07, October 2007.

- [I-D.mankamana-pim-bdr]
mishra, m., "PIM Backup Designated Router Procedure",
draft-mankamana-pim-bdr-00 (work in progress), June 2018.
- [I-D.mohanty-bess-ebgp-dmz]
satyamoh@cisco.com, s., Millisor, A., and A. Vayner,
"Cumulative DMZ Link Bandwidth and load-balancing", draft-
mohanty-bess-ebgp-dmz-00 (work in progress), March 2018.
- [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and
Multicast Next-Hop Selection", RFC 2991,
DOI 10.17487/RFC2991, November 2000,
<<https://www.rfc-editor.org/info/rfc2991>>.
- [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path
Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000,
<<https://www.rfc-editor.org/info/rfc2992>>.

Authors' Addresses

Satya Ranjan Mohanty
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA

Email: satyamoh@cisco.com

Mankamana Misra
Cisco Systems, Inc.
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: mankamis@cisco.com

Acee Lindem
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: acee@cisco.com

Ali Sajassi
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: sajassi@cisco.com

BESS Workgroup
Internet Draft
Intended status: Standards Track

J. Rabadan, Ed.
K. Nagaraj
Nokia

W. Lin
Juniper

A. Sajassi
Cisco

Expires: September 9, 2019

March 8, 2019

EVPN Multi-Homing Extensions for Split Horizon Filtering
draft-nr-bess-evpn-mh-split-horizon-00

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 or GENEVE. 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 September 12, 2019.

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)	7
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	10
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. An example is GENEVE, where the tunnel source IP address identifies the ingress NVE, and [EVPN-GENEVE] defines an Ethernet option TLV (Type Length Value) to encode an ESI label value. 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.

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

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

[DF] Rabadan, J., Mohanty, S., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for EVPN Designated Forwarder Election Extensibility", internet-draft draft-ietf-bess-evpn-df-election-framework-09.txt, January 2019.

[EVPN-GENEVE] Boutros, S., Sajassi, A., Drake, J., and J. Rabadan, "EVPN control plane for Geneve", Work in Progress, draft-boutros-bess-evpn-geneve-02, March 2018.

[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>>.

[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-08, October 2018.

[TUNNEL-ENCAP] Rosen, E., Ed., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", Work in Progress draft-ietf-idr-tunnel-encaps-09, February 2018.

[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>>.

9. Acknowledgments

10. Contributors

Authors' Addresses

Jorge Rabadan (Editor)
Nokia
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@nokia.com

Kiran Nagaraj
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA
Email: 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

BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: September 6, 2019

R. Bickhart
W. Lin
J. Drake
Juniper Networks
J. Rabadan
Nokia
March 5, 2019

Proxy IP->MAC Advertisement in EVPNs
draft-rbickhart-evpn-ip-mac-proxy-adv-00

Abstract

This document specifies procedures for EVPN PEs connected to a common multihomed site to generate proxy EVPN type 2 IP->MAC advertisements on behalf of other PEs to facilitate preservation of ARP/ND state across link or node failures.

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 September 6, 2019.

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. Conventions and Terminology	2
2. Introduction	2
3. IP->MAC Proxy Advertisements	3
3.1. Interoperation with Legacy PEs	5
3.2. Single-Active Multihoming Considerations	5
3.3. MAC Route Attribute Considerations	5
4. EVPN ARP/ND Extended Community	5
5. IANA Considerations	6
6. Operational Considerations	6
7. Security Considerations	7
8. Normative References	7
Authors' Addresses	7

1. Conventions and Terminology

This document assumes familiarity with the terminology used in EVPN [RFC7432]. A few key terms used in this document are defined below:

CE: Customer edge device.

PE: Provider edge device.

IP->MAC: An IP address associated with a MAC address.

ARP: Address Resolution Protocol.

ND: Neighbor Discovery Protocol.

DF: Designated Forwarder.

R-bit: Router Flag in NA messages, as per ND for IPv6 [RFC4861].

O-bit: Override Flag in NA messages, as per ND for IPv6 [RFC4861].

2. Introduction

EVPN [RFC7432] allows for the distribution of IP->MAC bindings of connected hosts learned through IPv4 ARP and IPv6 neighbor discovery (ND) using type 2 MAC/IP advertisement routes. When end hosts are connected to a multihomed site of an EVPN running in all-active mode, depending on the learning mechanisms of the multihoming PEs and the load balancing mechanisms implemented by the CE devices multihomed to the EVPN PEs, local learning of an IP->MAC may be limited to a subset

of the total number of multihoming PE's, possibly only a single PE device. In the steady state, the IP->MAC originally learned locally on one or more of the set of multihoming PE's is synchronized to all remaining PE's attached to the same multihoming site via the EVPN control plane. Once synchronized, the IP->MAC is available on each multihoming PE as well as other remote PE's not connected to the multihomed site for use in forwarding traffic or suppressing ARP or ND messaging in the EVPN.

In the event of the complete failure of a multihoming PE or the failure of a multihoming PE's link to the multihomed site, any IP->MAC locally learned on that PE or locally attached link will be invalidated and will be withdrawn from the EVPN control plane. If the source of the failure was the only origin of any particular IP->MAC, that IP->MAC will completely disappear from the EVPN until such time that one of the remaining multihoming PE's is able to relearn the IP->MAC that was lost. Traffic forwarding or other EVPN features like ARP/ND suppression may fail during the intermediate period between the loss of the IP->MAC from the original local learning PE and later learning and distribution of the IP->MAC from a new local learning PE.

This document specifies procedures to preserve IP->MAC state across the remaining multihoming PE's in the EVPN to bridge the gap between the initial failure and the subsequent relearning of the IP->MAC on one of the remaining multihoming PE's.

3. IP->MAC Proxy Advertisements

Preserving IP->MAC state in the EVPN until relearning and distribution of the new IP->MAC state to all PE's is completed can be accomplished by using IP->MAC proxy advertisements. When an IP->MAC for a host connected to a multihomed site is locally learned by a PE, the PE will advertise the IP->MAC via an EVPN MAC/IP route as usual. When other PE's learn that IP->MAC from the control plane upon reception of the MAC/IP route, they will install the ARP/ND state derived from the received IP->MAC for local use as usual. Additionally, if the receiving PE is locally connected to the same multihomed ethernet segment where the received IP->MAC originated and the IP->MAC was not previously locally learned and advertised, the receiving PE will inject its own EVPN MAC/IP route carrying the same IP->MAC (and with the same ESI) into the control plane and mark that injected route with a special proxy IP->MAC indication. Assuming that all PE's attached to the multihomed site support this proxy advertisement functionality, the result is that each PE attached to the site will originate the given IP->MAC using an EVPN MAC/IP route, some of the route advertisements possibly carrying the proxy

indication and at least one route advertisement not marked with the proxy indication.

A subsequent PE failure, link failure, or other event triggering the loss of all non-proxy IP->MAC state on a multihoming PE will cause that PE to start an aging timer for the proxy IP->MAC the PE had previously advertised. The aging timer should be initialized to the same age-time as the system default for ARP/ND aging, but an implementation may allow the initial age-time used for proxy a IP->MAC to be set administratively. While the aging timer is running, the multihoming PE will take no other actions and will continue using the proxy IP->MAC state for local forwarding and ARP/ND purposes and will continue to advertise the IP->MAC in the control plane with the proxy indication set. Remote PEs not connected to the multihomed site will ignore the proxy indication completely, and will be unaware of the difference between proxy and non-proxy IP->MAC advertisements. In this state, the EVPN will continue working as before the failure, with the exception of the failed link or PE being removed from the forwarding path.

In the event that one of the remaining multihoming PEs now learns the IP->MAC locally, it will restart the aging timer for the IP->MAC with the default ARP/ND age-time and remove the proxy indication from the EVPN MAC/IP route for the IP->MAC previously advertised in the control plane. When any other multihoming PE observes the removal of the proxy indication from at least one of the sources advertising the IP->MAC, that PE will stop the aging timer for the locally advertised proxy IP->MAC and continue advertising the IP->MAC with the proxy indication set as before.

In the event that a multihoming PE fails to learn the IP->MAC locally before the aging timer for the proxy IP->MAC expires, that PE will withdraw the EVPN MAC/IP route for proxy IP->MAC that it had advertised previously. In this way, if all multihoming PEs fail to learn the IP->MAC locally within the age-time, the proxy IP->MAC advertisements will expire on every PE and will be withdrawn, completely removing the IP->MAC from the EVPN.

In the case that a non-proxy IP->MAC is withdrawn from the EVPN because the original dynamically learned ARP/ND entry ages out due to end host inactivity or shutdown rather than a PE node or link failure, PEs which advertised a proxy IP->MAC will still follow the same procedures as above and retain their proxy IP->MAC advertisements until the age-time has passed. Implementations may allow the proxy IP->MAC age-time to be administratively specified separately from the regular system ARP/ND age-time to tune how fast stale proxy IP->MAC advertisements are cleared from the EVPN. Additionally, a PE may optionally use a mechanism like send-refresh

[I-D.ietf-bess-evpn-proxy-arp-nd] to probe the liveness of the IP->MAC and withdraw the proxy IP->MAC from the control plane before the age-time if the PE determines that the IP->MAC is no longer active.

3.1. Interoperation with Legacy PEs

A heterogeneous mix of new PEs supporting proxy IP->MAC advertisement and legacy PEs not supporting proxy IP->MAC advertisement is supported in the event of incremental configuration of the feature or incremental upgrades of PEs attached to the same ethernet segment. Although legacy PE devices will continue to operate with the traditional mechanisms and advertise only locally learned IP->MAC entries, they can make use of any remotely learned proxy IP->MAC advertised by other PEs supporting proxy advertisement.

3.2. Single-Active Multihoming Considerations

Proxy IP->MAC advertisement is not applicable to ethernet segments configured for single-active multihoming because MAC advertisements are the indication of which multihoming PE is the DF for remote PEs not directly connected ethernet segment. Advertisement of a proxy IP->MAC by a non-DF multihoming PE will prevent remote PEs not directly attached to the ethernet segment from determining the correct DF.

3.3. MAC Route Attribute Considerations

When a PE advertises a proxy IP->MAC that was originally learned from the control plane with a MAC mobility extended community attached with a nonzero sequence number, the PE should advertise the new proxy IP->MAC with the same sequence number as originally received. The presence or lack of a proxy indication for a received IP->MAC advertisement should not be used in active MAC determination for MAC mobility purposes.

When a PE advertises a proxy IP->MAC for an IPv6 address learned from the control plane that has the 'R' or 'O' bits set in the EVPN ND extended community, the new proxy IP->MAC should carry an EVPN ND extended community with the same 'R' and 'O' bits as originally received.

4. EVPN ARP/ND Extended Community

EVPN already provides an extended community to signal additional state relevant to IPv6 ND (the override and router bits). Because the proxy indication described in this document is equally applicable to both ARP and ND, we propose renaming the EVPN Neighbor Discovery

(ND) Extended Community (type 0x08) to EVPN ARP/ND Extended Community and allocating an additional bit from the flags field of the community to signal the proxy advertisement state.

```

      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=0x08 | Reserved | P | O | R | Reserved=0  |
      +-----+-----+-----+-----+-----+-----+-----+-----+
      |                                     Reserved=0                                     |
      +-----+-----+-----+-----+-----+-----+-----+-----+

```

The following bits in the flags field in the third octet of the extended community are defined. The remaining bits must be set to zero when sending and must be ignored when receiving this community.

Bit Name	Meaning
O, R	Defined in IPv6 Neighbor Advertisement Flags in EVPN [I-D.ietf-bess-evpn-na-flags]
P	Proxy IP->MAC advertisement defined in this draft

EVPN ARP/ND Extended Community Flags

5. IANA Considerations

This document requests the value of 0x04 of the EVPN ARP/ND extended community flags field to be assigned to the proxy IP->MAC advertisement (P) bit.

6. Operational Considerations

Depending on the number of multihoming PEs and MAC/IP scaling of an EVPN, proxy advertisement of IP->MAC entries by other PEs in addition to the devices initially learning IP->MAC entries locally in the data plane could cause scalability concerns for operators. Proxy advertisements would increase the total number of EVPN routes maintained in the route tables of PEs, as well as increase the time required for PEs to download all remotely learned EVPN routes. Protocol implementations should provide administrative controls for operators to limit proxy advertisement functionality to situations where the benefits are required and the scale overhead is acceptable.

7. Security Considerations

This draft does not introduce any new security considerations to EVPN.

8. Normative References

- [I-D.ietf-bess-evpn-na-flags]
Rabadan, J., Sathappan, S., and K. Nagaraj, "Propagation of IPv6 Neighbor Advertisement Flags in EVPN", draft-ietf-bess-evpn-na-flags-02 (work in progress), October 2018.
- [I-D.ietf-bess-evpn-proxy-arp-nd]
Rabadan, J., Sathappan, S., Nagaraj, K., Henderickx, W., Hankins, G., King, T., Melzer, D., and E. Nordmark, "Operational Aspects of Proxy-ARP/ND in EVPN Networks", draft-ietf-bess-evpn-proxy-arp-nd-05 (work in progress), November 2018.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <<https://www.rfc-editor.org/info/rfc4861>>.
- [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>>.

Authors' Addresses

Ryan Bickhart
Juniper Networks

Email: rbickhart@juniper.net

Wen Lin
Juniper Networks

Email: wlin@juniper.net

John Drake
Juniper Networks

Email: jdrake@juniper.net

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com

BESS Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

A. Sajassi, Ed.
A. Banerjee
S. Thoria
D. Carrel
B. Weis
Cisco

Expires: September 11, 2019

March 11, 2019

Secure EVPN
draft-sajassi-bess-secure-evpn-01

Abstract

The applications of EVPN-based solutions ([RFC7432] and [RFC8365]) have become pervasive in Data Center, Service Provider, and Enterprise segments. It is being used for fabric overlays and inter-site connectivity in the Data Center market segment, for Layer-2, Layer-3, and IRB VPN services in the Service Provider market segment, and for fabric overlay and WAN connectivity in Enterprise networks. For Data Center and Enterprise applications, there is a need to provide inter-site and WAN connectivity over public Internet in a secured manner with same level of privacy, integrity, and authentication for tenant's traffic as IPsec tunneling using IKEv2. This document presents a solution where BGP point-to-multipoint signaling is leveraged for 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.

Status of this Memo

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

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

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

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

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

Copyright and License Notice

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

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

Table of Contents

1	Introduction	6
2	Requirements	7
2.1	Tenant's Layer-2 and Layer-3 data & control traffic	7
2.2	Tenant's Unicast & Multicast Data Protection	7
2.3	P2MP Signaling for SA setup and Maintenance	7
2.4	Granularity of Security Association Tunnels	7
2.5	Support for Policy and DH-Group List	8
3	Solution Description	8
3.1	Inheritance of Security Policies	9
3.2	Distribution of Public Keys and Policies	10
3.2.1	Minimal DIM	10
3.2.2	Multiple Policies	10
3.2.2.1	Multiple DH-groups	11
3.2.2.2	Multiple or Single ESP SA policies	11
3.3	Initial IPsec SAs Generation	11
3.4	Re-Keying	12
3.5	IPsec Databases	12
4	Encapsulation	12
4.1	Standard ESP Encapsulation	13
4.2	ESP Encapsulation within UDP packet	13
5	BGP Encoding	15
5.1	The Base (Minimal Set) DIM Sub-TLV	15

5.2 Key Exchange Sub-TLV	16
5.3 ESP SA Proposals Sub-TLV	17
5.3.1 Transform Substructure	17
6 Applicability to other VPN types	18
7 Acknowledgements	19
8 Security Considerations	19
9 IANA Considerations	19
10 References	19
10.1 Normative References	19
10.2 Informative References	20
Authors' Addresses	21

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.

AC: Attachment Circuit.

ARP: Address Resolution Protocol.

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

BD Route Target: refers to the Broadcast Domain assigned Route Target [RFC4364]. In case of VLAN-aware bundle service model, all the BD instances in the MAC-VRF share the same Route Target.

BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per [RFC7432].

DGW: Data Center Gateway.

Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].

Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload. Examples of this type of tunnels are VXLAN or GENEVE.

EVI: EVPN Instance spanning the NVE/PE devices that are participating

on that EVPN, as per [RFC7432].

EVPN: Ethernet Virtual Private Networks, as per [RFC7432].

GRE: Generic Routing Encapsulation.

GW IP: Gateway IP Address.

IPL: IP Prefix Length.

IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload).

IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by EVPN and IP-VPN address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.

IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on an NVE/PE, as per [RFC7432]. A MAC-VRF is also an instantiation of an EVI in an NVE/PE.

ML: MAC address length.

ND: Neighbor Discovery Protocol.

NVE: Network Virtualization Edge.

GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].

NVO: Network Virtualization Overlays.

RT-2: EVPN route type 2, i.e., MAC/IP advertisement route, as defined in [RFC7432].

RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [EVPN-PREFIX].

SBD: Supplementary Broadcast Domain. A BD that does not have any ACs, only IRB interfaces, and it is used to provide connectivity among all the IP-VRFs of the tenant. The SBD is only required in IP-VRF- to-IP-VRF use-cases (see Section 4.4.).

SN: Subnet.

TS: Tenant System.

VA: Virtual Appliance.

VNI: Virtual Network Identifier. As in [RFC8365], the term is used as a representation of a 24-bit NVO instance identifier, with the understanding that VNI will refer to a VXLAN Network Identifier in VXLAN, or Virtual Network Identifier in GENEVE, etc. unless it is stated otherwise.

VTEP: VXLAN Termination End Point, as in [RFC7348].

VXLAN: Virtual Extensible LAN, as in [RFC7348].

This document also assumes familiarity with the terminology of [RFC7432], [RFC8365] and [RFC7365].

1 Introduction

The applications of EVPN-based solutions have become pervasive in Data Center, Service Provider, and Enterprise segments. It is being used for fabric overlays and inter-site connectivity in the Data Center market segment, for Layer-2, Layer-3, and IRB VPN services in the Service Provider market segment, and for fabric overlay and WAN connectivity in the Enterprise networks. For Data Center and Enterprise applications, there is a need to provide inter-site and WAN connectivity over public Internet in a secured manner with the same level of privacy, integrity, and authentication for tenant's traffic as used in IPsec tunneling using IKEv2. This document presents a solution where BGP point-to-multipoint signaling is leveraged for 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.

EVPN uses BGP as control-plane protocol for distribution of information needed for discovery of PEs participating in a VPN, discovery of PEs participating in a redundancy group, customer MAC addresses and IP prefixes/addresses, aliasing information, tunnel encapsulation types, multicast tunnel types, multicast group memberships, and other info. The advantages of using BGP control plane in EVPN are well understood including the following:

- 1) A full mesh of BGP sessions among PE devices can be avoided by using Route Reflector (RR) where a PE only needs to setup a single BGP session between itself and the RR as opposed to setting up N BGP sessions to N other remote PEs; therefore, reducing number of BGP sessions from $O(N^2)$ to $O(N)$ in the network. Furthermore, RR hierarchy can be leveraged to scale the number of BGP routes on the RR.
- 2) MP-BGP route filtering and constrained route distribution can be leveraged to ensure that the control-plane traffic for a given VPN is only distributed to the PEs participating in that VPN.

For setting up point-to-point security association (i.e., IPsec tunnel) between a pair of EVPN PEs, it is important to leverage BGP point-to-multipoint signaling architecture using the RR along with its route filtering and constrain mechanisms to achieve the performance and the scale needed for large number of security associations (IPsec tunnels) along with their frequent re-keying requirements. Using BGP signaling along with the RR (instead of peer-to-peer protocol such as IKEv2) reduces number of message exchanges needed for SAs establishment and maintenance from $O(N^2)$ to $O(N)$ in the network.

2 Requirements

The requirements for secured EVPN are captured in the following subsections.

2.1 Tenant's Layer-2 and Layer-3 data & control traffic

Tenant's layer-2 and layer-3 data and control traffic must be protected by IPsec cryptographic methods. This implies not only tenant's data traffic must be protected by IPsec but also tenant's control and routing information that are advertised in BGP must also be protected by IPsec. This in turn implies that BGP session must be protected by IPsec.

2.2 Tenant's Unicast & Multicast Data Protection

Tenant's layer-2 and layer-3 unicast traffic must be protected by IPsec. In addition to that, tenant's layer-2 broadcast, unknown unicast, and multicast traffic as well as tenant's layer-3 multicast traffic must be protected by IPsec when ingress replication or assisted replication are used. The use of BGP P2MP signaling for setting up P2MP SAs in P2MP multicast tunnels is for future study.

2.3 P2MP Signaling for SA setup and Maintenance

BGP P2MP signaling must be used for IPsec SAs setup and maintenance. The BGP signaling must follow P2MP signaling framework per [CONTROLLER-IKE] for IPsec SAs setup and maintenance in order to reduce the number of message exchanges from $O(N^2)$ to $O(N)$ among the participant PE devices.

2.4 Granularity of Security Association Tunnels

The solution must support the setup and maintenance of IPsec SAs at the following level of granularities:

- 1) Per PE: A single IPsec tunnel between a pair of PEs to be used for all tenants' traffic supported by the pair of PEs.
- 2) Per tenant: A single IPsec tunnel per tenant per pair of PEs. For example, if there are 1000 tenants supported on a pair of PEs, then 1000 IPsec tunnels are required between that pair of PEs.
- 3) Per subnet: A single IPsec tunnel per subnet (e.g., per VLAN/EVI) of a tenant on a pair of PEs.
- 4) Per IP address: A single IPsec tunnel per pair of IP addresses of a tenant on a pair of PEs.

5) Per MAC address: A single IPsec tunnel per pair of MAC addresses of a tenant on a pair of PEs.

6) Per Attachment Circuit: A single IPsec tunnel per pair of Attachment Circuits between a pair of PEs.

2.5 Support for Policy and DH-Group List

The solution must support a single policy and DH group for all SAs as well as supporting multiple policies and DH groups among the SAs.

3 Solution Description

This solution uses BGP P2MP signaling where an originating PE only send a message to the Route Reflector (RR) and then the RR reflects that message to the interested recipient PEs. The framework for such signaling is described in [CONTROLLER-IKE] and it is referred to as device-to-controller trust model. This trust model is significantly different than the traditional peer-to-peer trust model where a P2P signaling protocol such as IKEv2 [RFC7296] is used in which the PE devices directly authenticate each other and agree upon security policy and keying material to protect communications between themselves. The device-to-controller trust model leverages P2MP signaling via the controller (e.g., the RR) to achieve much better scale and performance for establishment and maintenance of large number of pair-wise Security Associations (SAs) among the PEs.

This device-to-controller trust model first secures the control channel between each device and the controller using peer-to-peer protocol such as IKEv2 [RFC7296] to establish P2P SAs between each PE and the RR. It then uses this secured control channel for P2MP signaling in establishment of P2P SAs between each pair of PE devices.

Each PE advertises to other PEs via the RR the information needed in establishment of pair-wise SAs between itself and every other remote PEs. These pieces of information are sent as Sub-TLVs of IPsec tunnel type in BGP Tunnel Encapsulation attribute. These Sub-TLVs are detailed in section 5 and are based on the DIM message components from [CONTROLLER-IKE] and the IKEv2 specification [RFC7296]. The IPsec tunnel TLVs along with its Sub-TLVs are sent along with the BGP route (NLRI) for a given level of granularity.

If only a single SA is required per pair of PE devices to multiplex user traffic for all tenants, then IPsec tunnel TLV is advertised along with IPv4 or IPv6 NLRI representing loopback address of the

originating PE. It should be noted that this is not a VPN route but rather an IPv4 or IPv6 route.

If a SA is required per tenant between a pair of PE devices, then IPsec tunnel TLV can be advertised along with EVPN IMET route representing the tenant or can be advertised along with a new EVPN route representing the tenant.

If a SA is required per tenant's subnet (e.g., per VLAN) between a pair of PE devices, then IPsec tunnel TLV is advertised along with EVPN IMET route.

If a SA is required between a pair of tenant's devices represented by a pair of IP addresses, then IPsec tunnel TLV is advertised along with EVPN IP Prefix Advertisement Route or EVPN MAC/IP Advertisement route.

If a SA is required between a pair of tenant's devices represented by a pair of MAC addresses, then IPsec tunnel TLV is advertised along with EVPN MAC/IP Advertisement route.

If a SA is required between a pair of Attachment Circuits (ACs) on two PE devices (where an AC can be represented by <VLAN, port>), then IPsec tunnel TLV is advertised along with EVPN Ethernet AD route.

3.1 Inheritance of Security Policies

Operationally, it is easy to configure a security association between a pair of PEs using BGP signaling. This is the default security association that is used for traffic that flows between peers. However, in the event more finer granularity of security association is desired on the traffic flows, it is possible to set up SAs between a pair of tenants, a pair of subnets within a tenant, a pair of IPs between a subnet, and a pair of MACs between a subnet using the appropriate EVPN routes as described above. In the event, there are no security TLVs associated with an EVPN route, there is a strict order in the manner security associations are inherited for such a route. This results in an EVPN route inheriting the security associations of the parent in a hierarchical fashion. For example, traffic between an IP pair is protected using security TLVs announced along with the EVPN IP Prefix Advertisement Route or EVPN MAC/IP Advertisement route as a first choice. If such TLVs are missing with the associated route, then one checks to see if the subnets the IPs are associated with has security TLVs with the EVPN IMET route. If they are present, those associations are used in securing the traffic. In the absence of them, the peer security associations are used. The order in which security associations are inherited are from the granular to the coarser, namely, IP/MAC associated TLVs with the

EVPN route being the first preference, and the subnet, the tenant, and the peer associations preferred in that fashion.

It should be noted that when a security association is made it is possible for it to be re-used by a large number of traffic flows. For example, a tenant security association may be associated with a number of child subnet routes. Clearly it is mandatory to keep a tenant security association alive, if there are one or more subnet routes that want to use that association. Logically, the security associations between a pair of entities creates a single secure tunnel. It is thus possible to classify the incoming traffic in the most granular sense {IP/MAC, subnet, tenant, peer} to a particular secure tunnel that falls within its route hierarchy. The policy that is applied to such traffic is independent from its use of an existing or a new secure tunnel. It is clear that since any number of classified traffic flows can use a security association, such a security association will not be torn down, if at least there is one policy using such a secure tunnel.

3.2 Distribution of Public Keys and Policies

One of the requirements for this solution is to support a single DH group and a single policy for all SAs as well as to support multiple DH groups and policies among the SAs. The following subsections describe what pieces of information (what Sub-TLVs) are needed to be exchanged to support a single DH group and a single policy versus multiple DH groups and multiple policies.

3.2.1 Minimal DIM

For SA establishment, at the minimum, a PE needs to advertise to other PEs, its DIM values as specified in [CONTROLLER-IKE]. These include:

ID	Tunnel ID
N	Nonce
RC	Rekey Counter
I	Indication of initial policy distribution
KE	DH public value.

When this minimal set of DIM values is sent, then it is assumed that all peer PEs share the same policy for which DH group to use, as well as which IPsec SA policy to employ. Section 5.1 defines the Minimal DIM sub-TLV as part of IPsec tunnel TLV in BGP Tunnel Encapsulation Attribute.

3.2.2 Multiple Policies

There can be scenarios for which there is a need to have multiple policy options. This can happen when there is a need for policy change and smooth migration among all PE devices to the new policy is required. It can also happen if different PE devices have different capabilities within the network. In these scenarios, PE devices need to be able to choose the correct policy to use for each other. This multi-policy scheme is described in section 6 of [CONTROLLER-IKE]. In order to support this multi-policy feature, a PE device MUST distribute a policy list. This list consists of multiple distinct policies in order of preference, where the first policy is the most preferred one. The receiving PE selects the policy by taking the received list (starting with the first policy) and comparing that against its own list and choosing the first one found in common. If there is no match, this indicates a configuration error and the PEs MUST NOT establish new SAs until a message is received that does produce a match.

3.2.2.1 Multiple DH-groups

It can be the case that not all peers use the same DH group. When multiple DH groups are supported, the peer may include multiple KE Sub-TLVs. The order of the KE Sub-TLVs determines the preference. The preference and selection methods are specified in Section 6 of [CONTROLLER-IKE].

3.2.2.2 Multiple or Single ESP SA policies

In order to specify an ESP SA Policy, a DIM may include one or more SA Sub-TLVs. When all peers are configured by a controller with the same ESP SA policy, they MAY leave the SA out of the DIM. This minimizes messaging when group configuration is static and known. However, it may also be desirable to include the SA. If a single SA is included, the peer is indicating what ESP SA policy it uses, but is not willing to negotiate. If multiple SA Sub-TLVs are included, the peer is indicating that it is willing to negotiate. The order of the SA Sub-TLVs determines the preference. The preference and selection methods are specified in Section 6 of [CONTROLLER-IKE].

3.3 Initial IPsec SAs Generation

The procedure for generation of initial IPsec SAs is described in section 3 of [CONTROLLER-IKE]. This section gives a summary of it in context of BGP signaling. When a PE device first comes up and wants to setup an IPsec SA between itself and each of the interested remote PEs, it generates a DH pair along for each [what word here? "tenant"?] using an algorithm defined in the IKEv2 Diffie-Hellman

Group Transform IDs [IKEv2-IANA]. The originating PE distributes the DH public value along with the other values in the DIM (using IPsec Tunnel TLV in Tunnel Encapsulation Attribute) to other remote PEs via the RR. Each receiving PE uses this DH public number and the corresponding nonce in creation of IPsec SA pair to the originating PE - i.e., an outbound SA and an inbound SA. The detail procedures are described in section 5.2 of [CONTROLLER-IKE].

3.4 Re-Keying

A PE can initiate re-keying at any time due to local time or volume based policy or due to the result of cipher counter nearing its final value. The rekey process is performed individually for each remote PE. If rekeying is performed with multiple PEs simultaneously, then the decision process and rules described in this rekey are performed independently for each PE. Section 4 of [CONTROLLER-IKE] describes this rekeying process in details and gives examples for a single IPsec device (e.g., a single PE) rekey versus multiple PE devices rekey simultaneously.

3.5 IPsec Databases

The Peer Authorization Database (PAD), the Security Policy Database (SPD), and the Security Association Database (SAD) all need to be setup as defined in the IPsec Security Architecture [RFC4301]. Section 5 of [CONTROLLER-IKE] gives a summary description of how these databases are setup for the controller-based model where key is exchanged via P2MP signaling via the controller (i.e., the RR) and the policy can be either signaled via the RR (in case of multiple policies) or configured by the management station (in case of single policy).

4 Encapsulation

Vast majority of Encapsulation for Network Virtualization Overlay (NVO) networks in deployment are based on UDP/IP with UDP destination port ID indicating the type of NVO encapsulation (e.g., VxLAN, GPE, GENEVE, GUE) and UDP source port ID representing flow entropy for load-balancing of the traffic within the fabric based on n-tuple that includes UDP header. When encrypting NVO encapsulated packets using IP Encapsulating Security Payload (ESP), the following two options can be used: a) adding a UDP header before ESP header (e.g., UDP header in clear) and b) no UDP header before ESP header (e.g., standard ESP encapsulation). The following subsection describe these encapsulation in further details.

4.1 Standard ESP Encapsulation

When standard IP Encapsulating Security Payload (ESP) is used (without outer UDP header) for encryption of NVO packets, it is used in transport mode as depicted below. When such encapsulation is used, for BGP signaling, the Tunnel Type of Tunnel Encapsulation TLV is set to ESP-Transport and the Tunnel Type of Encapsulation Extended Community is set to NVO encapsulation type (e.g., VxLAN, GENEVE, GPE, etc.). This implies that the customer packets are first encapsulated using NVO encapsulation type and then it is further encapsulated & encrypted using ESP-Transport mode.

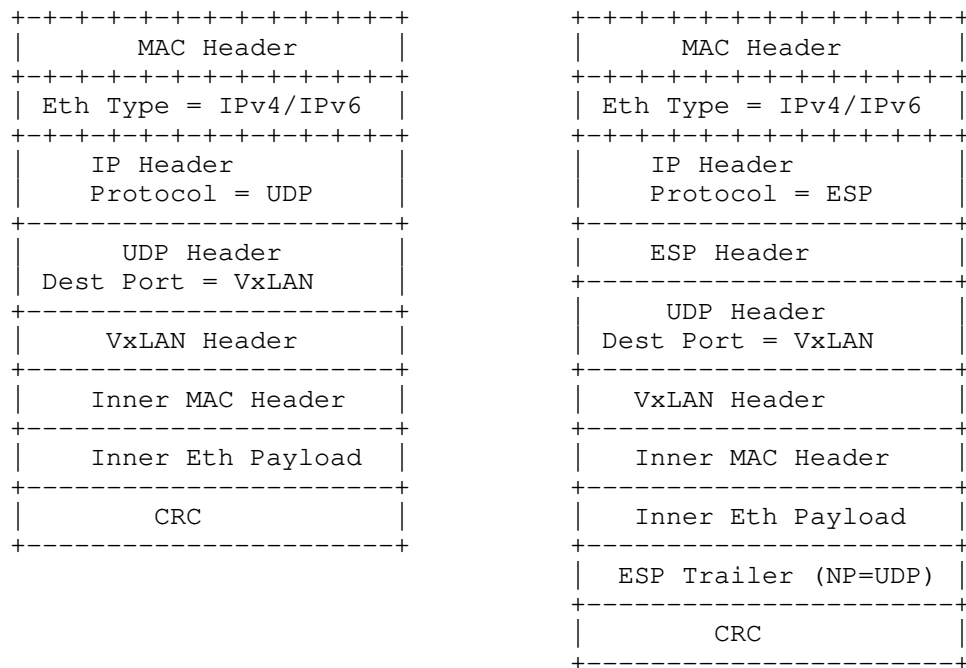


Figure 3: VxLAN Encapsulation within ESP

4.2 ESP Encapsulation within UDP packet

In scenarios where NAT traversal is required ([RFC3948]) or where load balancing using UDP header is required, then ESP encapsulation within UDP packet as depicted in the following figure is used. The ESP for NVO applications is in transport mode. The outer UDP header

(before the ESP header) has its source port set to flow entropy and its destination port set to 4500 (indicating ESP header follows). A non-zero SPI value in ESP header implies that this is a data packet (i.e., it is not an IKE packet). The Next Protocol field in the ESP trailer indicates what follows the ESP header, is a UDP header. This inner UDP header has a destination port ID that identifies NVO encapsulation type (e.g., VxLAN). Optimization of this packet format where only a single UDP header is used (only the outer UDP header) is for future study.

When such encapsulation is used, for BGP signaling, the Tunnel Type of Tunnel Encapsulation TLV is set to ESP-in-UDP-Transport and the Tunnel Type of Encapsulation Extended Community is set to NVO encapsulation type (e.g., VxLAN, GENEVE, GPE, etc.). This implies that the customer packets are first encapsulated using NVO encapsulation type and then it is further encapsulated & encrypted using ESP-in-UDP with Transport mode.

[RFC3948]

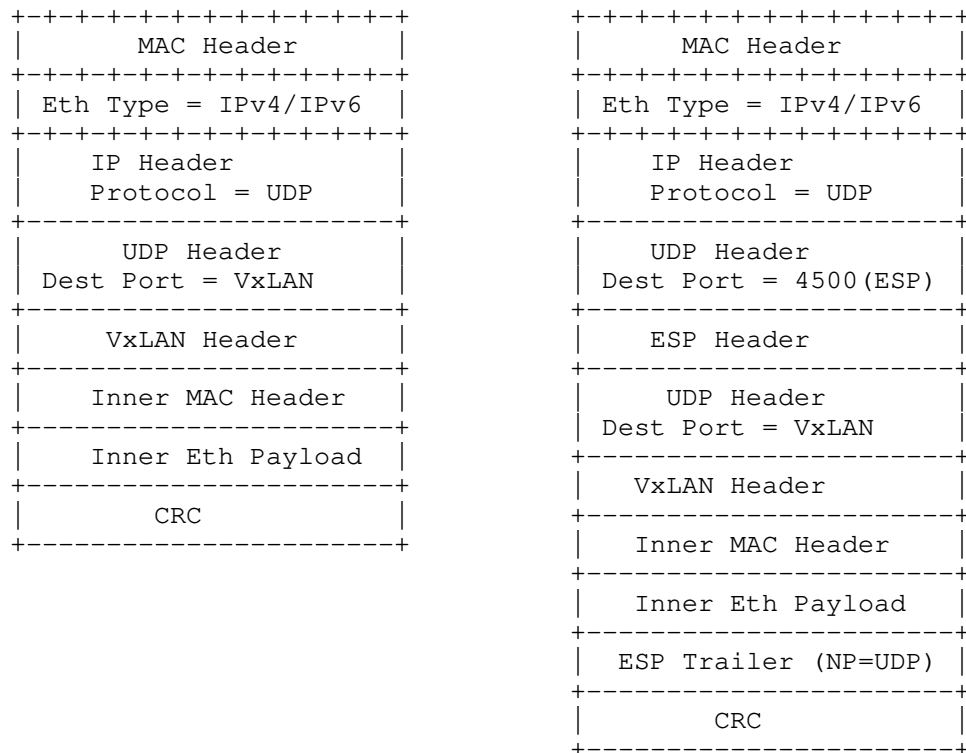


Figure 4: VxLAN Encapsulation within ESP Within UDP

5 BGP Encoding

This document defines two new Tunnel Types along with its associated sub-TLVs for The Tunnel Encapsulation Attribute [TUNNEL-ENCAP]. These tunnel types correspond to ESP-Transport and ESP-in-UDP-Transport as described in section 4. The following sub-TLVs apply to both tunnel types unless stated otherwise.

5.1 The Base (Minimal Set) DIM Sub-TLV

The Base DIM is described in 3.2.1. One and only one Base DIM may be sent in the IPSec Tunnel TLV.

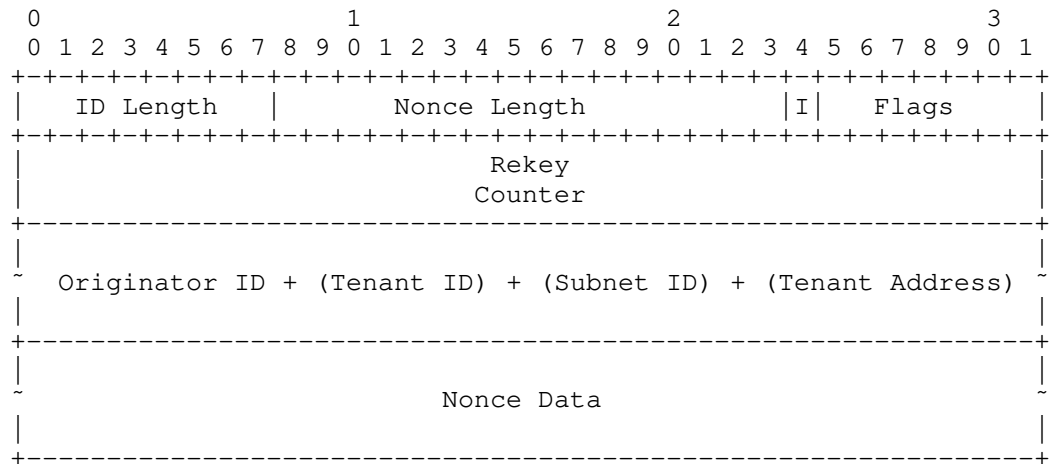


Figure 5: The Base DIM Sub-TLV

ID Length (16 bits) is the length of the Originator ID + (Tenant ID) + (Subnet ID) + (Tenant Address) in bytes.

Nonce Length (8 bits) is the length of the Nonce Data in bytes

I (1 bit) is the initial contact flag from [CONTROLLER-IKE]

Flags (7 bits) are reserved and MUST be set to zero on transmit and ignored on receipt.

The Rekey Counter is a 64 bit rekey counter as specified in [CONTROLLER-IKE]

The Originator ID + (Tenant ID) + (Subnet ID) + (Tenant Address) is the tunnel identifier and uniquely identifies the tunnel. Depending on the granularity of the tunnel, the fields in () may not be used - i.e., for a tunnel at the PE level of granularity, only Originator ID is required.

The Nonce Data is the nonce described in [CONTROLLER-IKE]. Its length is a multiple of 32 bits. Nonce lengths should be chosen to meet minimum requirements described in IKEv2 [RFC7296].

5.2 Key Exchange Sub-TLV

The KE Sub-TLV is described in 3.2.1 and 3.2.2.1. A KE is always required. One or more KE Sub-TLVs may be included in the IPSec Tunnel TLV.

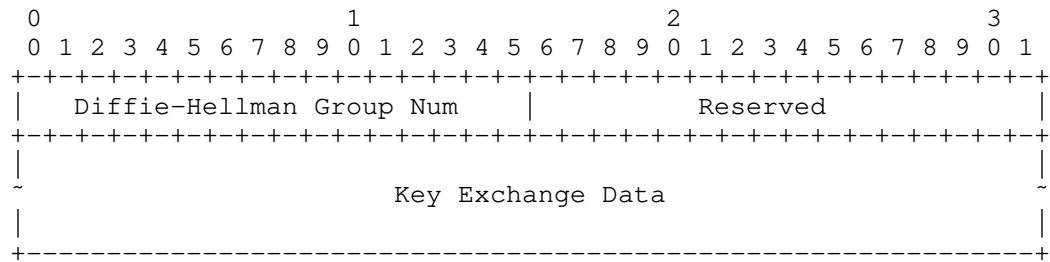


Figure 6: Key Exchange Sub-TLV

Diffie-Hellman Group Num 916 bits) identifies the Diffie-Hellman group in the Key Exchange Data was computed. Diffie-Hellman group numbers are discussed in IKEv2 [RFC7296] Appendix B and [RFC5114].

The Key Exchange payload is constructed by copying one's Diffie-Hellman public value into the "Key Exchange Data" portion of the payload. The length of the Diffie-Hellman public value is described for MOPD groups in [RFC7296] and for ECP groups in [RFC4753].

5.3 ESP SA Proposals Sub-TLV

The SA Sub-TLV is described in 3.2.2.2. Zero or more SA Sub-TLVs may be included in the IPSec Tunnel TLV.

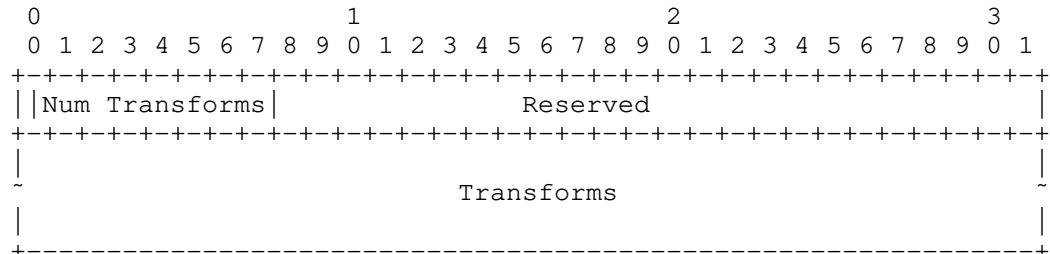


Figure 8: ESP SA Proposals Sub-TLV

Num Transforms is the number of transforms included.

Reserved is not used and MUST be set to zero on transmit and MUST be ignored on receipt.

5.3.1 Transform Substructure

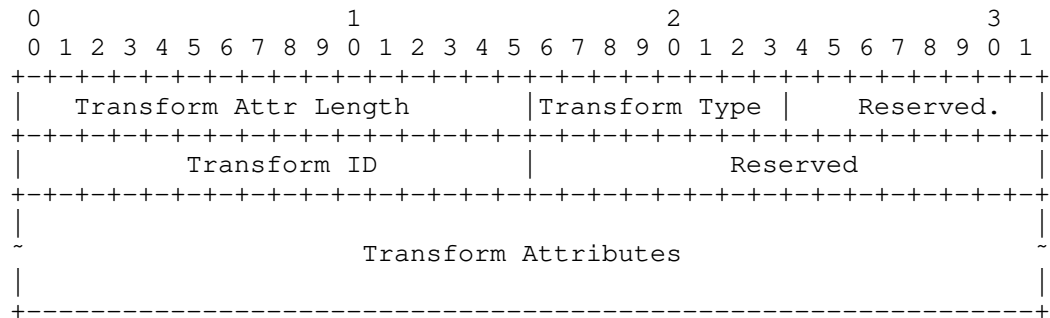


Figure 9: Transform Substructure Sub-TLV

The Transform Attr Length is the length of the Transform Attributes field.

The Transform Type is from Section 3.3.2 of [RFC7296] and [IKEV2IANA]. Only the values ENCR, INTEG, and ESN are allowed.

The Transform ID specifies the transform identification value from [IKEV2IANA].

Reserved is unused and MUST be zero on transmit and MUST be ignored on receipt.

The Transform Attributes are taken directly from 3.3.5 of [RFC7296].

6 Applicability to other VPN types

Although P2MP BGP signaling for establishment and maintenance of SAs among PE devices is described in this document in context of EVPN, there is no reason why it cannot be extended to other VPN technologies such as IP-VPN [RFC4364], VPLS [RFC4761] & [RFC4762], and MVPN [RFC6513] & [RFC6514] with ingress replication. The reason EVPN has been chosen is because of its pervasiveness in DC, SP, and Enterprise applications and because of its ability to support SA establishment at different granularity levels such as: per PE, Per tenant, per subnet, per Ethernet Segment, per IP address, and per MAC. For other VPN technology types, a much smaller granularity levels can be supported. For example for VPLS, only the granularity of per PE and per subnet can be supported. For per-PE granularity level, the mechanism is the same among all the VPN technologies as IPsec tunnel type (and its associated TLV and sub-TLVs) are sent along with the PE's loopback IPv4 (or IPv6) address. For VPLS, if per-subnet (per bridge domain) granularity level needs to be supported, then the IPsec tunnel type and TLV are sent along with

VPLS AD route.

The following table lists what level of granularity can be supported by a given VPN technology and with what BGP route.

Functionality	EVPN	IP-VPN	MVPN	VPLS
per PE	IPv4/v6 route	IPv4/v6 route	IPv4/v6 rte	IPv4/v6
per tenant	IMET (or new)	lpbk (or new)	I-PMSI	N/A
per subnet	IMET	N/A	N/A	VPLS AD
per IP	EVPN RT2/RT5	VPN IP rt	*,G or S,G	N/A
per MAC	EVPN RT2	N/A	N/A	N/A

7 Acknowledgements

8 Security Considerations

9 IANA Considerations

A new transitive extended community Type of 0x06 and Sub-Type of TBD for EVPN Attachment Circuit Extended Community needs to be allocated by IANA.

10 References

10.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017.

- [RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", RFC 7432, February, 2015.
- [RFC8365] Sajassi et al., "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, March, 2018.
- [TUNNEL-ENCAP] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-03, November 2016.
- [CONTROLLER-IKE] Carrel et al., "IPsec Key Exchange using a Controller", draft-carrel-ipsecme-controller-ike-00, July, 2018.
- [IKEV2IANA] IANA, "Internet Key Exchange Version 2 (IKEv2) Parameters", <<http://www.iana.org/assignments/ikev2-parameters/>>.
- [RFC3948] Huttunen et al., "UDP Encapsulation of IPsec ESP Packets", RFC 3948, January 2005.
- [IKEV2-IANA] IANA, "Internet Key Exchange Version 2 (IKEv2) Parameters", February 2016, www.iana.org/assignments/ikev2-parameters/ikev2-parameters.xhtml.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005.

10.2 Informative References

- [RFC4364] Rosen, E., et. al., "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4761] Kompella, K., et. al., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Kompella, K., et. al., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC6513] Rosen, E., et. al., "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Rosen, E., et. al., "BGP Encodings and Procedures for

Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

[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>>.

[802.1Q] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q(tm), 2014 Edition, November 2014.

[RFC7348] Mahalingam, M., et al., "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.

[GENEVE] Gross, J., et al., "Geneve: Generic Network Virtualization Encapsulation", Work in Progress, draft-ietf-nvo3-geneve-06, March 2018.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Ayan Banerjee
Cisco
Email: ayabaner@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com

David Carrel
Cisco
Email: carrel@cisco.com

Brian Weis
Cisco
Email: bew@cisco.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: September 7, 2019

W. Lin, Ed.
S. Sivaraj
V. Garg
Juniper Networks, Inc.
J. Rabadan
Nokia
March 6, 2019

Extended Procedures for EVPN Optimized Ingress Replication
draft-wsv-bess-extended-evpn-optimized-ir-01

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 September 7, 2019.

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-LEAFm. 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-mld-proxy-02.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: September 6, 2019

Z. Zhang
Y. Wang
G. Mirsky
ZTE Corporation
March 5, 2019

Bidirectional Forwarding Detection (BFD) for EVPN Ethernet Segment
Failover Use Case
draft-zwm-bess-es-failover-00.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 September 6, 2019.

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.

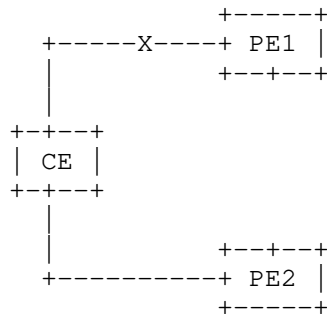
Table of Contents

1. Introduction	2
2. Proposal	3
3. Specification	3
4. Security Considerations	4
5. IANA Considerations	4
6. Normative References	4
Authors' Addresses	4

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.

[I-D.ietf-bess-evpn-df-election-framework] 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 multi-homed 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.jain-bess-evpn-lsp-ping].

2. 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.jain-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.jain-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 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.

3. Specification

[I-D.jain-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.

4. Security Considerations

This document does not introduce any new security considerations other than already discussed in [RFC7432] and [RFC5884].

5. IANA Considerations

There is no IANA consideration.

6. Normative References

- [I-D.ietf-bess-evpn-df-election-framework]
Rabadan, J., satyamoh@cisco.com, s., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for EVPN Designated Forwarder Election Extensibility", draft-ietf-bess-evpn-df-election-framework-09 (work in progress), January 2019.
- [I-D.jain-bess-evpn-lsp-ping]
Jain, P., Salam, S., Sajassi, A., Boutros, S., and G. Mirsky, "LSP-Ping Mechanisms for EVPN and PBB-EVPN", draft-jain-bess-evpn-lsp-ping-08 (work in progress), December 2018.
- [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>>.

Authors' Addresses

Zheng(Sandy) Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zzhang_ietf@hotmail.com

Yubao Wang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: wang.yubao2@zte.com.cn

Greg Mirsky
ZTE Corporation

Email: gregimirsky@gmail.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: June 28, 2019

Z. Zhang
L. Giuliano
Juniper Networks
K. Patel
Arrcus
I. Wijnands
M. Mishra
Cisco Systems
A. Gulko
Refinitiv
December 25, 2018

BGP Based Multicast
draft-zzhang-bess-bgp-multicast-02

Abstract

This document specifies a BGP address family and related procedures that allow BGP to be used for setting up multicast distribution trees. This document also specifies procedures that enable BGP to be used for multicast source discovery, and for showing interest in receiving particular multicast flows. Taken together, these procedures allow BGP to be used as a replacement for other multicast routing protocols, such as PIM or mLDp. The BGP procedures specified here are based on the BGP multicast procedures that were originally designed for use by providers of Multicast Virtual Private Network service.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 June 28, 2019.

Copyright 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 (<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
1.1.	Motivation	3
1.1.1.	Native/unlabeled Multicast	3
1.1.2.	Labeled Multicast	4
1.2.	Overview	4
1.2.1.	(x,g) Multicast	5
1.2.1.1.	Source Discovery for ASM	5
1.2.1.2.	ASM Shared-tree-only Mode	6
1.2.1.3.	Integration with BGP-MVPN	7
1.2.2.	BGP Inband Signaling for mLDP Tunnel	7
1.2.3.	BGP Sessions	7
1.2.4.	LAN and Parallel Links	8
1.2.5.	Transition	9
2.	Specification	10
2.1.	BGP NLRIs and Attributes	10
2.1.1.	S-PMSI A-D Route	11
2.1.2.	Leaf A-D Route	11
2.1.3.	Source Active A-D Route	12
2.1.4.	S-PMSI A-D Route for C-multicast mLDP	13
2.1.5.	Session Address Extended Community	13
2.2.	Procedures	14
2.2.1.	Source Discovery for ASM	14
2.2.2.	Originating Tree Join Routes	14
2.2.2.1.	(x,g) Multicast Tree	14
2.2.2.2.	BGP Inband Signaling for mLDP Tunnel	15
2.2.3.	Receiving Tree Join Routes	15

2.2.4. Withdrawl of Tree Join Routes	16
2.2.5. LAN procedures for (x,g) Unidirectional Tree	16
2.2.5.1. Originating S-PMSI A-D Routes	16
2.2.5.2. Receiving S-PMSI A-D Routes	17
2.2.6. Distributing Label for Upstream Traffic for Bidirectional Tree/Tunnel	17
3. Security Considerations	18
4. Acknowledgements	18
5. References	18
5.1. Normative References	18
5.2. Informative References	19
Authors' Addresses	20

1. Introduction

1.1. Motivation

This section provides some motivation for BGP signaling for native and labeled multicast. One target deployment would be a Data Center that requires multicast but uses BGP as its only routing protocol [RFC7938]. In such a deployment, it would be desirable to support multicast by extending the deployed routing protocol, without requiring the deployment of tree building protocols such as PIM, mLDP, RSVP-TE P2MP, and without requiring an IGP.

Additionally, compared to PIM, BGP based signaling has several advantage as described in the following section, and may be desired in non-DC deployment scenarios as well.

1.1.1. Native/unlabeled Multicast

Protocol Independent Multicast (PIM) has been the prevailing multicast protocol for many years. Despite its success, it has two drawbacks:

- o The ASM model, which is prevalent, introduces complexity in the following areas: source discovery procedures, need for Rendezvous Points (RPs) and group-to-RP mappings, need to switch between RP-rooted trees and source-rooted trees, etc.
- o Periodical protocol state refreshes due to soft state nature.

While PIM-SSM removes the complexity of PIM-ASM, it requires that multicast sources are known apriori. There have not been a good way of discovering sources, so its deployment has been limited. PIM-Port (PIM over Reliable Transport) solves the soft state issue, though its deployment has also been limited for two reasons:

- o It does not remove the ASM complexities.
- o In many of the scenarios where reliable transport is deemed important, BGP-based multicast (e.g. BGP-MVPN) has been used instead of PORT.

Partly because of the above mentioned problems, some Data Center operators have been avoiding deploying multicast in their networks.

BGP-MVPN [RFC6514] uses BGP to signal VPN customer multicast state over provider networks. It removes the above mentioned problems from the SP environment, and the deployment experiences have been encouraging. While RFC 6514 makes it possible for an SP to provide MVPN service without running PIM on its backbone, that RFC still assumes that PIM (or mLDP) runs on the PE-CE links. [draft-ietf-bess-mvpn-pe-ce] adapts the concept of BGP-MVPN to PE-CE links so that the use of PIM on the PE-CE links can be eliminated (though the PIM-ASM complexities still remains in the customer network), and this document extends it further to general topologies, so that they can be run on any router, as a replacement for PIM or mLDP.

With that, PIM can be completely eliminated from the network. PIM soft state is replaced by BGP hard state. For ASM, source specific trees are set up directly after simpler source discovery (data driven on FHRs and control driven elsewhere), all based on BGP. All the complexities related to source discovery and shared/source tree switch are also eliminated. Additionally, the trees can be setup with MPLS labels, with just minor enhancements in the signaling.

1.1.2. Labeled Multicast

There could be two forms of labeled multicast signaled by BGP. The first one is labeled (x,g) multicast where 'x' stands for either 's' or '*'. Basically, it is for BGP-signaled multicast tree as described in previous section but with labels. The second one is for mLDP tunnels with BGP signaling in part or whole through a BGP domain.

For both cases, BGP is used because other label distribution mechanisms like mLDP may not be desired by some operators. For example, a DC operator may prefer to have a BGP-only deployment.

1.2. Overview

1.2.1. (x,g) Multicast

PIM-like functionality is provided, using BGP-based join/prune signaling and BGP-based source discovery for ASM. The BGP-based join signaling supports both labeled multicast and IP multicast.

The same RPF procedures as in PIM are used for each router to determine the RPF neighbor for a particular source or RPA (in case of Bidirectional Tree). Except in the Bidirectional Tree case and a special case described in Section 1.2.1.2, no (*,G) join is used - LHR routers discover the sources for ASM and then join towards the sources directly. Data driven mechanisms like PIM Assert is replaced by control driven mechanisms (Section 1.2.4).

The joins are carried in BGP Updates with MCAST-TREE SAFI and S-PMSI/Leaf A-D routes defined in this document. The updates are targeted at the upstream neighbor by use of Route Targets. [Note - earlier version of this draft uses C-multicast route to send joins. We're now switching to S-PMSI/Leaf routes for three reasons. a) when the routes go through RRs, we have to distinguish different routes based on upstream router and downstream router. This leads to Leaf routes. b) for labeled bidirectional trees, we need to signal "upstream fec". S-PMSI suits this very well. c) we may want to allow the option of setting up trees from the roots instead of from the leaves. S-PMSI suits that very well.]

If the BGP updates carry labels (via Tunnel Encapsulation Attribute [I-D.ietf-idr-tunnel-encaps]), then (s,g) multicast traffic can use the labels. This is very similar to mLDP Inband Signaling [RFC6826], except that there are no corresponding "mLDP tunnels" for the PIM trees. Similar to mLDP, labeled traffic on transit LANs are point to point. Of course, traffic sent to receivers on a LAN by a LHR is native multicast.

For labeled bidirectional (*,g) trees, downstream traffic (away from the RPA) can be forwarded as in the (s,g) case. For upstream traffic (towards RPA), the upstream neighbor needs to advertise a label for its downstream neighbors. The same label that the upstream neighbor advertises to its upstream is the same one that it advertises to its downstreams, using an S-PMSI A-D route.

1.2.1.1. Source Discovery for ASM

This document does not support ASM via shared trees (aka RP Tree, or RPT) with one exception discussed in the next section. Instead, FHRs, LHRs, and optionally RRs work together to propagate/discover source information via control plane and LHRs join source specific Shortest Path Trees (SPT) directly.

A FHR originates Source Active A-D routes upon discovering sources for particular flows and advertise them to its peers. It is desired that the SA routes only reach LHRs that are interested in receiving the traffic. To achieve that, the SA routes carry an IPv4 or IPv6 address specific Route Target. The Global Administrator field is set the group address of the flow, and the Local Administrator field is set to 0. An LHR advertises Route Target Membership routes, with the Route Target field in the NLRI set according to the groups it wants to receive traffic for, as how a FHR encode the Route Target in its Source Active routes. The propagation of the SA routes is subject to cooperative export filtering as specified in [RFC4684] and referred to as RTC mechanism in this document. That way, the LHR only receives Source Active routes for groups that it is interested in.

Typically, a set of RRs are used and they maintains all Source Active routes but only distribute to interested LHRs on demand (upon receiving corresponding Route Target Membership routes, which are triggered on LHRs when they receive IGMP/MLD membership routes). The rest of the document assumes that RRs are used, even though that is not required.

1.2.1.2. ASM Shared-tree-only Mode

It may be desired that only a shared tree is used to distribute all traffic for a particular ASM group from its RP to all LHRs, as described in Section 4.1 "PIM Shared Tree Forwarding" of [RFC7438]. This will significantly cut down the number of trees and works out very well in certain deployment scenarios. For example, all the sources could be connected to the RP, or clustered close the to RP. In the latter case, either the path from FHRs to the RP do not intersect the shared tree so native forwarding can be used between the FHRs and the RP, or other means outside of this document could be used to forward traffic from FHRs to the RP.

For native forwarding from FHRs to the RP, SA routes may be used to announce the sources so that the RP can join source specific trees to pull traffic, but the group specific Route Target is not needed. The LHRs do not advertise the group specific Route Target Membership routes as they do not need the SA routes.

To establish the shared tree, (*,g) Leaf A-D routes are used as in the bidirectional tree case, though no forwarding state is established to forward traffic from downstream neighbors.

1.2.1.3. Integration with BGP-MVPN

For each VPN, the Source Active routes distribution in that VPN do not have to involve PE's at all unless there are sources/receivers directly connected to some PE's and they are independent of MVPN SA routes. For example, FHRs and LHRs establish BGP sessions with RRs of that particular VPN for the purpose of SA distribution.

After source discovery, BGP multicast signaling is done from LHRs towards the sources. When the signaling reaches an egress PE, BGP-MVPN signaling takes over, as if a PIM (s,g) join/prune was received on the PE-CE interface. When the BGP-MVPN signaling reaches the ingress PE, BGP multicast signaling as specified in this document takes over, similar to how BGP-MVPN triggers PIM (s,g) join/prune on PE-CE interfaces.

1.2.2. BGP Inband Signaling for mLDP Tunnel

Part of an (or the whole) mLDP tunnel can also be signaled via BGP and seamlessly integrated with the rest of mLDP tunnel signaled natively via mLDP. All the procedures are similar to mLDP except that the signaling is done via BGP. The mLDP FEC is encoded as the BGP NLRI, with MCAST-TREE SAFI and S-PMSI/Leaf A-D Routes for C-multicast mLDP defined in this document. The Leaf A-D routes correspond to mLDP Label Mapping messages, and the S-PMSI A-D routes are used to signal upstream FEC for MP2MP mLDP tunnels, similar to the bidirection (*,g) case.

1.2.3. BGP Sessions

In order for two BGP speakers to exchange MCAST-TREE NLRI, they must use BGP Capabilities Advertisement [RFC5492] to ensure that they both are capable of properly processing the MCAST-TREE NLRI. This is done as specified in [RFC4760], by using a capability code 1 (multiprotocol BGP) with an AFI of IPv4 (1) or IPv6 (2) and a SAFI of MCAST-TREE with a value to be assigned by IANA.

How the BGP peer sessions are provisioned, whether EBGp or IBGP, whether statically, automatically (e.g., based on IGP neighbor discovery), or programmably via an external controller, is outside the scope of this document.

In case of IBGP, it could be that every router peering with Route Reflectors, or hop by hop IBGP sessions could be used to exchange MCAST-TREE NLRIs for joins. In the latter case, unless desired otherwise for reasons outside of the scope of this document, the hop by hop IBGP sessions SHOULD only be used to exchange MCAST-TREE NLRIs.

When multihop BGP is used, a router advertises its local interface addresses, for the same purposes that the Address List TLV in LDP serves. This is achieved by advertising the interface address as host prefixes with IPv4/v6 Address Specific ECs corresponding to the router's local addresses used for its BGP sessions (Section 2.1.5).

Because the BGP Capability Advertisement is only between two peers, when the sessions are only via RRs, a router needs another way to determine if its neighbor is capable of signaling multicast via BGP. The interface address advertisement can be used for that purpose - the inclusion of a Session Address EC indicates that the BGP speaker identified in the EC supports the C-Multicast NLRI.

FHRs and LHRs may also establish BGP sessions to some Route Reflectors for source discovery purpose (Section 1.2.1.1).

With the traditional PIM, the FHRs and LHRs refer to the PIM DRs on the source or receiver networks. With BGP based multicast, PIM may not be running at all, and the FHRs and LHRs refer to the IGMP/MLD queriers, or the DF elected per [I-D.wijnands-bier-mld-lan-election]. Alternatively, if it is known that a network only has senders then no IGMP/MLD or DF election is needed - any router may generate SA routes. That will not cause any issue other than redundant SA routes being originated.

1.2.4. LAN and Parallel Links

There could be parallel links between two BGP peers. A single multihop session, whether IBGP or EBGP, between loopback addresses may be used. Except for LAN interfaces in case of unlabeled (x,g) unidirectional trees (note that transit LAN interface is not supported for BGP signaled (*,g) bidirectional tree and for mLDP tunnels, traffic on transit LAN is point to point between neighbors), any link between the two peers can be automatically used by a downstream peer to receive traffic from the upstream peer, and it is for the upstream peer to decide which link to use. If one of the links goes down, the upstream peer switches to a different link and there is no change needed on the downstream peer.

For unlabeled (x,g) unidirectional trees, the upstream peer MAY prefer LAN interfaces to send traffic, since multiple downstream peers may be reached simultaneously, or it may make a decision based on local policy, e.g., for load balancing purpose. Because different downstream peers might choose different upstream peers for RPF, when an upstream peer decides to use a LAN interface to send traffic, it originates an S-PMSI A-D route indicating that one or more LAN interface will be used. The route carries Route Targets specific to the LANs so that all the peers on the LANs import the route. If more

than one router originate the route specifying the same LAN for the same (s,g) or (*,g) flow, then assert procedure based on the S-PMSI A-D routes happens and assert losers will stop sending traffic to the LAN.

1.2.5. Transition

A network currently running PIM can be incrementally transitioned to BGP based multicast. At any time, a router supporting BGP based multicast can use PIM with some neighbors (upstream or downstream) and BGP with some other neighbors. PIM and BGP MUST not be used simultaneously between two neighbors for multicast purpose, and routers connected to the same LAN MUST be transitioned during the same maintenance window.

In case of PIM-SSM, any router can be transitioned at any time (except on a LAN all routers must be transitioned together). It may receive source tree joins from a mixed set of BGP and PIM downstream neighbors and send source tree joins to its upstream neighbor using either PIM or BGP signaling.

In case of PIM-ASM, the RPs are first upgraded to support BGP based multicast. They learn sources either via PIM procedures from PIM FHRs, or via Source Active A-D routes from BGP FHRs. In the former case, the RPs can originate proxy Source Active A-D routes. There may be a mixed set of RPs/RRs - some capable of both traditional PIM RP functionalities while some only redistribute SA routes.

Then any routers can be transitioned incrementally. A transitioned LHR router will pull Source Active A-D routes from the RPs/RRs when they receive IGMP/MLD (*,G) joins for ASM groups, and may send either PIM (s,g) joins or BGP Source Tree Join routes. A transitioned transit router may receive (*,g) PIM joins but only send source tree joins after pulling Source Active A-D routes from RPs/RRs.

Similarly, a network currently running mLDP can be incrementally transitioned to BGP signaling. Without the complication of ASM, any router can be transitioned at any time, even without the restriction of coordinated transition on a LAN. It may receive mixed mLDP label mapping or BGP updates from different downstream neighbors, and may exchange either mLDP label mapping or BGP updates with its upstream neighbors, depending on if the neighbor is using BGP based signaling or not.

2. Specification

2.1. BGP NLRI's and Attributes

The BGP Multiprotocol Extensions [RFC4760] allow BGP to carry routes from multiple different "AFI/SAFIs". This document defines a new a new SAFI known as a MCAST-TREE SAFI with a value to be assigned by the IANA. This SAFI is used along with the AFI of IPv4 (1) or IPv6 (2).

The MCAST-TREE NLRI defined below is carried in the BGP UPDATE messages [RFC4271] using the BGP multiprotocol extensions [RFC4760] with a AFI of IPv4 (1) or IPv6 (2) assigned by IANA and a MCAST-TREE SAFI with a value to be assigned by the IANA.

The Next hop field of MP_REACH_NLRI attribute SHALL be interpreted as an IPv4 address whenever the length of the Next Hop address is 4 octets, and as an IPv6 address whenever the length of the Next Hop is address is 16 octets.

The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix with a maximum length of 12 octets for IPv4 AFI and 36 octets for IPv6 AFI. The following is the format of the MCAST-TREE NLRI:

```

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

The Route Type field defines encoding of the rest of the MCAST-TREE NLRI. (Route Type specific MCAST-TREE NLRI).

The Length field indicates the length in octets of the Route Type specific field of MCAST-TREE NLRI.

The following new route types are defined:

- 3 - S-PMSI A-D Route for (x,g)
- 4 - Leaf A-D Route
- 5 - Source Active A-D Route
- 0x43 - S-PMSI A-D Route for C-multicast mLDP

Except for the Source Active A-D routes, the routes are to be consumed by targeted upstream/downstream neighbors, and are not

propagated further. This can be achieved by outbound filtering based on the RTs that lead to the importation of the routes.

The Type-3/4 routes MAY carry a Tunnel Encapsulation Attribute (TEA) [I-D.ietf-idr-tunnel-encaps]. The Type-0x43 route MUST carry a TEA. When used for mLDP, the Type-4 route MUST carry a TEA. Only the MPLS tunnel type for the TEA is considered. Others are outside the scope of this document.

2.1.1. S-PMSI A-D Route

Similar to defined in RFC 6514, an S-PMSI A-D Route Type specific MCAST-TREE NLRI consists of the following, though it does not have an RD:

-----+-----
Multicast Source Length (1 octet)
-----+-----
Multicast Source (variable)
-----+-----
Multicast Group Length (1 octet)
-----+-----
Multicast Group (variable)
-----+-----
Upstream Router's IP Address
-----+-----

If the Multicast Source (or Group) field contains an IPv4 address, then the value of the Multicast Source (or Group) Length field is 32. If the Multicast Source (or Group) field contains an IPv6 address, then the value of the Multicast Source (or Group) Length field is 128.

Usage of other values of the Multicast Source Length and Multicast Group Length fields is outside the scope of this document.

There are two usages for S-PMSI A-D route. They're described in Section 2.2.5 and Section 2.2.6 respectively.

2.1.2. Leaf A-D Route

Similar to the Leaf A-D route in [RFC6514], a MCAST-TREE Leaf A-D route's route key includes the corresponding S-PMSI NLRI, plus the Originating Router's IP Addr. The difference is that there is no RD.

	S-PMSI NLRI	
	Originating Router's IP Address	

For example, the entire NLRI of a Leaf A-D route for (x,g) tree is as following:

			Route Type - 4 (Leaf A-D)			
			Length (1 octet)			
L E A F N L R I	L E A F		Route Type - 3 (S-PMSI A-D)		S P M S I N L R I	
			Length (1 octet)			
	R O U T E		Multicast Source Length (1 octet)			
			Multicast Source (variable)			
	K E Y		Multicast Group Length (1 octet)			
			Multicast Group (variable)			
			Upstream Router's IP Address			
			Originating Router's IP Address			

Even though the MCAST-TREE Leaf A-D route is unsolicited, unlike the Leaf A-D route for GTM in [RFC7524], it is encoded as if a corresponding S-PMSI A-D route had been received.

When used for signaling mLDP tunnels, even though the Leaf A-D route is unsolicited, unlike the "Route-type 0x44 Leaf A-D route for C-multicast mLDP" as in [RFC7441], it is Route-type 4 and encoded as if a corresponding S-PMSI A-D route had been received.

2.1.3. Source Active A-D Route

Similar to defined in RFC 6514, a Source Active A-D Route Type specific MCAST NLRI consists of the following:

Multicast Source Length (1 octet)

Multicast Source (variable)

Multicast Group Length (1 octet)

Multicast Group (variable)

The definition of the source/length and group/length fields are the same as in the S-PMSI A-D routes.

Usage of Source Active A-D routes is described in Section 1.2.1.1.

2.1.1.4. S-PMSI A-D Route for C-multicast mLDP

The route is used to signal upstream FEC for an MP2MP mLDP tunnel. The route key include the mLDP FEC and the Upstream Router's IP Address field. The encoding is similar to the same route in [RFC7441], though there is no RD.

2.1.1.5. Session Address Extended Community

For two BGP speakers to determine if they are directly connected, each will advertise their local interface addresses, with an Session Address Extended Community. This is an Address Specific EC, with the Global Admin Field set to the local address used for its multihop sessions and the Local Admin Field set to the prefix length corresponding to the interface's network mask.

For example, if a router has two interfaces with address 10.10.10.1/24 and 10.12.0.1/16 respectively (notice the different network mask), and a loopback address 11.11.11.1/32 that is used for BGP sessions, then it will advertise prefix 10.10.10.1/32 with a Session Address EC 11.11.11.1:24 and 10.12.0.1/32 with a Session Address EC 11.11.11.1:16. If it also uses another loopback address 11.11.11.11/32 for other BGP sessions, then the routes will additionally carry Session Address EC 11.11.11.11:24 and 11.11.11.11:16 respectively.

This achieves what the Address List TLV in LDP Address Messages achieves, and can also be used to indicate that a router supports the BGP multicast signaling procedures specified in this document.

Only those interface addresses that will be used as resolved nexthops in the RIB need to be advertised with the Session Address EC. For example, the RPF lookup may say that the resolved nexthop address is

A1, so the router needs to find out the corresponding BGP speaker with address A1 through the (interface address, session address) mapping built according to the interface address NLRI with the Session Address EC. For comparison with LDP, this is done via the (interface address, session address) mapping that is built by the LDP Address Messages.

2.2. Procedures

2.2.1. Source Discovery for ASM

When a FHR first receives a multicast packet addressed to an ASM group, it originates a Source Active route. It carries a IP/IPv6 Address Specific RT, with the Global Admin Field set to the group address and the Local Admin Field set to 0. The route is advertised to its peers, who will re-advertise further based on the RTC mechanisms. Note that typically the route is advertised only to the RRs.

The FHRs withdraws the Source Active route after a certain amount of time since it last received a packet of an (s,g) flow. The amount of time to wait is a local matter.

2.2.2. Originating Tree Join Routes

Note that in this document, tree join routes are S-PMSI/Leaf A-D routes.

2.2.2.1. (x,g) Multicast Tree

When a router learns from IGMP/MLD or a downstream PIM/BGP peer that it needs to join a particular (s,g) tree, it determines the RPF nexthop address wrt the source, following the same RPF procedures as defined for PIM. It further finds the BGP router that advertised the nexthop address as one of its local addresses.

If the RPF neighbor supports MCAST-TREE SAFI, this router originates a Leaf A-D route. Although it is unsolicited, it is constructed as if there was a corresponding S-PMSI A-D route. The Upstream Router's IP Address field is set to the RPF neighbor's session address (learnt via the EC attached to the host route for the RPF nexthop address). An Address Specific RT corresponding to the session address is attached to the route, with the Global Administrative Field set to the session address and the local administrative field set to 0.

Similarly, when a router learns that it needs to join a bi-directional tree for a particular group, it determines the RPF neighbor wrt the RPA. If the neighbor supports MCAST-TREE SAFI, it

originates a Leaf A-D Route and advertises the route to the RPF neighbor (in case of EBGP or hop-by-hop IBGP), or one or more RRs.

When a router first learns that it needs to receive traffic for an ASM group, either because of a local (*,g) IGMP/MLD report or a downstream PIM (*,g) join, it originates a RTC route with the NLRI's AS field set to its AS number and the Route Target field set to an address based RT, with the Global Administrator field set to group address and the Local Administrator field set to 0. The route is advertised to its peers (most practically some RRs), so that the router can receive matching Source Active A-D routes. Upon the receiving of the Source Active A-D routes, the router originates Leaf A-D routes as described above, as long as it still needs to receive traffic for the flows (i.e., the corresponding IGMP/MLD membership exists or join from downstream PIM/BGP neighbor exists).

When a Leaf A-D route is originated by this router, it sets up corresponding forwarding state such that the expected incoming interface list includes all non-LAN interfaces directly connecting to the upstream neighbor. LAN interfaces are added upon receiving corresponding S-PMSI A-D route (Section 2.2.5.2). If the upstream neighbor is not directly connected, tunnels may be used - details to be included in future revisions.

When the upstream nbr changes, the previously advertised Leaf A-D route is withdrawn. If there is a new upstream neighbor, a new Leaf A-D route is originated, corresponding to the new neighbor. Because NLRIs are different for the old and new Leaf A-D routes, make-before-break can be achieved, so can MoFRR [RFC7431].

2.2.2.2. BGP Inband Signaling for mLDP Tunnel

The same mLDP procedures as defined in [RFC6388] are followed, except that where a label mapping message is sent in [RFC6388], a Leaf A-D route is sent if the the upstream neighbor supports BGP based signaling.

2.2.3. Receiving Tree Join Routes

A router (auto-)configures Import RTs matching itself so that it can import tree join routes from their peers. Note that in this document, tree join routes are S-PMSI/Leaf A-D routes.

When a router receives a tree join route and imports it, it determines if it needs to originate its own corresponding route and advertise further upstream wrt the source/RPA or mLDP tunnel root. If itself is the FHR or is on the RPL or is the tunnel root, then it

does not need to. Otherwise the procedures in Section 2.2.2 are followed.

Additionally, the router sets up its corresponding forwarding state such that traffic will be sent to the downstream neighbor, and received from the downstream neighbor in case of bidirectional tree/tunnel. If the downstream neighbor is not directly connected, tunnels may be used - details to be included in future revisions.

2.2.4. Withdrawal of Tree Join Routes

For a particular tree or tunnel, if a downstream neighbor withdraws its Leaf A-D route, the neighbor is removed from the corresponding forwarding state. If all downstream neighbors withdraw their tree join routes and this router no longer has local receivers, it withdraws the tree join routes that it previously originated.

As mentioned earlier, when the upstream neighbor changes, the previously advertised Leaf A-D route is also withdrawn. The corresponding incoming interfaces are also removed from the corresponding forwarding state.

2.2.5. LAN procedures for (x,g) Unidirectional Tree

For a unidirectional (x,g) multicast tree, if there is a LAN interface connecting to the downstream neighbor, it MAY be preferred over non-LAN interfaces, but an S-PMSI A-D route MUST be originated to facilitate the analog of the Assert process (Section 2.2.5.1).

2.2.5.1. Originating S-PMSI A-D Routes

If this router chooses to use a LAN interface to send traffic to its neighbors for a particular (s,g) or (*,g) flow, it MUST announce that by originating a corresponding S-PMSI A-D route. The Tunnel Type in the PMSI Tunnel Attribute (PTA) is set to 0 (no tunnel information Present). The LAN interface is identified by an IP address specific RT, with the Global Administrative Field set to the LAN interface's address prefix and the Local Administrative Field set to the prefix length. The RT also serves the purpose of restricting the importing of the route by all routers on the LAN. An operator MUST ensure that RTs encoded as above are not used for other purposes. Practically that should not be unreasonable.

If multiple LAN interfaces are to be used (to reach different sets of neighbors), then the route will include multiple RTs, one for each used LAN interface as described above.

The S-PMSI A-D routes may also be used to announce tunnels that could be used to send traffic to downstream neighbors that are not directly connected. Details may be added in future revisions.

2.2.5.2. Receiving S-PMSI A-D Routes

A router (auto-)configures an Import RT for each of its LAN interfaces over which BGP is used for multicast signaling. The construction of the RT is described in the previous section.

When a router R1 imports an S-PMSI A-D route for flow (x,g) from router R2, R1 checks to see if it also originates an S-PMSI A-D route with the same NLRI except the Upstream Router's IP Address field. When a router R1 originates an S-PMSI A-D route, it checks to see if it also has installed an S-PMSI A-D route, from some other router R2, with the same NLRI except the Upstream Router's IP Address field. In either case, R1 checks to see if the two routes have an RT in common and the RT is encoded as in Section 2.2.5.1. If so, then there is a LAN attached to both R1 and R2, and both routers are prepared to send (S,G) traffic onto that LAN. This kicks off the assert procedure to elect a winner - the one with the highest Upstream Router's IP Address in the NLRI wins. An assert loser will not include the corresponding LAN interface in its outgoing interface list, but it keeps the S-PMSI A-D route that it originates.

If this router does not have a matching S-PMSI route of its own with some common RTs, and the originator of the received S-PMSI route is a chosen upstream neighbor for the corresponding flow, then this router updates its forwarding state to include the LAN interface in the incoming interface list. When the last S-PMSI route with a RT matching the LAN is withdrawn later, the LAN interface is removed from the incoming interface list.

Note that a downstream router on the LAN does not participate in the assert procedure. It adds/keeps the LAN interface in the expected incoming interfaces as long as its chosen upstream peer originates the S-PMSI AD route. It does not switch to the assert winner as its upstream. An assert loser MAY keep sending joins upstream based on local policy even if it has no other downstream neighbors (this could be used for fast switch over in case the assert winner would fail).

2.2.6. Distributing Label for Upstream Traffic for Bidirectional Tree/Tunnel

For MP2MP mLDP tunnels or labeled (*,g) bidirectional trees, an upstream router needs to advertise a label to all its downstream neighbors so that the downstream neighbors can send traffic to itself.

For MP2MP mLDP tunnels, the same procedures for mLDP are followed except that instead of MP2MP-U Label Mapping messages, S-PMSI A-D Routes for C-Multicast mLDP are used.

For labeled (*,g) bidirectional trees, for a Leaf A-D route received from a downstream neighbor, a corresponding S-PMSI A-D route is sent back to the downstream router.

In both cases, a single S-PMSI A-D route is originated for each tree from this router, but with multiple RTs (one for each downstream neighbor on the tree). A TEA specifies a label allocated by the upstream router for its downstream neighbors to send traffic with. Note that this is still a "downstream allocated" label (the upstream router is "downstream" from traffic direction point of view).

The S-PMSI routes do not carry a PTA, unless a P2MP tunnel is used to reach downstream neighbors. Such use case is out of scope of this document for now and may be specified in the future.

3. Security Considerations

This document does not introduce new security risks?

4. Acknowledgements

The authors thank Marco Rodrigues for his initial idea/ask of using BGP for multicast signaling beyond MVPN. We thank Eric Rosen for his questions, suggestions, and help finding solutions to some issues. We also thank Luay Jalil and James Uttaro for their comments and support for the work.

5. References

5.1. Normative References

- [I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10 (work in progress), August 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>>.

- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <<https://www.rfc-editor.org/info/rfc4601>>.
- [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>>.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <<https://www.rfc-editor.org/info/rfc5015>>.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <<https://www.rfc-editor.org/info/rfc6514>>.
- [RFC7441] Wijnands, IJ., Rosen, E., and U. Joerde, "Encoding Multipoint LDP (mLDP) Forwarding Equivalence Classes (FECs) in the NLRI of BGP MCAST-VPN Routes", RFC 7441, DOI 10.17487/RFC7441, January 2015, <<https://www.rfc-editor.org/info/rfc7441>>.

5.2. Informative References

- [I-D.ietf-bess-mvpn-pe-ce] Patel, K., Rosen, E., and Y. Rekhter, "BGP as an MVPN PE-CE Protocol", draft-ietf-bess-mvpn-pe-ce-01 (work in progress), October 2015.
- [I-D.wijnands-bier-mld-lan-election] Wijnands, I., Pfister, P., and Z. Zhang, "Generic Multicast Router Election on LAN's", draft-wijnands-bier-mld-lan-election-01 (work in progress), July 2016.
- [RFC6826] Wijnands, IJ., Ed., Eckert, T., Leymann, N., and M. Napierala, "Multipoint LDP In-Band Signaling for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6826, DOI 10.17487/RFC6826, January 2013, <<https://www.rfc-editor.org/info/rfc6826>>.

- [RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <<https://www.rfc-editor.org/info/rfc7431>>.
- [RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <<https://www.rfc-editor.org/info/rfc7938>>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Lenny Giuliano
Juniper Networks

EMail: lenny@juniper.net

Keyur Patel
Arrcus

EMail: keyur@arrcus.com

IJsbrand Wijnands
Cisco Systems

EMail: ice@cisco.com

Mankamana Mishra
Cisco Systems

EMail: mankamis@cisco.com

Arkadiy Gulko
Refinitiv

EMail: arkadiy.gulko@refinitiv.com

BESS
Internet-Draft
Updates: 6513, 6514 (if approved)
Intended status: Standards Track
Expires: January 9, 2017

Z. Zhang
R. Kebler
W. Lin
E. Rosen
Juniper Networks
July 8, 2016

MVPN/EVPN C-Multicast Routes Enhancements
draft-zzhang-bess-mvpn-evpn-cmcast-enhancements-00

Abstract

[RFC6513] and [RFC6514] specify procedures for originating, propagating, and processing "C-multicast routes". However, there are a number of MVPN use cases that are not properly or optimally handled by those procedures. This document describes those use cases, and specifies the additional procedures needed to handle them. Some of the additional procedures are also applicable to EVPN SMET routes [I-D.sajassi-bess-evpn-igmp-mld-proxy].

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 9, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Terminology	3
1.2. MVPN C-Bidir Support with VPN Backbone being RPL	3
1.2.1. C-multicast Routes for the MVPN-RPL Method of C-BIDIR support	4
1.2.2. Optional use of MVPN-RPL RD with mLDP/PIM Provider Tunnels	5
1.2.3. MVPN C-ASM Support without CE Routers	6
1.3. Inter-AS Propagation of MVPN C-Multicast Routes	6
1.4. EVPN Selective Multicast Ethernet Tag (SMET) Routes	8
1.5. Provider Tunnel Segmentation with Explicit-Tracking C-Multicast Routes	9
1.5.1. Conventional Tunnel Segmentation	9
1.5.2. Selective Tunnel Segmentation with Untargeted Explicit-Tracking C-multicast Routes	9
2. Specifications	10
2.1. MVPN C-Bidir Support with VPN Backbone being RPL	10
2.1.1. Constructing C-Multicast Share Tree Join route	10
2.1.2. Setting Up the MVPN-RPL	12
2.2. Inter-AS Propagation of MVPN C-Multicast Routes	12
2.2.1. Procedures in Section 11.2 of [RFC6514]	12
2.2.2. Ordinary BGP Propagation Procedures	13
2.3. Provider Tunnel Segmentation with Explicit-Tracking C-Multicast Routes	13
2.3.1. Egress PEs and RBRs	14
2.3.2. Transit RBRs	15
2.3.3. Ingress RBRs	15
2.3.4. Setting Up Forwarding State on RBRs	16
2.3.5. Other Types of Tunnels	16
3. Security Considerations	16
4. Acknowledgements	17

5. References	17
5.1. Normative References	17
5.2. Informative References	18
Authors' Addresses	18

1. Introduction

[RFC6513] and [RFC6514] specify procedures for originating, propagating, and processing "C-multicast routes". However, there are a number of MVPN use cases that are not properly or optimally handled by those procedures. This document describes those use cases, and specifies the additional procedures needed to handle them.

Some of the additional procedures are also applicable to EVPN SMET routes [I-D.sajassi-bess-evpn-igmp-mld-proxy]; this is discussed in Section 1.4.

1.1. Terminology

This document uses terminology from MVPN and EVPN. It is expected that the audience is familiar with the concepts and procedures defined in [RFC6513], [RFC6514], [RFC7524], [RFC7432], [I-D.zzhang-bess-evpn-bum-procedure-updates], and [I-D.sajassi-bess-evpn-igmp-mld-proxy]. Some terms are listed below for references.

- o PMSI: P-Multicast Service Interface - a conceptual interface for a PE to send customer multicast traffic to all or some PEs in the same VPN.
- o I-PMSI: Inclusive PMSI - to all PEs in the same VPN.
- o S-PMSI: Selective PMSI - to some of the PEs in the same VPN.
- o C-G-BIDIR: A bidirectional multicast group address (i.e., a group address whose IP multicast distribution tree is built by BIDIR-PIM) in customer address space.
- o RBR: Regional Border Router. A provider tunnel could be segmented, with one segment in each region. A region could be an AS, an IGP area, or even a subarea.

1.2. MVPN C-Bidir Support with VPN Backbone being RPL

In BIDIR-PIM [RFC5015], every group is associated with a "Rendezvous Point Link" (RPL). The RPL for a given group G is at the root of the BIDIR-PIM distribution tree. Links of the distribution tree that lead towards the RPL are considered to be "upstream" links, and links that lead away from the RPL are considered to be "downstream" links.

Every node on the distribution tree has one upstream link and zero or more downstream links.

Data addressed to a BIDIR-PIM group may enter the distribution tree at any node. The entry node sends the data on the upstream links and the downstream links. A node that receives the data from a downstream link sends it on its upstream link and on its other downstream links. A node that receives the data from its upstream link sends it on its downstream links. When a node that is attached to the RPL receives data from a downstream link, it forwards the data onto the RPL (as well as onto any other downstream links.) When node attached to the RPL receives data from the RPL, it forwards the data downstream.

The above is a simplified description, and ignores the fact that every link except the RPL has a Designated Forwarder (DF). Only the DF forwards traffic onto the link. However, the RPL has no DF; any node can forward traffic onto the RPL.

1.2.1. C-multicast Routes for the MVPN-RPL Method of C-BIDIR support

Section 11.1 of [RFC6513] describes a method of providing MVPN support for customers that use BIDIR-PIM. This is known as "MVPN C-BIDIR support". In this method of C-BIDIR support, the VPN backbone itself functions as the RPL. Thus this method is known as the "MVPN-RPL" method. The RPL is actually an I-PMSI or S-PMSI. The PE routers treat the I-PMSI or S-PMSI as their upstream link, and treat their VRF interfaces as downstream links.

If the MVPN-RPL method of C-BIDIR support is being used in a particular MVPN, all the PEs attached to that MVPN must be provisioned to use this method.

In the context of a given VPN, a PE with interest in receiving a particular C-BIDIR group (call it C-G-BIDIR) advertises this interest to the other PEs by originating a C-multicast Shared Tree Join route. When any PE receives traffic for the C-G-BIDIR on its PE-CE interface, it sends the data to the MVPN-RPL if and only if it has received corresponding (C-*,C-G-BIDIR) C-multicast Shared Tree Join route. Other PEs receive the traffic on the MVPN-RPL and forward to their downstream receivers. However, the procedure for constructing the C-multicast Shared Tree Join route in this case is not fully specified in [RFC6513] or [RFC6514]. The proper set of procedures are specified in Section 2.1.1 of this document.

Compared to other C-Multicast routes specified in [RFC6514], these are "untargeted" in that the RT allows all PEs in the same MVPN to

import them, while those other C-Multicast routes use a RT that identifies a VRF on a particular Upstream Multicast Hop (UMH) PE.

If a PE wants to use selective tunnel to send traffic to only a subset of the PEs on MVPN-RPL, i.e., those with downstream (C-*,C-G-BIDIR) state, per [RFC6513] [RFC6514] the PE needs to advertise a corresponding (C-*,C-G-BIDIR) S-PMSI A-D route, whose PTA specifies the tunnel to be used. In case of RSVP-TE P2MP, Ingress Replication (IR), or BIER tunnel, the Leaf Information Required (LIR) bit in the S-PMSI route's PTA is set to solicit corresponding Leaf A-D routes from those PEs with downstream (C-*,C-G-BIDIR) state. Every PE that wants to use selective tunnel for the (C-*,C-G-BIDIR) will advertise its own S-PMSI A-D route, each triggering a set of corresponding Leaf A-D routes.

Notice that the (C-*,C-G-BIDIR) C-Multicast routes from different PEs all have their own RDs so Route Reflectors (RRs) will reflect every one of them, and they already serve explicit tracking purpose (the BGP Next Hop identifies the originator of the route in non-segmentation case) - there is no need to use Leaf A-D routes triggered by the LIR bit in S-PMSI A-D routes. In case of RSVP-TE P2MP tunnel, the S-PMSI A-D routes are still needed to announce the tunnel but the LIR bit does not need to be set. In case of IR/BIER, there is no need for S-PMSI A-D routes at all.

1.2.2. Optional use of MVPN-RPL RD with mLDP/PIM Provider Tunnels

When mLDP/PIM tunnels are used, there is no need for explicit tracking as the leaves will simply send mLDP label Mapping or PIM Join messages. As a result, it's unnecessary for a PE to retain each C-Multicast route from each PE for the same C-G-BIDIR. If there is a Route Reflector (RR) in use, and it is known apriori that all the PEs/RRs/ASBRs involved in the propagation of the C-Multicast routes support BGP ADD-PATH [I-D.ietf-idr-add-paths], then the PEs could use a common RD to construct the C-Multicast routes. That way, the routes from different PEs for the same C-G-BIDIR will be considered paths for the same route and the RRs will reflect N paths to each PE. If N is significantly smaller than the number of PEs that advertises the routes, then the burden is significantly reduced for the PEs.

The reason for the need for ADD-PATH is shown with this example: both PE1 and PE2 advertise the same (C-*,C-G-BIDIR) C-Multicast route and the RR chooses the one from PE1 as the active path. Without ADD-PATH, the RR won't reflect any (C-*,C-G-BIDIR) path back to PE1, causing PE1 to think there is no other PE interested in receiving the C-G-BIDIR traffic. With ADD-PATH, it is guaranteed that even the originator of the active path will receive at least one other path. For this reason, ADD-PATH is needed and N=2 is well enough.

1.2.3. MVPN C-ASM Support without CE Routers

Current MVPN specifications is based on the fact that CEs are routers and in case of ASM one or more of the routers in customer address space, which could be a CE, a PE's VRF, or another non-PE/CE router, serves as RP. Traffic may be delivered on shared trees, switch to source specific trees, or switch back to shared trees depending the situation. There are two modes of MVPN to support ASM, all involving (C-S,C-G) MVPN Source Active (SA) A-D routes, individual (C-S,C-G) control/forwarding plane state and procedures that are not needed for a special scenario where CEs are not routers but just hosts.

From a logical point of view, this special scenario is when a VPN only involves customer networks directly connected to the PEs and no customer routers are used.. A practical example is EVPN inter-subnet multicast [I-D.lin-bess-evpn-irb-mcast], when EVPN is used to connect only servers and no customer routers are involved. In this case, it does not make sense to introduce the RP concept into the deployment and involve the MVPN SA procedures. Rather, this could be modeled as C-Bidir with MVPN-RPL and all the above discussed optimizations apply.

1.3. Inter-AS Propagation of MVPN C-Multicast Routes

Section 11.2 of [RFC6514] specifies the procedure used to propagate C-multicast routes from one AS to another. However, there are a number of problems with the procedures as specified in that RFC.

RFC6514 presumes that C-multicast routes are propagated through the ASBRs. This is analogous to RFC 4364's "Inter-AS option b". However, in some deployment scenarios, the C-multicast routes are propagated through Route Reflectors, in a manner analogous to RFC 4364's "Inter-AS option c". Strictly speaking, RFC 6514 does not allow this deployment scenario. This document updates RFC 6514 by allowing this deployment scenario to be used in place of the procedures of Section 11.2 of RFC 6514.

In some deployment scenarios, the propagation of C-multicast routes is controlled by the "Route Target Constraint" procedures of [RFC4684]. Strictly speaking, RFC 6514 does not allow this deployment scenario. This document updates RFC 6514 by allowing this deployment scenario to be used in place of the procedures of Section 11.2 of RFC 6514.

Per [RFC6514], an MVPN C-Multicast route is targeted at a particular PE, and its inter-as propagation towards the PE follows a series of ASBRs (in the reverse order) on the propagation path of one of the following:

- o The Intra-AS I-PMSI A-D route from the targeted PE, if the deployment is using non-segmented tunnels. In this scenario, the IP address of the targeted PE is encoded into the four-octet "Source AS" field (!) of the C-multicast route's NLRI.
- o The Inter-AS I-PMSI A-D route for the AS that the targeted PE is in, if the deployment is using segmented tunnel. In this scenario, the AS number of the source PE is encoded into the "Source AS" field of the C-multicast route's NLRI.

In both cases, the corresponding I-PMSI A-D route is found by looking for an I-PMSI A-D route whose NLRI consists of the C-multicast route's RD prepended to the contents of the C-multicast route's "Source AS" field. If neither Inter-AS nor Intra-AS I-PMSI A-D route is used, e.g. (C-*,C-*) S-PMSI A-D route is used, then the specified procedure will not work.

It must be noted that the RFC 6514 Section 11.2 propagation procedures cannot be applied to untargeted C-multicast routes, and cannot be applied even to targeted C-multicast routes if the infrastructure is based on IPv6 rather than IPv4.

This document updates RFC 6514 by declaring that the procedure of Section 11.2 of that document is only applicable in the case that (1) the C-multicast routes are being propagated through the ASBRs, AND (2) the propagation of those routes is not under the control of the Route Target Constraint procedures. It also updates the procedures of Section 11.2 of [RFC6514] to allow it to work without relying on I-PMSI A-D routes, whether IPv4 or IPv6 infrastructure is used.

This document also updates RFC 6514 by declaring that C-multicast routes MAY be propagated using ordinary BGP propagation procedures, which do not rely on the presence of I-PMSI A-D routes. For targeted C-multicast routes, this will result in a less optimal propagation path, but it does work in all cases. The Route Target Constraint procedures can always be used to obtain a more optimal path.

The selection of the propagation procedure for C-multicast routes is determined by provisioning.

In Section 1.2.1, the explicit tracking using C-multicast route relies on that the route's next hop is not changed so that the next hop can identify the originator. If the c-multicast routes are propagated through ASBRs, the next hop will be changed. With tunnel segmentation, this is not a problem (see Section 1.5) but if non-segmented tunnels are used, either the C-multicast route propagation must follow the Optoin C procedures and the next hop is not changed by the RRs, or the routes must carry an EC to identify the

originator. Or, the RD of a C-multicast route can be used to locate an I/S-PMSI route from the same PE, in which the Originator IP Address can be found.

1.4. EVPN Selective Multicast Ethernet Tag (SMET) Routes

[I-D.sajassi-bess-evpn-igmp-mld-proxy] defines a new EVPN route type known as an "SMET route".

The EVPN SMET routes are analogous to the MVPN C-multicast routes, in that both type of routes are used to disseminate the information that a particular egress PE has interest in a particular multicast C-flow or set of C-flows.

An EVPN SMET route contains, in its NLRI, the RD associated with the VRF from which the SMET route was originated. In addition, it is disseminated to all PEs of a given EVI. In this way, SMET routes are analogous to the MVPN C-multicast routes that are used for C-BIDIR support.

An EVPN SMET route contains, in its NLRI, the IP address of the originating PE. In this way, they are analogous to the MVPN Leaf A-D routes (They really combine the function of the MVPN C-multicast routes and the MVPN Leaf A-D routes). Similarly, they are also analogous to the C-multicast route for MVPN-RPL that carries an EC that identifies the originating PE.

In EVPN, as in MVPN, explicit tracking is required when selective tunnels are realized using IR, BIER, or RSVP-TE P2MP. The EVPN SMET routes provide this explicit tracking, so in these cases EVPN does not need explicit Leaf A-D routes. With IR/BIER, there is no need for S-PMSI route either. However, when SMET routes are used with segmented IR/BIER tunnels, more procedures are needed, just like the C-multicast route in MVPN-RPL case (Section 1.5). For that reason, given the similarity between SMET and C-Multicast routes, in this document we will use the same term C-Multicast route for EVPN SMET route as well. The two may be used interchangeably in case of EVPN.

If selective tunnels are set up using procedures that do not require explicit tracking, e.g. mLDP or PIM, the following optimization could be done, similar to MVPN-RPL with mLDP/PIM tunnels (Section 1.2.2):

- o When constructing an SMET route, put 0 as the Originator Router Address.
- o When constructing an SMET route in the context of a given EVI, have all PEs of that EVI set the RD field of the NLRI to the same

value (This is analogous to "MVPN-RPL RD" discussed in Section 1.2.2).

- o When a Route Reflector distributes the SMET routes, it uses BGP ADD-PATH to distribute at least two "paths" for a given NLRI.

1.5. Provider Tunnel Segmentation with Explicit-Tracking C-Multicast Routes

For the above MVPN-RPL and EVPN cases where C-multicast routes are used for explicit tracking without requiring corresponding S-PMSI A-D routes in case of IR/BIER selective tunnel, it works well when there is no tunnel segmentation. With tunnel segmentation [RFC6514] [RFC7524], [I-D.zzhang-bess-evpn-bum-procedure-updates] additional procedures are needed.

1.5.1. Conventional Tunnel Segmentation

Multicast forwarding needs to follow a rooted tree. With segmentation, the tree is divided into segments, with each segment rooted at either the ingress PE or a Regional Border Router (RBR). A segment is contained in a region, which could be an AS, an area, or a sub-area. The root of a segment only needs to track the leaves in its region, which are PEs or RBRs in that region. With the traditional PMSI/Leaf A-D procedures, an ingress PE/RBR sends out an I/S-PMSI route, propagated by RBRs (segmentation points), who change the tunnel identifier along the way to identify the tunnels for their segments. The Leaf A-D routes from PEs are not propagated by the RBRs. Rather, a RBR will proxy the Leaf AD routes it receives from its downstream towards its upstream RBR or PE, following the I/S-PMSI A-D routes received in the upstream region, as specified in [RFC6514] [RFC7524] [I-D.zzhang-bess-evpn-bum-procedure-updates].

1.5.2. Selective Tunnel Segmentation with Untargeted Explicit-Tracking C-multicast Routes

Without segmentation, the untargeted explicit-tracking C-Multicast routes are sent to every PE, and each PE adds the originator of the routes as leaves of the tunnel rooted at the PE.

With segmentation, untargeted explicit-tracking C-Multicast routes are propagated through segmentation points towards all ingress PEs or ASes and are merged along the way. This is like the traditional PMSI/Leaf A-D procedures but with one difference.

With the traditional PMSI/Leaf A-D procedures, the propagation is towards the originator of the PMSI A-D route and a single tree is formed. With untargeted C-Multicast routes, multiple trees are

formed, each being rooted at the ingress PE (if per-region aggregation [I-D.zhang-bess-evpn-bum-procedure-updates] is not used) or ingress RBR (if per-region aggregation is used). The roots of those trees are either the ingress PEs or the ingress RBRs, identified by all the per-PE or per-region I-PMSI A-D routes.

To form those multiple trees without requiring S-PMSI A-D routes from the ingress PEs/RBRs, this document proposes that the RBRs convert a C-multicast route originated in its own region to Leaf A-D routes, as if corresponding S-PMSI A-D routes had been received from ingress PEs/RBRs. The details are provided in Section 2.2.

2. Specifications

This section provides detailed specifications for the optional enhancements introduced above.

2.1. MVPN C-Bidir Support with VPN Backbone being RPL

2.1.1. Constructing C-Multicast Share Tree Join route

In the context of a particular VRF, a PE with downstream state for the group C-G-BIDIR originates a C-multicast Shared Tree Join route, referred to as "MVPN-RPL C-multicast Join", when the MVPN-RPL method of C-BIDIR support is being used.

The fields of the route are set as follows:

- o RD: See Section 2.1.1.2.
- o Source AS: set to zero.
- o Multicast Source Length: 4 or 16.
- o Multicast Source: set to RPA.
- o Multicast Group Length: 4 or 16.
- o Multicast Group: BIDIR-PIM group address.

Note that the RD field, and the Route Targets that are attached to the C-multicast route are different than what is specified in [RFC6514]. See following two sections.

2.1.1.1. Setting the Route Targets

Per [RFC6514], when a PE originates a C-multicast route, it "targets" the route to a specific one of the other PEs attached to the same VPN. The IP address of the targeted PE is encoded into a Route Target and attached to the C-multicast route. This ensures that the C-multicast route is processed only by the PE to which it is targeted.

However, C-multicast routes used by the MVPN-RPL method are not targeted. Rather, they must be processed by all the other PEs attached to the same MVPN. Thus we refer to these routes as "untargeted". The Route Targets attached to these routes must be such as to cause the routes to be propagated to all the other PEs of the given MVPN. By default, these will be the same Route Targets that are attached to the I-PMSI A-D routes of the MVPN.

2.1.1.2. Setting the Route Distinguisher

Per [RFC6514], the RD in a C-multicast Join Route is the RD of a VRF on the PE to which the route is targeted. However, in an MVPN-RPL C-multicast Join, the RD is set differently.

If PIM/mLDP provider tunnels are used, and it is known that all the PEs/RRs/ASBRs involved in the propagation of C-multicast routes support BGP ADD-PATH, the RD MAY be set to a value that is specially configured to be used as the RD for MVPN-RPL in a given VPN. Call this the "MVPN-RPL" RD for that VPN. In that case, all the C-multicast Joins that are providing C-BIDIR support (for a given VPN) using the MVPN-RPL method will have the same RD. This MVPN-RPL RD of a given VPN MUST NOT be used for any other purpose, or by any other VPN. See Section 1.2.2 for a discussion of when it may be advantageous to use an MVPN-RPL RD.

For other provider tunnel types, or if the above mentioned MVPN-RPL RD in case of PIM/mLDP tunnel is not feasible (e.g. BGP ADD-PATH is not supported), the RD in the C-multicast route is that of the VRF from which the route is originated.

For Global Table Multicast (GTM) using MVPN procedures [RFC7116], RFC 7116 specifies that MVPN routes use a special 0:0 RD. This document specifies that GTM use non-0:0 RDs for C-Multicast routes for C-Bidir, when the backbone is used as RPL and provider tunnels are not set up by PIM/mLDP.

2.1.2. Setting Up the MVPN-RPL

By default, the I-PMSI or (C-*,C-BIDIR) S-PMSI plays the role of MVPN-RPL. When (C-*,C-G-BIDIR) S-PMSI is used for a particular C-G-BIDIR, the following procedures are followed, depending on the type of provider tunnel used.

2.1.2.1. Ingress Replication or BIER

If Ingress Replication or BIER is used, there is no need for the ingress PE to advertise (C-*,C-G-BIDIR) S-PMSI A-D route. The ingress PE identifies the tunnel leaves to send traffic to by the C-multicast routes it receives, because each such route has a different RD and serves explicit tracking purpose. In case of IR, the label in the Intra-AS I-PMSI A-D route or (C-*,C-*) S-PMSI A-D route from a leaf is used to send traffic to the leaf. In case of BIER, the label in the same route from the ingress PE is used to send traffic.

2.1.2.2. RSVP-TE P2MP

With RSVP-TE P2MP tunnel, the ingress PE advertises (C-*,C-G-BIDIR) S-PMSI A-D route without setting the LIR bit in the route's PTA. It identifies the tunnel leaves from the C-multicast routes it receives.

2.1.2.3. PIM/mLDP

With PIM or mLDP P2MP provider tunnel, procedures in [RFC6514] are followed.

2.2. Inter-AS Propagation of MVPN C-Multicast Routes

This specification allows two methods of Inter-AS propagation for MVPN C-multicast routes. The choice of which method is used is by provisioning.

2.2.1. Procedures in Section 11.2 of [RFC6514]

The procedures in Section 11.2 of [RFC6514] are extended with the following.

The Source AS field in the NLRI of C-multicast route is set to the AS number of the UMH PE if and only if segmented inter-AS tunnels and per-AS aggregation (via Inter-AS I-PMSI A-D routes) are used. The existing procedures are used as is in this case.

Otherwise, when an egress PE constructs a C-Multicast route and the upstream PE is in a different AS from the local PE, it finds in its

VRF an Intra-AS I-PMSI A-D route or any S-PMSI A-D route from the upstream PE (the Originating Router's IP Address field of that route has the same value as the one carried in the VRF Route Import of the (unicast) route to the address carried in the Multicast Source field). The RD of the found I/S-PMSI A-D route is used as the RD of the advertised C-multicast route. The Source AS field in the C-multicast route is set to 0. If the Next Hop of the found I/S-PMSI A-D route is an EBGP neighbor of the local PE, then the PE advertises the C-multicast route to that neighbor. Otherwise the PE advertises the C-multicast route into IBGP.

When an ASBR receives a C-multicast route with the Source AS field set to 0, it uses the RD of the C-multicast route to locate an Intra-AS I-PMSI A-D route or any S-PMSI A-D route, and propagate the C-multicast route to the bgp neighbor from which the found I/S-PMSI A-D route is learned.

2.2.2. Ordinary BGP Propagation Procedures

This document specifies that C-multicast routes MAY be propagated using ordinary BGP propagation procedures, which do not rely on the presence of any I/S-PMSI A-D routes. With this method, the Source AS field in the C-Multicast route SHOULD be set to 0. For targeted C-multicast routes, this will result in a less optimal propagation path, but it does work in all cases. The Route Target Constraint procedures can always be used to obtain a more optimal path.

2.3. Provider Tunnel Segmentation with Explicit-Tracking C-Multicast Routes

This section applies when IR/BIER are used for MVPN/EVPN selective tunnels.

If per-region aggregation [I-D.zhang-bess-evpn-bum-procedure-updates] is used, this document specifies that the per-region I-PMSI A-D route MUST carry a VRF Route Import EC to identify the originator of the per-region I-PMSI A-D route. Note that, while it borrows "VRF Route Import EC" from the UMH routes, it is only used to identify the originator.

If per-region aggregation is not used, this document specifies that either per-PE I-PMSI or (C-*,C-*) S-PMSI A-D routes MUST be originated by every PE.

2.3.1. Egress PEs and RBRs

An egress PE originates MVPN C-multicast routes for MVPN-RPL as specified in previous sections of this document, or EVPN SMET routes as specified in [I-D.sajassi-bess-evpn-igmp-mld-proxy]. Recall that EVPN SMET routes may also be referred to C-Multicast routes in this document.

Explicit-tracking C-multicast routes must be processed by segmentation points, which are referred to as RBRs. When a RBR receives a C-multicast route from within its own region, and the route does not carry a flag bit that indicates the route is converted from a downstream Leaf A-D route (see descriptions below), it converts the C-multicast route into one or more Leaf A-D routes, as if it had received corresponding S-PMSI A-D routes. When a converted Leaf A-D routes reaches the ingress region, the RBR converts it back to C-multicast routes.

With per-region aggregation, the RBR in an egress region finds all active per-region I-PMSI A-D route that the RBR has in the corresponding VRF. For each of them, it makes up a (C-S,C-G) or (C-*,C-G) S-PMSI A-D route as following.

- o RD: set to the RD from the per-region I-PMSI A-D route.
- o Source/Group length/address fields: set according to the received C-multicast route.
- o Originator's IP Address: set according to the VRF Route Import EC in the per-region I-PMSI A-D route
- o Ethernet Tag ID in case of EVPN: set according to the received SMET route (which is also referred to as C-multicast route).
- o Next Hop: set according to the per-region I-PMSI A-D route.

Without per-region aggregation, a RBR finds all active per-PE I-PMSI or (C-*,C-*) S-PMSI A-D route in the VRF. For each of them it makes up a (C-S,C-G) or (C-*,C-G) S-PMSI A-D route similar to the per-region aggregation case. The only difference is that the Originator's IP Address field is set to the same as in the per-PE I-PMSI or (C-*,C-*) S-PMSI A-D route.

The made up S-PMSI A-D route is for local use only, and not propagated anywhere. A corresponding Leaf A-D route is then generated and propagated to the upstream identified by the BGP next hop in the made up S-PMSI A-D route, following existing PMSI/Leaf A-D route procedures.

2.3.2. Transit RBRs

When an upstream RBR receives a (C-S,C-G) or (C-*,C-G) Leaf A-D route, It locates the active per-PE/region I-PMSI or (C-*,C-*) S-PMSI A-D route whose RD matches the received Leaf A-D route. If no such route exists, the received Leaf A-D route is ignored until such a route appears later. It also tries to locate a corresponding active (C-S,C-G) or (C-*,C-G) S-PMSI A-D route, which could be a real one received from an upstream PE/RBR, or could be a made up one triggered by a Leaf A-D route from a different downstream. If such route exists, existing PMSI/Leaf A-D route procedures are followed.

If no such corresponding active (C-S,C-G) or (C-*,C-G) S-PMSI A-D route exists, and the located active I-PMSI or (C-*,C-*) S-PMSI A-D route has a next hop different from the Originator IP Address in the per-PE I-PMSI A-D route or (C-*,C-*) I-PMSI A-D route, or different from the address in the VRF Route Import EC in the per-region I-PMSI A-D route, the ingress region corresponding to the I-PMSI or (C-*,C-*) S-PMSI A-D route has not been reached. The RBR then makes up a (C-S,C-G) or (C-*,C-G) S-PMSI A-D route. as specified earlier, and proxies Leaf A-D routes further up.

2.3.3. Ingress RBRs

If the BGP next hop in the located active I-PMSI or (C-*,C-*) S-PMSI A-D route matches the Originator IP Address in the per-PE I/S-PMSI A-D route or the IP address in the per-region I-PMSI A-D route's VRF Route Import EC, it means the ingress region has been reached. If the corresponding (C-S,C-G) or (C-*,C-G) S-PMSI A-D route is a made up one and not actually advertised by an ingress PE/RBR, the RBR reconverts the Leaf A-D route back to C-multicast route, with a CV ("Converted") flag bit indicating that the route is not from local state learned on PE-CE interface but from state learned further downstream. The flag bit prevents other RBRs in this region to trigger Leaf A-D routes from this converted C-multicast route.

The converted C-multicast route is constructed as following:

- o RD: set to the RD of the RBR for the related IP/MAC VRF.
- o Source/Group length/address fields: set according to the received Leaf A-D route.
- o Ethernet Tag ID in case of EVPN: set according to the received Leaf A-D route.
- o Next Hop: set to the RBR's local IP Address.

The RT of the converted C-multicast route is set to the RT used for VRF but the route is only propagated to PEs/RBRs in the local region.

For EVPN SMET routes, the flag bit is part of the existing Flags field in the NLRI:

```

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

```

The IE/v3/v2/v1 are existing bits and the CV bit is the new bit to indicate that this is converted from state learned from downstream.

For MVPN C-Multicast route, the CV bit is part of a new MVPN Flag EC, to be specified in a future revision.

2.3.4. Setting Up Forwarding State on RBRs

As a RBR follows the PMSI/Leaf A-D route procedures (even though the S-PMSI A-D route may be made up and not real), it sets up forwarding state accordingly [I-D.ietf-bess-ir] [I-D.ietf-bier-mvpn]. If IR is used in the upstream region, a downstream allocated label is advertised in the PTA of the Leaf A-D route sent upstream. If BIER is used in a region, the root RBR for the segment in that region MUST advertise an S-PMSI A-D route, whether the route is actually received from upstream or made up based on received C-multicast route or Leaf A-D route, with the PTA's label field set to a label upstream-allocated by the root RBR of the segment. This allows label switching by the RBR instead of relying on (C-S,C-G) lookup based forwarding in the VRF.

2.3.5. Other Types of Tunnels

The inter-region segmented tunnel can consists of different types of tunnels, like PIM/mLDP/RSVP-TE P2MP tunnels that require advertised S-PMSI A-D routes. This is just like BIER case mentioned in the above section. The only difference is that in BIER case it is the upstream allocated label that needs to be advertised by the S-PMSI A-D routes and in PIM/mLDP/RSVP-TE P2MP case it is the tunnel identity and optionally the upstream allocated label that need to be advertised by the S-PMSI A-D routes.

3. Security Considerations

This document does not seem to introduce new security risks, though this may be revised after further review and scrutiny.

4. Acknowledgements

The authors thank Vinay Nallamothu and Kevin Wang for their comments and suggestions.

5. References

5.1. Normative References

- [I-D.ietf-bess-ir]
Rosen, E., Subramanian, K., and Z. Zhang, "Ingress Replication Tunnels in Multicast VPN", draft-ietf-bess-ir-03 (work in progress), April 2016.
- [I-D.ietf-bier-mvpn]
Rosen, E., Sivakumar, M., Aldrin, S., Dolganow, A., and T. Przygienda, "Multicast VPN Using BIER", draft-ietf-bier-mvpn-03 (work in progress), June 2016.
- [I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-15 (work in progress), May 2016.
- [I-D.sajassi-bess-evpn-igmp-mld-proxy]
Sajassi, A., Patel, K., Thoria, S., Yeung, D., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-sajassi-bess-evpn-igmp-mld-proxy-00 (work in progress), October 2015.
- [I-D.zzhang-bess-evpn-bum-procedure-updates]
Zhang, Z., Lin, W., Rabadan, J., and K. Patel, "Updates on EVPN BUM Procedures", draft-zzhang-bess-evpn-bum-procedure-updates-03 (work in progress), April 2016.
- [RFC2119] 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>>.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszkuk, 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, <<http://www.rfc-editor.org/info/rfc4684>>.

- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <<http://www.rfc-editor.org/info/rfc5015>>.
- [RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <<http://www.rfc-editor.org/info/rfc6513>>.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <<http://www.rfc-editor.org/info/rfc6514>>.
- [RFC7116] Scott, K. and M. Blanchet, "Licklider Transmission Protocol (LTP), Compressed Bundle Header Encoding (CBHE), and Bundle Protocol IANA Registries", RFC 7116, DOI 10.17487/RFC7116, February 2014, <<http://www.rfc-editor.org/info/rfc7116>>.
- [RFC7524] Rekhter, Y., Rosen, E., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area Point-to-Multipoint (P2MP) Segmented Label Switched Paths (LSPs)", RFC 7524, DOI 10.17487/RFC7524, May 2015, <<http://www.rfc-editor.org/info/rfc7524>>.

5.2. Informative References

- [I-D.lin-bess-evpn-irb-mcast]
Lin, W., Zhang, Z., Drake, J., and J. Rabadan, "EVPN Inter-subnet Multicast Forwarding", draft-lin-bess-evpn-irb-mcast-02 (work in progress), March 2016.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

E-Mail: zzhang@juniper.net

Robert Kebler
Juniper Networks

E-Mail: rkebler@juniper.net

Wen Lin
Juniper Networks

EMail: wlin@juniper.net

Eric Rosen
Juniper Networks

EMail: erosen@juniper.net

BESS
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2019

Z. Zhang
Juniper Networks
J. Rabadan, Ed.
Nokia
A. Sajassi
Cisco Systems
March 11, 2019

MVPN/EVPN Composite Tunnel
draft-zzhang-bess-mvpn-evpn-composite-tunnel-00

Abstract

EVPN E-Tree defines a composite tunnel to be used for a Root PE to simultaneously indicate a non-Ingress-Replication tunnel (e.g., P2MP tunnel) in the transmit direction and an Ingress Replication tunnel in the receive direction for BUM traffic. This document extends it to more generic use in both MVPN and general EVPN.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 September 12, 2019.

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. Terminologies	2
2. Introduction	2
2.1. P2MP Tunnels	3
2.2. MP2MP Tunnels	4
3. Specifications	4
3.1. General MVPN/EVPN Use of Composite Tunnels	4
3.2. EVPN-IP Assisted Replication with BIER-IR Composite Tunnel	6
4. Use-cases	6
4.1. BIER-IR Composite Tunnels in EVPN Networks	7
4.2. Assisted Replication and BIER Composite Tunnels	8
5. Security Considerations	10
6. IANA Considerations	10
7. Acknowledgements	11
8. References	11
8.1. Normative References	11
8.2. Informative References	11
Authors' Addresses	12

1. Terminologies

Familiarity with BIER/MVPN/EVPN protocols and procedures is assumed. Some terminologies are listed below for convenience.

[To be added].

2. Introduction

The composite tunnel defined in [RFC8317] is specifically designed for the particular use case of EVPN E-Tree in that the Root PE only needs to receive on the Ingress Replication (IR) tunnel and transmit on the non-IR tunnel encoded in the PMSI Tunnel Attribute (PTA) that specifies a Composite Tunnel, hence the following language quoted from [RFC8317]:

Composite tunnel type is advertised by the Root PE to simultaneously indicate a non-Ingress-Replication tunnel (e.g., P2MP tunnel) in the transmit direction and an Ingress Replication tunnel in the receive direction for the BUM traffic.

However, the underlying principal of MVPN PMSI A-D route, EVPN IMET route and the PTA that the routes carry allows the composite tunnel to be used in more generic use cases for both MVPN and EVPN, as explained in Section 2.1 and Section 2.2.

The EVPN IMET route is the equivalent of MVPN I-PMSI A-D route. In the rest of the document, unless explicitly stated, I-PMSI A-D route refers to MVPN Intra-AS I-PMSI A-D route and/or EVPN IMET route.

2.1. P2MP Tunnels

As specified in [RFC6514] [RFC7432], an I-PMSI A-D route advertises a PE's membership in a VPN or Broadcast Domain (BD). The route carries a PTA, whose Tunnel Identifier field specifies the tunnel that the advertising PE uses to send traffic unless the tunnel type is either "No tunnel information present" or "Ingress Replication". A PE that imports the route into a VRF/BD/EVI will join the specified tunnel if it needs to receive traffic from the advertising PE.

As specified in [RFC6514] and clarified in [RFC7988], if the tunnel type is Ingress Replication, and the Leaf Information Required (LIR) bit in the PTA's Flags field is set to 0, the advertise PE is actually not indicating that it uses IR to send traffic, but that it will receive traffic using the label that is part of the Tunnel Identifier field of the PTA. A PTA with tunnel type set to IR and LIR bit set to 1 does indicate that the advertising router will use IR to send traffic. In that case, the label field in the Tunnel Identifier is set to 0 and receiving PEs will need to send a Leaf A-D route to "join" the IR tunnel. The label value in the Tunnel Identifier of the Leaf A-D route's PTA is used when sending traffic to the advertiser of the Leaf A-D route.

While [RFC7988] is MVPN specific, the above IR procedures and clarifications are also applicable to EVPN, as the EVPN IMET route is the equivalent of an I-PMSI A-D route with the LIR flag set to 0.

In summary, w/o considering composite tunnel, when IR is specified in I-PMSI A-D routes w/ the LIR bit NOT set, the tunnel is used to receive traffic (even from PEs not advertising IR in its PMSI A-D routes). The composite tunnel introduced in [RFC8317] combines a transmitting non-IR tunnel and a receiving IR tunnel, but a PE advertising a composite tunnel should be still be able to send to certain PEs using IR.

2.2. MP2MP Tunnels

When the PTA specifies one of the MP2MP tunnels (BIDIR-PIM, mLDP MP2MP, BIER), it means the advertising PE will use the MP2MP tunnel for both sending and receiving. While [RFC8317] specifies composite tunnel only as transmitting non-IR + receiving IR, an MP2MP tunnel can also be part of composite tunnel to receive traffic. The rest of the document focuses on BIER, but it equally applies to mLDP MP2MP or BIDIR-PIM.

3. Specifications

While not previously done so, this document makes it explicit that, an MVPN/EVPN PE1 advertising a non-IR tunnel for sending traffic can also send to another PE2 using IR if that PE2 advertises to receive traffic with IR (whether PE advertises IR standalone or as part of a composite tunnel), as long as it is known that PE2 does not also join the non-IR tunnel on which PE1 is also sending the same data.

3.1. General MVPN/EVPN Use of Composite Tunnels

This document extends the use of composite tunnel to appropriate general MVPN/EVPN scenarios where a PE advertises a composite tunnel in its I-PMSI A-D route to receive traffic on IR tunnel and send traffic on non-IR tunnel.

This document also allows an MP2MP tunnel to be part of a composite tunnel so that the advertising PE can use both the MP2MP tunnel and IR to receive traffic.

For a regular, non-composite tunnel in the PMSI Tunnel Attribute (PTA) of a PMSI/Leaf A-D route, the PTA includes an "MPLS Label" field between the "Tunnel Type" field and the "Tunnel Identifier" field. The label is for "tunnel aggregation" purpose - traffic on the same tunnel could carry different labels for multiplexing purpose (e.g. for different VPNs/BDs). For an IR tunnel, the label is downstream- assigned; for non-IR tunnels, the label is either 0 (no aggregation) or upstream-assigned, or from a Domain-wide Common Block (DCB) [I-D.ietf-bess-mvpn-evpn-aggregation-label].

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

PTA Fields [RF6514]

[RFC8317] specifies that the "Tunnel Identifier" field includes a three-octet label before the actual identifier of the non-IR tunnel, though the text/diagram about the roles of the labels is unclear/confusing. For easier reference this document moves the added label out, so that the "Tunnel Identifier" is the actual identifier of the non-IR tunnel:

Flags (1 octet)
Tunnel Type (1 octet)
Non-IR Tunnel Aggregation Label (3 octets)
Ingress Replication MPLS Label (3 octets)
Non-IR Tunnel Identifier (variable)

PTA Fields for Composite Tunnel [This Document]

An example of composite tunnel is BIER-IR tunnel, where the tunnel type is set to 0x8B, and BIER Tunnel Aggregation Label and BIER Tunnel Identifier are as specified in [I-D.ietf-bier-mvpn].

Section 4.1 gives an example application of using BIER-IR tunnel for BIER capable EVPN PEs to send/receive BUM traffic via BIER, and receive BUM traffic from BIER incapable PEs via IR; BIER incapable PEs send BUM traffic using IR; BIER traffic from BIER capable PEs will have the BIER header popped off by a Penultimate Hop before reaching BIER incapable PEs [I-D.ietf-bier-php]. The same can be used for MVPN as well.

3.2. EVPN-IP Assisted Replication with BIER-IR Composite Tunnel

For the example in Section Section 4.1, instead of having BIER incapable PEs send BUM traffic using IR to every PE, Assisted Replication (AR) [I-D.ietf-bess-evpn-optimized-ir] can be used for a BIER incapable PE to send BUM traffic to a BIER capable Assisted Replication Replicator (AR-R) via IR, who will then relay to other PEs via BIER.

The same concept applies to MVPN as well, though AR for MVPN is via Virtual Hub and Spoke (VHS) [RFC7024], similar to AR for EVPN-MPLS [I-D.keyupate-bess-evpn-virtual-hub] ([I-D.ietf-bess-evpn-optimized-ir] is for EVPN-IP). The procedures for those two cases will be specified in separate documents.

EVPN-IP AR with BIER-IR composite tunnels follows similar procedures as in [I-D.ietf-bess-evpn-optimized-ir], with the following differences:

- o The IMET route from a BIER capable AR-Replicator that has the IR-IP address in the Originating PE field encodes BIER tunnel in the PTA, as specified in [EVPN-BIER].
- o An AR-Leaf originates an IMET route with BIER-IR tunnel with AR-Leaf flag. If it is BIER capable, it both sends and receives BM traffic via BIER. If it is not BIER capable, it sends BM traffic via IR to the AR-Replicator, who will then relay to other PEs using BIER.
- o The AR-R does NOT relay traffic that arrive with BIER encapsulation.
- o Only non-selective mode is supported.

The above rules are illustrated in further details in Section Section 4.2. Notice that, composite-tunnel is used because [I-D.ietf-bess-evpn-optimized-ir] requires falling back to IR when the AR-Replicator is not available.

4. Use-cases

This section describes some Composite Tunnel use-cases. We refer to BIER-IR as the PTA's Tunnel Type with the high-order bit set and BIER type, I.e., Tunnel Type = 0x8B, as per [RFC8317]. In this section, a BIER non-capable PE is assumed to be a PE that does not support BIER tunnel data plane transmission or termination. However, these BIER non-capable PEs support the required control plane extensions to advertise BIER tunnel information in the IMET PTAs.

4.1. BIER-IR Composite Tunnels in EVPN Networks

BIER-IR composite tunnels may be used in a group of PEs attached to the same EVPN tenant network. This would allow some of those PEs to use BIER P-tunnels where other PEs in the same group may use Ingress Replication (IR). While these BIER-IR composite tunnels can be used in a similar use case as described in [RFC8317], they can also be used along with PHP as a way to introduce BIER in EVPN networks where some of the PEs do not support BIER data plane.

Figure 1 illustrates an example of an EVPN BD where the PE1 and PE2 support BIER data plane, but PE3/PE4/PE5 do not. The network could still benefit of BIER if the BFRs connected to the receiver PEs (BFR1 and BFR2), either directly or through a tunnel, pop the BIER header and send the EVPN payload natively to PE3/PE4/PE5, as described in [I-D.ietf-bier-php].

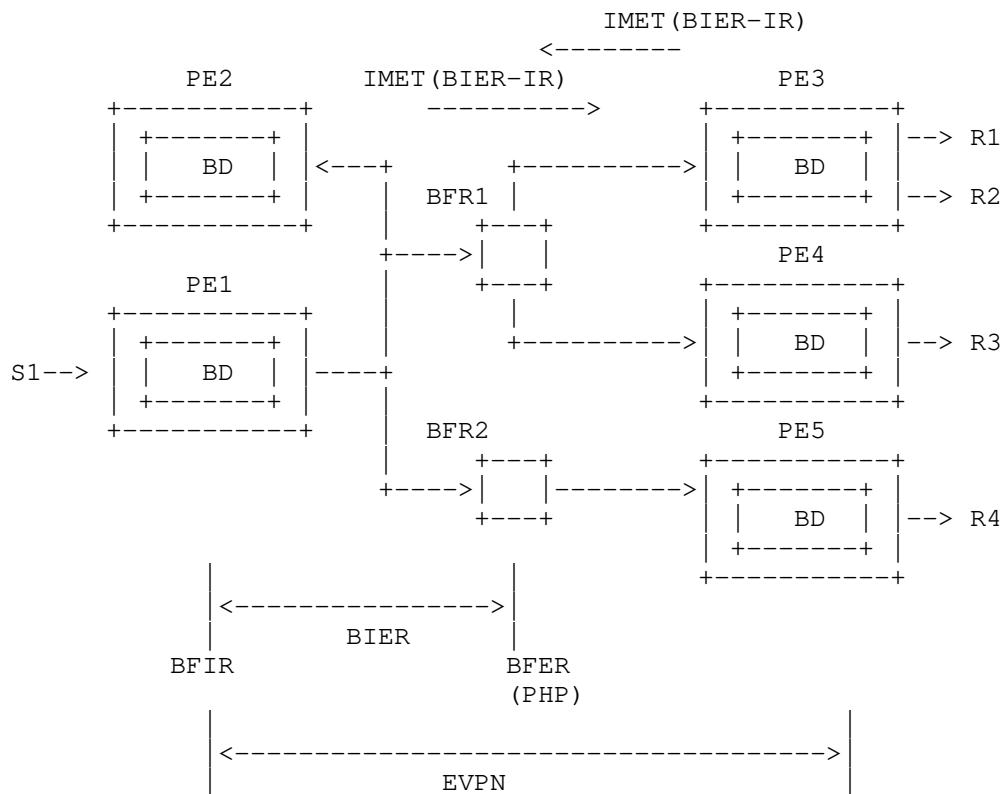


Figure 1 - BIER Composite Tunnels in EVPN

In this example:

- o All the PEs advertise an IMET route containing a BIER-IR composite tunnel in the PTA (PMSI Tunnel Attribute):
 - * The Tunnel Type has a value of 0x8B (BIER-IR composite tunnel).
 - * The BIER Tunnel Identifier (composed Sub-Domain ID, BFR-ID and BFR-Prefix) and Flags are populated as in [EVPN-BIER]. The IR Label is a downstream allocated Label that allows remote PEs to send BUM traffic to the advertising PE using Ingress Replication, as in [RFC8317].
- o When PE1/PE2 need to transmit BUM packets, they follow the procedures in [EVPN-BIER]. BUM packets received on PE1/PE2 from other BIER capable PEs will be received with a BIER encapsulation and procedures in [EVPN-BIER] will be followed. PHP nodes pop the BIER header before delivering the EVPN packets to PE3/PE4/PE5.
- o When PE3/PE4/PE5 need to send BUM packets to each other or to PE1/PE2, they use Ingress Replication and the IR label that is received from the other PEs as part of the composite tunnel Tunnel Identifier.

4.2. Assisted Replication and BIER Composite Tunnels

The use case in section Section 4.1 allows the introduction of BIER in EVPN tenant networks where BIER capable and BIER non-capable PEs are attached to the same EVPN tenant network. However, BIER non-capable PEs still send multiple copies of the same BUM packet to reach the other PEs.

In overlay networks, the use case can be optimized so that the BIER non-capable PEs send a single copy per packet by using Assisted Replication along with BIER-IR composite tunnels. Figure 2 illustrates this use case with an example, where AR-L1/AR-L2/AR-L3 and AR-L4 are Assisted Replication Leaf routers [AR] that do not support BIER data plane. AR-R1/AR-R2 are Non-Selective Assisted Replication Replicators [AR] that do support BIER data plane and are connected to other BFRs, such as BFR1 and BFR2.

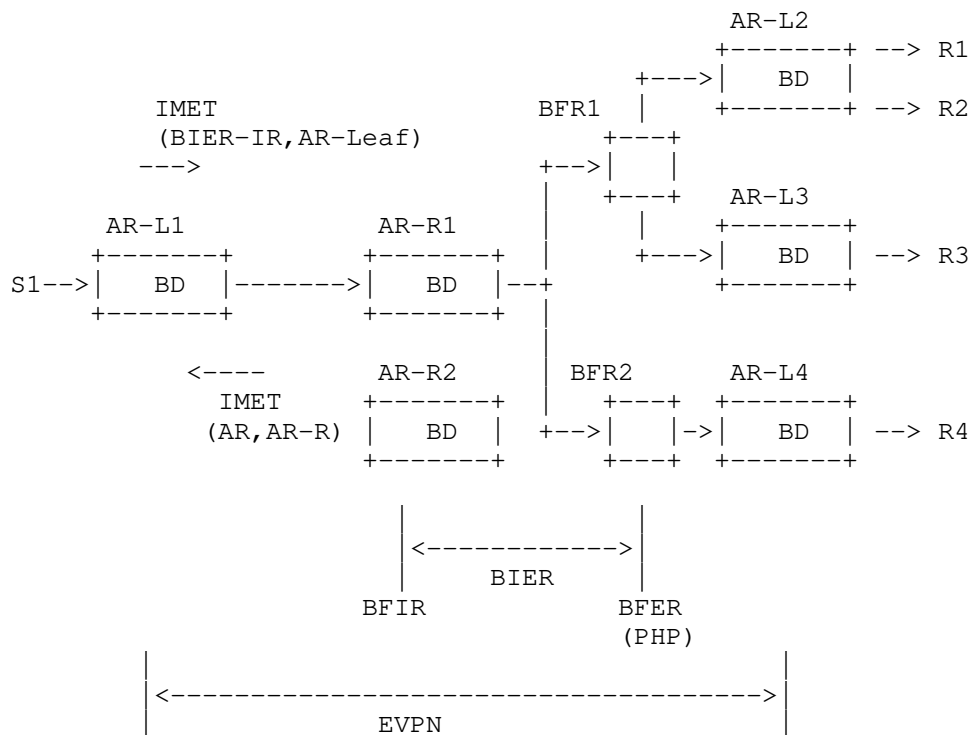


Figure 2 - BIER-IR Composite Tunnels and AR

In this example:

- o The AR-R PE's issue two IMET routes each:
 - * An IMET route that includes the AR-IP in the Originating PE, Tunnel Type AR, IR label, Flags Type = 01 (AR-Replicator) and L = 0 (no Leaf Information Required). The IR Label is a downstream allocated Label that will be used by the AR-L PE's that transmit BUM traffic to the receiving AR-R for replication to remote AR-L PE's. No change with respect to [AR].
 - * And an IMET route that includes the IR-IP in the Originating PE field, Tunnel Type and Tunnel Identifier with BIER information, as in [EVPN-BIER].
- o Each AR-L PE issues an IMET route with:
 - * The Flags field populated as in [AR] with AR Type set to AR-LEAF.

- * The Tunnel Type and Tunnel Identifier have composite BIER-IR information, as in Section Section 4.1. The IR Label is a downstream allocated Label that will be used by the remote AR-L PEs when IR is used for unknown unicast traffic. The MPLS Label field in the PTA MAY be zero.

When an AR-R receives a BM packet encapsulated in an overlay tunnel, it will do a tunnel destination IP lookup and if the destination IP is the AR-R IR-IP Address, the AR-R will proceed as in [AR]. If the destination IP is the AR-R AR-IP Address, the AR-R MUST forward the packet to the BIER network and any local AC (if any). When creating the BIER header, the AR-R will behave as a BFIR and will include all the remote AR-L and AR-R in the BIER header, excluding the AR-L from which the BM packet was received.

If an AR-R receives a BM packet encapsulated in BIER, it will follow the procedures in [EVPN-BIER] as any other BIER PE. It MUST NOT send the BM packets to any overlay tunnels, only to local ACs.

In the example of Figure 2, when AR-R1 receives a BM packet from AR-L1 in an overlay tunnel with its AR-IP as tunnel destination address, it will forward the packet encapsulated with a BIER header that includes AR-L2, AR-L3 and AR-L4 as BFERs, but not AR-L1.

As in [AR], if the AR-L does not discover any AR-R in the service, it MUST use IR to send BM traffic to the remote AR-L PEs and AR-R PEs with local ACs. If there is one or more AR-Rs (discovered by tracking the received AR-R routes) the AR-L selects a AR-R to send the BM traffic to. The selection rules are described in [AR]. The AR-L encapsulates the BM packets into an overlay tunnel that uses the AR-IP and AR Label advertised by the selected AR-R. In the example of Figure 2, AR-L1 selects AR-R1 as the AR-R. If AR-R1 becomes unavailable, AR-R2 is selected. If no AR-R is available, AR-L1 would use IR to send the BM packets to the remote AR-L PEs.

AR-L PEs receive the BUM packets without a BIER header (since it is popped by the PHP node) and with the MPLS Label / VNI imposed by the AR-R (or the source AR-L if there is no AR for the packet).

5. Security Considerations

This specification does not introduce additional security concerns.

6. IANA Considerations

7. Acknowledgements

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>>.
- [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <<https://www.rfc-editor.org/info/rfc8279>>.
- [RFC8317] Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J., Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/RFC8317, January 2018, <<https://www.rfc-editor.org/info/rfc8317>>.

8.2. Informative References

- [I-D.ietf-bess-evpn-optimized-ir]
Rabadan, J., Sathappan, S., Lin, W., Katiyar, M., and A. Sajassi, "Optimized Ingress Replication solution for EVPN", draft-ietf-bess-evpn-optimized-ir-06 (work in progress), October 2018.
- [I-D.ietf-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-ietf-bess-mvpn-evpn-aggregation-label-02 (work in progress), December 2018.
- [I-D.ietf-bier-evpn]
Zhang, Z., Przygienda, T., Sajassi, A., and J. Rabadan, "EVPN BUM Using BIER", draft-ietf-bier-evpn-01 (work in progress), April 2018.
- [I-D.ietf-bier-mvpn]
Rosen, E., Sivakumar, M., Aldrin, S., Dolganow, A., and T. Przygienda, "Multicast VPN Using BIER", draft-ietf-bier-mvpn-11 (work in progress), March 2018.

- [I-D.ietf-bier-php]
Zhang, Z., "BIER Penultimate Hop Popping", draft-ietf-bier-php-01 (work in progress), November 2018.
- [I-D.keyupate-bess-evpn-virtual-hub]
Patel, K., Sajassi, A., Drake, J., Zhang, Z., and W. Henderickx, "Virtual Hub-and-Spoke in BGP EVPNs", draft-keyupate-bess-evpn-virtual-hub-01 (work in progress), October 2018.
- [RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <<https://www.rfc-editor.org/info/rfc6513>>.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <<https://www.rfc-editor.org/info/rfc6514>>.
- [RFC7024] Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", RFC 7024, DOI 10.17487/RFC7024, October 2013, <<https://www.rfc-editor.org/info/rfc7024>>.
- [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>>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Jorge Rabadan (editor)
Nokia

EMail: jorge.rabadan@nokia.com

Ali Sajassi
Cisco Systems

EMail: sajassi@cisco.com

BESS
Internet-Draft
Updates: 6513, 6514, 7524 (if approved)
Intended status: Standards Track
Expires: June 23, 2019

Z. Zhang
Juniper Networks
J. Xie
Huawei
December 20, 2018

MVPN/EVPN Segmentated Forwarding Options
draft-zzhang-bess-mvpn-evpn-segmented-forwarding-00

Abstract

[RFC6513] and [RFC6514] specify MVPN Inter-AS Segmentation procedures. [RFC7524] specifies MVPN Inter-Area Segmentation procedures. [I-D.ietf-bess-evpn-bum-procedure-updates] specifies EVPN BUM Inter-Region Segmentation Procedures. Several other documents also touch upon the segmentation topic. The forwarding at the segmentation points has been assumed to be label switching, subject to certain limitations. The purpose of this document is to provide a review of segmentation points' available forwarding options and limitations, and to clarify and expand some procedures.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 June 23, 2019.

Copyright 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 (<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	2
2. Introduction	3
2.1. MPLS Label Switching at Segmentation Points	3
2.2. IP Processing at Segmentation Points	5
3. Specifications	6
4. References	6
4.1. Normative References	6
4.2. Informative References	7
Authors' Addresses	8

1. Terminology

This document uses terminology from MVPN and EVPN. It is expected that the audience is familiar with the concepts and procedures defined in [RFC6513], [RFC6514], [RFC7524], [RFC7432], [I-D.ietf-bess-evpn-bum-procedure-updates], and [I-D.ietf-bess-evpn-igmp-mld-proxy]. Some terms are listed below for references.

- o PMSI: P-Multicast Service Interface - a conceptual interface for a PE to send customer multicast traffic to all or some PEs in the same VPN. A PMSI A-D route is a BGP MVPN/EVPN auto-discovery route that announces the PMSI and optionally the tunnel that instantiates the PMSI.
- o I-PMSI: Inclusive PMSI - to all PEs in the same VPN.
- o S-PMSI: Selective PMSI - to some of the PEs in the same VPN.
- o Leaf A-D routes: For explicit leaf tracking purpose. Triggered by S-PMSI A-D routes and targeted at triggering route's originator.

- o IMET A-D route: Inclusive Multicast Ethernet Tag A-D route. The EVPN equivalent of MVPN Intra-AS I-PMSI A-D route.
- o SMET A-D route: Selective Multicast Ethernet Tag A-D route. The EVPN equivalent of MVPN Leaf A-D route but unsolicited and untargeted.

2. Introduction

[RFC6513] and [RFC6514] specify MVPN Inter-AS Segmentation procedures. [RFC7524] specifies MVPN Inter-Area Segmentation procedures. [I-D.ietf-bess-evpn-bum-procedure-updates] specifies EVPN BUM Inter-Region Segmentation Procedures. Several other documents also touch upon the segmentation topic.

2.1. MPLS Label Switching at Segmentation Points

It has been assumed that the forwarding across a segmentation point is label based. The upstream segment of a PMSI tunnel is stitched to the downstream segment via label switching and no IP processing is done. This is true even if the segmentation point also has a VRF with PE-CE interfaces, where IP processing is done to decide if a packet should be forwarded out of a PE-CE interface but label switching is used for forwarding traffic to receivers connected by downstream segments.

This label switching is based on the assumption/requirement that each PMSI tunnel has its own unique label (in the simplest case - this can be relaxed as specified in [RFC7988] in case of Ingress Replication). The following is a breakdown of the various situations:

- o If an aggregated RSVP-TE or mLDP P2MP tunnel, or BIER is used for the upstream (or downstream) segment, the x-PMSI A-D route received (or re-advertised, in case of downstream segment) by the segmentation point carries a per-PMSI label in the PMSI Tunnel Attribute (PTA). The BIER case is specified in [I-D.ietf-bier-mvpn] and [I-D.ietf-bier-evpn].
- o If a unique RSVP-TE or mLDP P2MP tunnel is used for for each upstream segment, the segmentation point advertises a unique label for each tunnel to the upstream node on the tunnel. Similarly, in the downstream segment case, the segmentation point must receive a unique tunnel label.
- o If Ingress Replication is used for the upstream segment, the segmentation point may either simply advertise a different label in each Leaf A-D route that it advertises, or use a more elaborate procedure to decide how labels could be advertised while still

allow correct label switching procedure, as specified in Section 7.2 of [RFC7988].

Notice that in the case of P2MP tunnel, x-PMSI A-D routes are required to advertise the tunnel identification and in case of tunnel aggregation (BIER or aggregated P2MP tunnel) the x-PMSI A-D routes are required to advertise the per-PMSI label. However, [I-D.ietf-bess-mvpn-expl-track] introduces a "Leaf Information Required per Flow" bit (LIR-pF) in the flags field of the PTA of wildcard S-PMSI A-D routes, so that an ingress PE does not have to advertise individual more specific S-PMSI A-D routes even if it wants to explicitly track the leaves for more specific flows. This can be used for RSVP-TE P2MP, Ingress Replication and BIER.

For EVPN, explicit tracking is based on unsolicited Selective Multicast Ethernet Tag (SMET) A-D routes and LIR-pF is not used. However, that is as if the LIR-pF flag was set in an implicit (C-*, C-*) wildcard S-PMSI A-D route.

Both [I-D.ietf-bier-mvpn] and [I-D.ietf-bier-evpn] specify that the LIR-pF flag MUST not be used with segmentation. That's because with LIR-pF while an ingress PE can send a flow to only leaves tracked for the flow, it does not advertise the label bound to the corresponding PMSI for the flow (as the LIR-pF removes the need to advertise the more specific S-PMSI routes).

The same restriction also applies if aggregated RSVP-TE P2MP tunnels are used (the same tunnel could be used for multiple more specific S-PMSIs but a per-PMSI label would be associated with each S-PMSI). The LIR-pF flag removes the need for those more specific S-PMSI A-D routes so no S-PMSI specific labels could be advertised for the segmentation points to do label switching with.

The restriction does not apply to Ingress Replication because the per-PMSI label is advertised in the Leaf A-D routes.

The restriction with BIER and aggregated RSVP-TE P2MP tunnel can be lifted if the LIR-pF triggered more specific MVPN Leaf A-D routes or the unsolicited EVPN SMET routes can trigger corresponding S-PMSI A-D routes, so that the per-PMSI labels can be advertised. The concept of triggering S-PMSI A-D routes by Leaf/SMET A-D routes is already present in [RFC7524] and [I-D.zhang-bess-mvpn-evpn-cmcast-enhancements].

It may be argued that triggering S-PMSI A-D route from Leaf/SMET A-D routes for more specific flows has the following concerns (which leads to the consideration for forwarding option described in Section 2.2):

- o Flooding of those extra more specific S-PMSI A-D routes
- o Delay in setting up the forwarding state (as the segmentation points now have to wait for the corresponding S-PMSI A-D route from its upstream).

The first concern can be discounted that the burden of those extra S-PMSI A-D routes are mainly in the control plane. The forwarding plane does need to maintain additional per-PMSI labels but it's much better than the alternative described in the following section.

The second concern can be mitigated by having the ingress PE delay switching traffic over to the more specific S-PMSI. That way, traffic will continue to be forwarded on the less specific PMSI (and label switched by segmentation points) for a short period before being moved to the more specific S-PMSI.

2.2. IP Processing at Segmentation Points

If the above mentioned discount/mitigation are not enough to address the two concerns, IP processing can be used at segmentation points. This will allow the use of LIR-pF with segmentation without triggering those more specific S-PMSI A-D routes [I-D.xie-bier-mvpn-segmented] .

Basically, a segmentation point will create an IP multicast forwarding table for each "context", which could be for an EVPN Broadcast Domain (BD), a L3 VPN, an L3 VPN Extranet, or even something of smaller scope. An incoming packet on an upstream segment is decapsulated and a corresponding IP multicast forwarding table is identified. An IP lookup is performed and forwarded into downstream segments accordingly.

While this does not require the S-PMSI A-D routes triggered by Leaf/SMET routes (and corresponding label forwarding state), additional IP forwarding tables and lookup are needed, which requires additional memory and cycles in the forwarding path, additional code to maintain the RIB/FIB tables, and additional OPEX to monitor them.

Nonetheless, if IP processing on a segmentation point is desired for the reason of LIR-pF bit, the following could be done.

- o Wildcard S-PMSI A-D routes with the LIR-pF flag are assigned with different labels from those in x-PMSI routes w/o the flag, and they lead to IP lookup. The labels can either be upstream assigned or assigned from a Domain-wide Common Block (DCB) [I-D.ietf-bess-mvpn-evpn-aggregation-label].

- o Labels in x-PMSI routes w/o the LIR-pF flag, which are different from those in routes with the flag, lead to label switching.
- o A Leaf A-D route with LIR-pF flag triggers corresponding (C-S, C-G) or (C-*, C-G) routes used for IP lookup, if there is no corresponding S-PMSI A-D route with LIR-pF flag.
- o Upstream PE/ABR uses the label advertised in the matching x-PMSI routes to send traffic (so the packets will either be label switched or ip forwarded by segmentation points).

On a PE, there are already VRFs or BDs configured so the IP RIBs/FIBs are just in those VRFs/BDs. On a segmentation point, most likely there are no VRFs/BDs. How IP RIBs/FIBs are managed is local behavior and implementation dependent. While it is outside the scope of this document, one method could be to maintain one IP RIB/FIB for each label carried in a wildcard S-PMSI A-D route with the LIR-pF flag. .

3. Specifications

Detail specification for the above summary will be added in upcoming revisions.

4. References

4.1. Normative References

[I-D.ietf-bess-evpn-bum-procedure-updates]

Zhang, Z., Lin, W., Rabadan, J., Patel, K., and A. Sajassi, "Updates on EVPN BUM Procedures", draft-ietf-bess-evpn-bum-procedure-updates-05 (work in progress), December 2018.

[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-02 (work in progress), June 2018.

[I-D.ietf-bess-mvpn-expl-track]

Dolganow, A., Kotalwar, J., Rosen, E., and Z. Zhang, "Explicit Tracking with Wild Card Routes in Multicast VPN", draft-ietf-bess-mvpn-expl-track-13 (work in progress), November 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>>.
- [RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <<https://www.rfc-editor.org/info/rfc6513>>.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <<https://www.rfc-editor.org/info/rfc6514>>.
- [RFC7524] Rekhter, Y., Rosen, E., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area Point-to-Multipoint (P2MP) Segmented Label Switched Paths (LSPs)", RFC 7524, DOI 10.17487/RFC7524, May 2015, <<https://www.rfc-editor.org/info/rfc7524>>.

4.2. Informative References

- [I-D.ietf-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-ietf-bess-mvpn-evpn-aggregation-label-02 (work in progress), December 2018.
- [I-D.ietf-bier-evpn]
Zhang, Z., Przygienda, T., Sajassi, A., and J. Rabadan, "EVPN BUM Using BIER", draft-ietf-bier-evpn-01 (work in progress), April 2018.
- [I-D.ietf-bier-mvpn]
Rosen, E., Sivakumar, M., Aldrin, S., Dolganow, A., and T. Przygienda, "Multicast VPN Using BIER", draft-ietf-bier-mvpn-11 (work in progress), March 2018.
- [I-D.xie-bier-mvpn-segmented]
Xie, J., Geng, L., Wang, L., McBride, M., and G. Yan, "Segmented MVPN Using IP Lookup for BIER", draft-xie-bier-mvpn-segmented-06 (work in progress), October 2018.
- [I-D.zzhang-bess-mvpn-evpn-cmcast-enhancements]
Zhang, Z., Kebler, R., Lin, W., and E. Rosen, "MVPN/EVPN C-Multicast Routes Enhancements", draft-zzhang-bess-mvpn-evpn-cmcast-enhancements-00 (work in progress), July 2016.

[RFC7988] Rosen, E., Ed., Subramanian, K., and Z. Zhang, "Ingress Replication Tunnels in Multicast VPN", RFC 7988, DOI 10.17487/RFC7988, October 2016, <<https://www.rfc-editor.org/info/rfc7988>>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Jingrong Xie
Huawei

EMail: xiejingrong@huawei.com