

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2022

LA. Burdet, Ed.
P. Brissette
Cisco
T. Miyasaka
KDDI Corporation
October 25, 2021

EVPN Fast Reroute
draft-burdet-bess-evpn-fast-reroute-00

Abstract

This document summarises EVPN convergence mechanisms and specifies procedures for EVPN networks to achieve sub-second and scale-independant convergence.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 28, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Specification of Requirements	3
3. Terminology	3
4. Requirements	4
5. Solution	5
5.1. Pre-selection of Backup Path	6
5.2. Failure Detection and Traffic Restoration	6
5.2.1. Simultaneous Failures in ES	7
5.2.2. Successive and Cascading Failures in ES	8
6. Redirect Labels: Forwarding Attributes	8
6.1. Bypassing DF-Election Attribute	9
6.2. Terminal Disposition Attribute	9
6.3. Broadcast, Unknown Unicast and Multicast	10
7. Controlled Recovery Sequence	10
8. Transport Underlay	11
9. BGP Extensions	11
10. Security Considerations	11
11. IANA Considerations	12
12. References	12
12.1. Normative References	12
12.2. Informative References	12
Appendix A. Acknowledgments	13
Appendix B. Contributors	13
Authors' Addresses	13

1. Introduction

EVPN convergence and failure recovery methods from different types of network failures is described in [RFC7432] Section 17. Similarly for EVPN-VPWS, [RFC8214] briefly evokes an egress link protection mechanism at the end of Section 5.

The fundamentals of EVPN convergence rely on a mass-withdraw technique of the Ethernet A-D per ES route to unresolve all the associated forwarding paths ([RFC7432] Section 9.2.2 'Route Resolution'). The mass-withdraw grouping approach results in suitable EVPN convergence at lower scale, but is not sufficient to meet stricter sub-second requirements. Other control-plane enhancements such as route-prioritisation ([I-D.ietf-bess-rfc7432bis]) help further but still provide no guarantees.

EVPN convergence using only control-plane approaches is constrained by BGP route propagation delays, routes processing times in software and hardware programming. These are additionally often performed

sequentially and linearly given the potential large scale of EVPN routes present in control plane.

This document presents a mechanism for fast reroute to minimise packet loss in the case of a link failure using EVPN redirect labels (ERLs) with special forwarding attributes. Multiple-failures where loops may occur are addressed, as are cascading failures. A mechanism for distributing redirect labels (ERLs) alongside EVPN service labels (ESLs) is shown.

The main objective is to achieve sub-second convergence in EVPN networks without relying on control plane actions. The procedures in this document apply equally to EVPN services (EVPN [RFC7432], EVPN-VPWS [RFC8214] and EVPN-IRB [RFC9135]), and all Ethernet-Segment load-balancing modes.

2. Specification of Requirements

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

3. Terminology

Some of the terminology in this document is borrowed from [RFC8679] for consistency across fast reroute frameworks.

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

PE: Provider Edge device.

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

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

Egress link: Specific Ethernet link connecting a given PE-CE, which forms part of an Ethernet Segment.

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

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

DF-Election: Designated Forwarder election, as in [RFC7432] and [RFC8584].

DF: Designated Forwarder.

Backup-DF (BDF): Backup-Designated Forwarder.

Non-DF (NDF): Non-Designated Forwarder.

AC: Attachment Circuit.

ERL: Special-use EVPN redirect label, described in this document.

ESL: EVPN service label, as in [RFC7432], [RFC8214] and [RFC9135].

4. Requirements

1. EVPN multihoming is often described as 2 peering PEs. The solution **MUST** be generic enough to apply multiple peering PE and no artificial limit imposed on the number of peering PEs.
2. The solution **MUST** apply to all EVPN load-balancing modes.
3. The solution **MUST** be robust enough to tolerate failures of the same ES at multiple PEs. Simultaneous as well as cascading failures on the same ES must be addressed.
4. The solution **MUST** support EVPN [RFC7432], EVPN-VPWS [RFC8214] and EVPN-IRB [RFC9135] services.
5. The solution **MUST** meet stringent sub-second and often 50 millisecond requirements for traffic loss of EVPN services.
6. The solution **MUST** allow redirected-traffic to bypass port blocking states resulting from DF-Election (BDF or NDF).
7. The solution **MUST** be scale-independant and agnostic of EVPN route types, scale or choice of underlay.
8. The solution **MUST** address egress link (PE-CE link) failures.
9. The solution **MUST** be loop-free, and once-redirected traffic **MUST** never be repeatedly redirected.

10. The solution **MUST** not rely on pushing an additional label onto the label stack.
11. The solution **SHOULD** address Broadcast, unknown unicast and multicast (BUM) traffic.

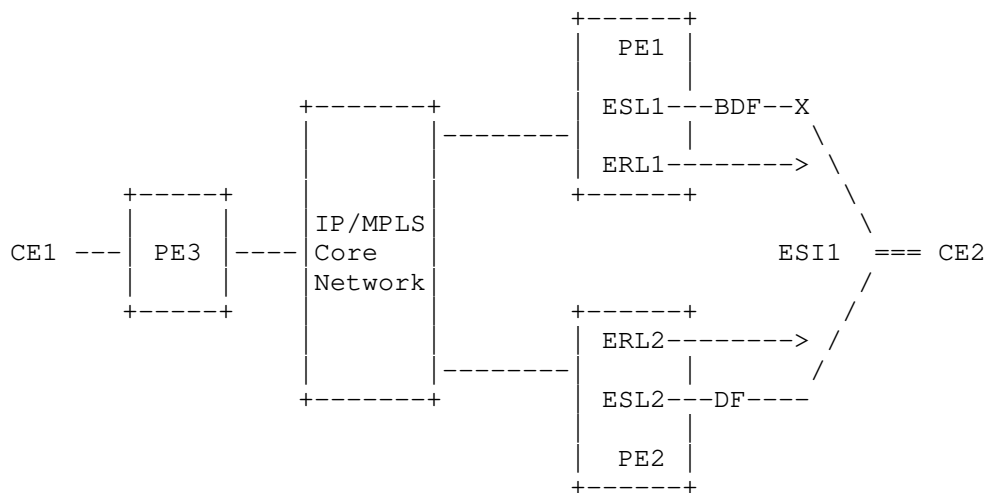
5. Solution

Sub-second convergence in EVPN networks is achieved using a combined approach to minimising traffic loss:

- o Local failure detection and restoration of traffic flows in minimal time using a pre-computed redirect path ;
- o Restoration of optimal traffic paths, and reconvergence of EVPN control plane with EVPN mass withdraw.

The solution presented in this document addresses the local failure detection and restoration, without impeding on or impacting existing EVPN control plane convergence mechanisms.

Consider the following EVPN topology where PE1 and PE2 are multihoming PEs on a shared ES, ESI1. EVPN (known unicast) or EVPN-VPWS traffic from CE1 to CE2 is sent to PE1 and PE2 using EVPN service labels ESL1 and/or ESL2 (depending on load-balancing mode of the ESI1 interfaces).



EVPN Multihoming with service and redirect labels

Figure 1

Alongside the service labels ESL1 and ESL2, two redirect labels ERL1 and ERL2 are allocated with special forwarding attributes, as detailed in Section 6. Fast-reroute and use of the ERLs is shown in Section 5.2

5.1. Pre-selection of Backup Path

EVPN DF-Election lends itself well to the selection of a pre-computed path amongst any given number of peering PEs by providing a DF-Elected and BDF-Elected node at the <EVI, ESI> granularity ([RFC8584] and [I-D.ietf-bess-rfc7432bis]).

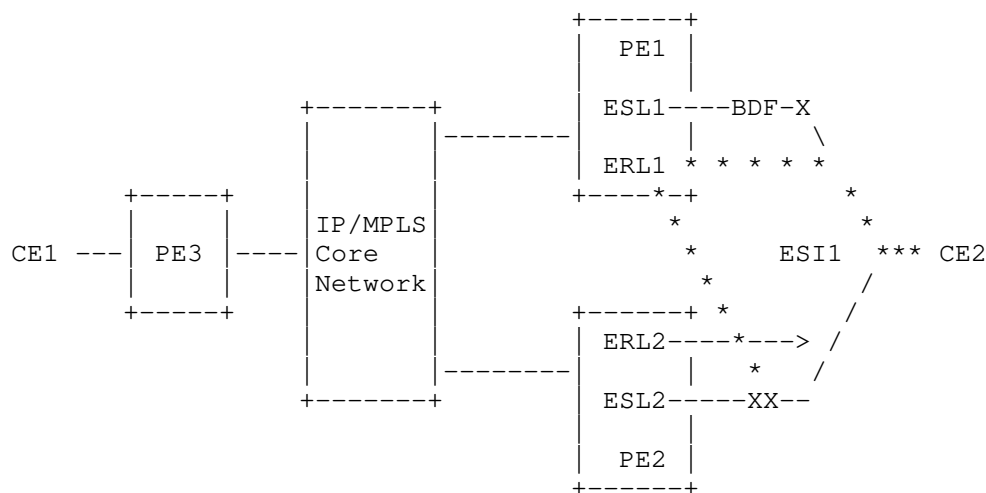
In All-active mode, all PEs in the Ethernet Segment are actively forwarding known unicast traffic to the CE. In Single-active mode, only a single PE in the Ethernet Segment is actively forwarding known unicast traffic to the CE: the DF-Elected PE. The BDF-Elected PE is next to be elected in the redundancy group and is already known.

For consistency across PEs and load-balancing modes, the backup path selected should be in order of {DF, BDF, NDF1, NDF2, ...}. The DF-Elected PE selects the next-best BDF-Elected as backup and all BDF- and NDF-Elected nodes select the best DF-Elected for the protection of their egress links.

- o PE1 (DF) -> ERL(PE2),
- o PE2 (BDF) -> ERL(PE1),
- o PE..n (NDF) -> ERL(PE1),

The number of peering PEs is not limited by existing DF-Election algorithms. A solution based on DF-Election supports subsequent redirection upon multiple cascading failures, once a new DF-Election has occurred. Pre-selection of a backup path is supported by all current DF-Election algorithms, and more generally by all algorithms supporting BDF-Election, as recommended in ([I-D.ietf-bess-rfc7432bis]).

5.2. Failure Detection and Traffic Restoration



EVPN Multihoming failure scenario

Figure 2

The procedures for forwarding known unicast packets received from a remote PE on the local redirect label largely follow [RFC7432] Section 13.2.2.

Consider the EVPN multihoming topology in Figure 1, and a traffic flow from CE1 to CE2 which is currently using EVPN service label ESL2 and forwarded through the core arriving at PE2. When the local AC representing the <EVI,ESI> pair is protected using the fast-reroute solution, the pre-computed backup path's redirect label (i.e. ERL1 from BDF-Elected PE1) is installed against the AC.

Under normal conditions, PE2 disposition using ESL2 will result in forwarding the packet to the CE by selecting the local AC associated with the EVPN service label (EVPN-VPWS) or MAC address lookup (EVPN). When this local AC is in failed state, the fast-reroute solution at PE2 will begin rerouting packets using the BDF-Elected peer's nexthop and ERL1. ERL1 is chosen for redirection and not ESL1 for the redirected traffic to prevent loops and overcome DF-Election timing as described in Sections 6.2 and 6.1 respectively.

5.2.1. Simultaneous Failures in ES

In EVPN multihoming where the CE connects to peering PEs through link aggregation (LAG), a single LAG failure at the CE may manifest as multiple ES failures at all peering PEs simultaneously.

As all peering PEs would enable simultaneously the fast-reroute mechanism, redirection would be permanent causing a traffic storm or until TTL expires.

Once-redirectioned traffic may not be redirectioned again, according to the terminal nature of ERLs described in Section 6.2

5.2.2. Successive and Cascading Failures in ES

Trying to support cascading failures by redirection once-redirectioned traffic is substantially equivalent to simultaneous failures above.

Once-redirectioned traffic may not be redirectioned again, according to the terminal nature of ERLs described in Section 6.2 and loss is to be expected until EVPN control plane reconverges for double-failure scenarios.

In a scenario with 3 peering PEs (PE1-DF, PE2-BDF, PE3-NDF) where PE1 fails, followed by a PE2 failure before control-plane reconvergence, there is no reroute of traffic towards PE3 because the reroute-label is terminal.

In such rapid-succession failures, it is expected that control plane must first correct for the initial failure and DF-Elect PE2 as new-DF and PE3 as the new-BDF. PE2 to PE3 redirection would then begin, unless control-plane is rapid enough to correct directly, and elect PE3 new-DF.

6. Redirect Labels: Forwarding Attributes

The EVPN redirect labels MUST be downstream assigned, and it is directly associated with the <EVI,ESI> AC being egress protected. The special forwarding characteristics and use of an EVPN redirect label (ERL) described below, are a matter of local significance only to the advertising PE (which is also the disposition PE).

Special-attributes to the ERLs do not affect any other PEs or transit P nodes. There are no extra labels appended to the label stack in the IP/MPLS network and the ERL appears to label-switching transit nodes as would any other EVPN service label.

- o Traffic redirection and use of reroute labels may create routing loops upon multiple failures. Such loops are detrimental to the network and may cause congestion between protected PEs.
- o Local restoration and redirection is meant to occur much faster than control-plane operations, meaning redirectioned packets may

arrive at the BDF PE long before a DF-Election operation unblocks the egress link.

Two special forwarding characteristics of EVPN redirect labels are described below to mitigate these issues.

6.1. Bypassing DF-Election Attribute

Local detection and restoration at PE2 will begin rapidly redirecting traffic onto the backup path.

Redirected packets will arrive at the Backup-DF port much faster than control plane DF-Election at the Backup-DF peer is capable of unblocking its local egress link for the shared ES (ESI1). All redirected traffic would drop at Backup-DF and no net reduction in traffic loss achieved.

Traffic restoration remains dependant upon ES route or Ethernet A-D per ES routes withdrawal for a DF-Election operation and for PE1 to assume the traffic forwarding role. This is especially important in single-active load-balancing mode where known unicast traffic is blocked.

To mitigate this, the redirect labels allocated must carry a special attribute in the local forwarding and decapsulation chain: for traffic received on the ERL when the AC is up, an override to the DF-Election is applied and traffic from the ERL will bypass the local Backup-DF blocking state. Once EVPN control plane reconverges, traffic from the ERL will cease and the optimal forwarding path based on ESLs will resume.

The EVPN redirect label MUST carry a context locally, such that from disposition to egress redirected packets are allowed to bypass the BDF blocking state that would otherwise drop. Similarly, this may open the gate to the traffic in the reverse direction.

6.2. Terminal Disposition Attribute

The reroute scheme is susceptible to loops and persistant redirects between peering PEs which have setup FRR redirection. Consider the scenario where both CE-facing interfaces fail simultaneously, fast reroute will be activated at both PE1 and PE2 effectively bouncing a redirected packet between the two PEs indefinitely (or until the TTL expires) causing a traffic storm.

To prevent this, a distinction is made between 'regular' EVPN service labels for disposition (i.e. known unicast EVI label or EVPN-VPWS label) and reroute labels with terminal disposition.

At the redirecting PE2, we consider the case of ESL2 vs. ERL2 , where both are locally allocated and provided in EVPN routes (downstream allocation) to BGP peers:

1. EVPN Service label, ESL2:

- * Regular MAC-lookup or traffic forwarding occurs towards the access AC.
- * If the AC is up, traffic will exit the interface, subject to local blocking state on the AC from DF-Election.
- * If the AC is down and fast-reroute procedures are enabled, traffic may be re-encapsulated using BDF peer's redirect label ERL1 (if received).

2. EVPN Reroute label, ERL2:

- * Regular MAC-lookup or traffic forwarding occurs towards the access AC.
- * If the AC is up, traffic will apply an override to DF-Election and bypass the local blocking state on the AC.
- * If the AC is down, traffic is dropped. No reroute must occur of once-rerouted traffic. Redirecting towards peer's redirect label ERL1 is explicitly prevented.

The ERL acts like a local cross-connect by providing a direct channel from disposition to the AC. ERLs are terminal-disposition and prevents once-redirection packets from being redirected again. With this forwarding attribute on ERLs, known only locally to the downstream-allocating PE, redirection is achieved without growing the label stack with another special purpose label.

6.3. Broadcast, Unknown Unicast and Multicast

BUM traffic is treated using EVPN defaults. There is no further extension to exiting procedure as of now, this work is left for future study.

7. Controlled Recovery Sequence

Fast reroute mechanisms such as the one described in this document generally provide a way to preserve traffic flows at failure time. Use of fast reroute in EVPN, however, permits setting up a controlled recovery sequence to shorten the period of loss between an interface

coming up and the EVPN DF-Election procedures and default timers for peer discovery.

The benefit of a controlled recovery sequence is amplified when used in conjunction with [I-D.ietf-bess-evpn-fast-df-recovery] (synchronised DF-Election)>

8. Transport Underlay

The solution is agnostic to transport underlays, for instance similar behaviour is carried forward for VXLAN and SRv6

9. BGP Extensions

There are no new BGP extensions required to advertise the redirect label(s) used for EVPN egress link protection. The ESI Label Extended Community defined in [RFC7432] Section 7.5 may be advertised along with Ethernet A-D routes:

- o When advertised with an Ethernet A-D per ES route, it enables split-horizon procedures for multihomed sites as described in [RFC7432] Section 8.3 ;
- o When advertised with an Ethernet A-D per EVI route, it enables link protection and fast-reroute procedures for multihomed sites as described in this document. The label value represents the per-<EVI,ESI> EVPN redirect label (ERL). The Flags field SHOULD NOT be set and MUST be ignored.

Remote PEs SHALL NOT use the ERLs as a substitution for ESLs in route resolution, and is especially not to be confused with the aliasing and backup path ESL as described and used in [RFC7432] Section 8.4.

10. Security Considerations

The mechanisms in this document use the EVPN control plane as defined in [RFC7432] and [RFC8214], and the security considerations described therein are equally applicable. Reroute labels redistributed in EVPN control plane are meant for consumption by the peering PE in a same ES. It is, however, visible in the EVPN control plane to remote peers. Care shall be taken when installing reroute labels, since their use may result in bypassing DF-Election procedures and lead to duplicate traffic at CEs if incorrectly installed.

11. IANA Considerations

This document makes no specific requests to IANA.

12. References

12.1. 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>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <<https://www.rfc-editor.org/info/rfc8214>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

12.2. Informative References

- [I-D.ietf-bess-evpn-fast-df-recovery] Brissette, P., Sajassi, A., Burdet, L., Drake, J., and J. Rabadan, "Fast Recovery for EVPN DF Election", draft-ietf-bess-evpn-fast-df-recovery-02 (work in progress), July 2021.
- [I-D.ietf-bess-rfc7432bis] Sajassi, A., Burdet, L., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", draft-ietf-bess-rfc7432bis-01 (work in progress), July 2021.
- [RFC8679] Shen, Y., Jeganathan, M., Decraene, B., Gredler, H., Michel, C., and H. Chen, "MPLS Egress Protection Framework", RFC 8679, DOI 10.17487/RFC8679, December 2019, <<https://www.rfc-editor.org/info/rfc8679>>.

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <<https://www.rfc-editor.org/info/rfc9135>>.

Appendix A. Acknowledgments

Appendix B. Contributors

In addition to the authors listed on the front page, the following co-authors have also contributed to this document:

Authors' Addresses

Luc Andre Burdet (editor)
Cisco

Email: lburdet@cisco.com

Patrice Brissette
Cisco

Email: pbrisset@cisco.com

Takuya
KDDI Corporation

Email: ta-miyasaka@kddi.com

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2022

A. Sajassi
K. Thiruvenkatasamy
S. Thoria
Cisco
A. Gupta
VMware
L. Jalil
Verizon
October 22, 2021

Seamless Multicast Interoperability between EVPN and MVPN PEs
draft-ietf-bess-evpn-mvpn-seamless-interop-03

Abstract

Ethernet Virtual Private Network (EVPN) solution is becoming pervasive for Network Virtualization Overlay (NVO) services in data center (DC) networks and as the next generation VPN services in service provider (SP) networks.

As service providers transform their networks in their Central Offices (COs) towards the next generation data center with Software Defined Networking (SDN) based fabric and Network Function Virtualization (NFV), they want to be able to maintain their offered services including Multicast VPN (MVPN) service between their existing network and their new Service Provider Data Center (SPDC) network seamlessly without the use of gateway devices. They want to have such seamless interoperability between their new SPDCs and their existing networks for a) reducing cost, b) having optimum forwarding, and c) reducing provisioning. This document describes a unified solution based on RFCs 6513 & 6514 for seamless interoperability of Multicast VPN between EVPN and MVPN PEs. Furthermore, it describes how the proposed solution can be used as a routed multicast solution in data centers with only EVPN PEs.

Status of This Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 25, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Requirements Language	5
3. Terminology	5
4. Requirements	7
4.1. Optimum Forwarding	7
4.2. Optimum Replication	7
4.3. All-Active and Single-Active Multi-Homing	8
4.4. Inter-AS Tree Stitching	8
4.5. EVPN Service Interfaces	8
4.6. Distributed Anycast Gateway	8
4.7. Selective & Aggregate Selective Tunnels	8
4.8. Tenants' (S,G) or (*,G) states	9
4.9. Zero Disruption upon BD/Subnet Addition	9
4.10. No Changes to Existing EVPN Service Interface Models	9
4.11. External source and receivers	9
4.12. Tenant RP placement	9
5. Solution Overview	9
5.1. IRB Unicast versus IRB Multicast	10
5.1.1. IRB multicast in seamless interop mode	10
5.2. Operational Model for EVPN IRB PEs	11
5.3. Unicast Route Advertisements for IP multicast Source	13
5.4. Multi-homing of IP Multicast Source and Receivers	15
5.4.1. Single-Active Multi-Homing	15
5.4.2. All-Active Multi-Homing	16
5.5. Mobility for Tenant's Sources and Receivers	18
6. Control Plane Operation	18

6.1.	Intra-ES Subnet Tunnel	18
6.2.	Intra-Subnet BUM Tunnel	20
6.3.	Inter-Subnet IP Multicast Tunnel	20
6.4.	IGMP Hosts as TSes	21
6.5.	PIM Routers as TSes	21
7.	Data Plane Operation	22
7.1.	Intra-Subnet L2 Switching	23
7.2.	Inter-Subnet L3 Routing	23
8.	DCs with only EVPN PE	23
8.1.	Setup of overlay multicast delivery	24
8.2.	Handling of different encapsulations	25
8.2.1.	MPLS Encapsulation	26
8.2.2.	VxLAN Encapsulation	26
8.2.3.	Other Encapsulation	26
9.	DCI with MPLS in WAN and VxLAN in DCs	26
9.1.	Control plane inter-connect	27
9.2.	Data plane inter-connect	28
10.	Interop with L2 EVPN PE	28
10.1.	Interaction with L2EVPN PE and Seamless interop capable PE	29
10.2.	Network having L2EVPN PE, Seamless interop capable PE and MVPN PE	31
11.	Connecting external Multicast networks or PIM routers.	31
12.	TS RP options	32
13.	IANA Considerations	32
14.	Security Considerations	32
15.	Acknowledgements	32
16.	References	33
16.1.	Normative References	33
16.2.	Informative References	34
Appendix A.	Supporting application with TTL value 1	34
A.1.	Policy based model	34
A.2.	Exercising BUM procedure for VLAN/BD	35
A.3.	Intra-subnet bridging	35
Authors' Addresses	36

1. Introduction

Ethernet Virtual Private Network (EVPN) solution is becoming pervasive for Network Virtualization Overlay (NVO) services in data center (DC) networks and as the next generation VPN services in service provider (SP) networks.

As service providers transform their networks in their Central Offices (COs) towards the next generation data center with Software Defined Networking (SDN) based fabric and Network Function Virtualization (NFV), they want to be able to maintain their offered services including Multicast VPN (MVPN) service between their

existing network and their new SPDC network seamlessly without the use of gateway devices. There are several reasons for having such seamless interoperability between their new DCs and their existing networks:

- Lower Cost: gateway devices need to have very high scalability to handle VPN services for their DCs and as such need to handle large number of VPN instances (in tens or hundreds of thousands) and very large number of routes (e.g., in tens of millions). For the same speed and feed, these high scale gateway boxes are relatively much more expensive than the edge devices (e.g., PEs and TORs) that support much lower number of routes and VPN instances.
- Optimum Forwarding: in a given Central Office(CO), both EVPN PEs and MVPN PEs can be connected to the same fabric/network (e.g., same IGP domain). In such scenarios, the service providers want to have optimum forwarding among these PE devices without the use of gateway devices. Because if gateway devices are used, then the IP multicast traffic between an EVPN and MVPN PEs can no longer be optimum and in some case, it may even get tromboned. Furthermore, when an SPDC network spans across multiple LATA (multiple geographic areas) and gateways are used between EVPN and MVPN PEs, then with respect to IP multicast traffic, only one GW can be designated forwarder (DF) between EVPN and MVPN PEs. Such scenarios not only result in non-optimum forwarding but also it can result in tromboning of IP multicast traffic between the two LATAs when both source and destination PEs are in the same LATA and the DF gateway is elected to be in a different LATA.
- Less Provisioning: If gateways are used, then the operator need to configure per-tenant info on the gateways. In other words, for each tenant that is configured, one (or maybe two) additional touch points are needed.

In datacenter deployments, inter-subnet multicast traffic within an EVPN based fabric/data center is unoptimized. When there are multiple receivers in different bridge domains of the same tenant system, a router attached to an EVPN PE would send multiple copies into the EVPN fabric resulting bandwidth wastage. [RFC9135] only covers procedures for efficient inter-subnet connectivity among these Tenant Systems and End Devices while maintaining the multi-homing capabilities of EVPN only for unicast traffic. There is a need to support efficient inter-subnet multicast forwarding within the data center.

This document describes a unified solution based on [RFC6513] and [RFC6514] for seamless interoperability of multicast VPN between EVPN and MVPN PEs. Furthermore, it describes how the proposed solution

can be used as a routed multicast solution in data centers with only EVPN PEs (e.g., routed multicast VPN only among EVPN PEs) to do optimized multicast forwarding.

The document is organized such that seamless interop mode covered first followed by how the same model can be used as an optimized multicast forwarding solution for data center networks.

Section 5 provides the solution overview in detail. This section assumes that all EVPN PEs have IRB capability and operating in IRB mode for both unicast and multicast traffic (e.g., all EVPN PEs are homogenous in terms of their capabilities and operational modes). Section 6 and 7 covers control plane and data plane respectively.

Section 8 describes how the proposed solution can be used to achieve optimized multicast forwarding within the EVPN domain/Data center only networks. Section 9 discusses DCI usecases.

An EVPN network can consist of a mix of L2 and L3 PEs. The multicast operation of such heterogeneous EVPN network will be an extension of an EVPN homogenous network. Section 10 discusses the multicast IRB solution description for the EVPN heterogeneous network.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without any normative meaning.

3. Terminology

Most of the terminology used in this document comes from [RFC8365]

Broadcast Domain (BD): In a bridged network, the broadcast domain corresponds to a Virtual LAN (VLAN), where a VLAN is typically represented by a single VLAN ID (VID) but can be represented by several VIDs where Shared VLAN Learning (SVL) is used per [802.1Q].

Bridge Table (BT): An instantiation of a broadcast domain on a MAC-VRF.

VXLAN: Virtual Extensible LAN

PoD: Point of Delivery

NV: Network Virtualization

NVO: Network Virtualization Overlay

NVE: Network Virtualization Endpoint

NVGRE: Network Virtualization using Generic Routing Encapsulation

GENEVE: Generic Network Virtualization Encapsulation

VNI: Virtual Network Identifier (for VXLAN)

EVPN: Ethernet VPN

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

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

IP-VRF: A Virtual Routing and Forwarding table for Internet Protocol (IP) addresses on a PE

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

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

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

PE: Provider Edge device.

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

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

PIM-SM: Protocol Independent Multicast - Sparse-Mode

PIM-SSM: Protocol Independent Multicast - Source Specific Multicast

Bidir PIM: Bidirectional PIM

FHR: First Hop Router

LHR: Last Hop Router

CO: Central Office of a service provider

SPDC: Service Provider Data Center

LATA: Local Access and Transport Area

Border Leafs: A set of EVPN-PE acting as exit point for EVPN fabric.

EC: BGP Extended Community

UMH: Upstream Multicast Hop

TS: Tenant Systems

4. Requirements

This section describes the requirements specific in providing seamless multicast VPN service between MVPN and EVPN capable networks.

4.1. Optimum Forwarding

The solution SHALL support optimum multicast forwarding between EVPN and MVPN PEs within a network. The network can be confined to a CO or it can span across multiple LATAs. The solution SHALL support optimum multicast forwarding with both ingress replication tunnels and P2MP tunnels.

4.2. Optimum Replication

For EVPN PEs with IRB capability, the solution SHALL use only a single multicast tunnel among EVPN and MVPN PEs for IP multicast traffic, when both PEs use the same tunnel type. Multicast tunnels can be either ingress replication tunnels or P2MP tunnels. The solution MUST support optimum replication for both Intra-subnet and Inter-subnet IP multicast traffic:

- Non-IP traffic SHALL be forwarded per EVPN baseline [RFC7432] or [RFC8365]
- If a Multicast VPN spans across both Intra and Inter subnets, then for Ingress replication regardless of whether the traffic is Intra or

Inter subnet, only a single copy of IP multicast traffic SHALL be sent from the source PE to the destination PE.

- If a Multicast VPN spans across both Intra and Inter subnets, then for P2MP tunnels regardless of whether the traffic is Intra or Inter subnet, only a single copy of multicast data SHALL be transmitted by the source PE. Source PE can be either EVPN or MVPN PE and receiving PEs can be a mix of EVPN and MVPN PEs - i.e., a multicast VPN can be spread across both EVPN and MVPN PEs.

4.3. All-Active and Single-Active Multi-Homing

The solution MUST support multi-homing of source devices and receivers that are sitting in the same subnet (e.g., VLAN) and are multi-homed to EVPN PEs. The solution SHALL allow for both Single-Active and All-Active multi-homing.

4.4. Inter-AS Tree Stitching

The solution SHALL support multicast tree stitching when the tree spans across multiple Autonomous Systems.

4.5. EVPN Service Interfaces

The solution MUST support all EVPN service interfaces listed in section 6 of [RFC7432]:

- o VLAN-based service interface
- o VLAN-bundle service interface
- o VLAN-aware bundle service interface.

4.6. Distributed Anycast Gateway

The solution SHALL support distributed anycast gateways for tenant workloads on NVE devices operating in EVPN-IRB mode.

4.7. Selective & Aggregate Selective Tunnels

The solution SHALL support selective and aggregate selective P-tunnels as well as inclusive and aggregate inclusive P-tunnels. When selective tunnels are used, multicast traffic SHOULD only be forwarded to the remote PEs that have receivers - i.e., if there are no receivers at a remote PE, the multicast traffic SHOULD NOT be forwarded to that PE. If there are no receivers on any remote PEs, then the multicast traffic SHOULD NOT be forwarded to the core.

4.8. Tenants' (S,G) or (*,G) states

The solution SHOULD store (C-S,C-G) and (C-*,C-G) states only on PE devices that have interest in such states hence reducing memory and processing requirements - i.e., PE devices that have sources and/or receivers interested in such multicast groups.

4.9. Zero Disruption upon BD/Subnet Addition

In DC environments, various Bridge Domains are provisioned and removed on regular basis due to host mobility, policy and tenant changes. Such change in BD configuration should not affect existing flows within the same BD or any other BD in the network.

4.10. No Changes to Existing EVPN Service Interface Models

VLAN-aware bundle service as defined in [RFC7432] typically does not require any VLAN ID translation from one tenant site to another - i.e., the same set of VLAN IDs are configured consistently on all tenant segments. In such scenarios, EVPN-IRB multicast service MUST maintain the same mode of operation and SHALL NOT require any VLAN ID translation.

4.11. External source and receivers

The solution SHALL support sources and receivers external to the tenant domain. i.e., multicast source inside the tenant domain can have receiver outside the tenant domain and vice versa.

4.12. Tenant RP placement

The solution SHALL support a tenant to have RP anywhere in the network. RP can be placed inside the EVPN network or MVPN network or external domain.

5. Solution Overview

This section describes a multicast VPN solution based on [RFC6513] and [RFC6514] for EVPN PEs operating in IRB mode that want to perform seamless interoperability with their counterparts MVPN PEs.

In order to enable seamless integration of EVPN and MVPN PEs, traffic originated/received from EVPN PE needs to be modelled very similar to MVPN PE. Hence, there are some differences in handling IRB multicast defined in this document in comparison to IRB unicast defined in [RFC9135]. The next section covers differences.

5.1. IRB Unicast versus IRB Multicast

[RFC9135] describes the operation for EVPN PEs in IRB mode for unicast traffic. The same IRB model used for unicast traffic, where an IP-VRF in an EVPN PE is attached to one or more bridge tables (BTs) via virtual IRB interfaces, is also applicable for multicast traffic.

For unicast traffic, the intra-subnet traffic is bridged within the MAC-VRF associated with that subnet (i.e., a lookup based on MAC-DA is performed); whereas, the inter-subnet traffic is routed in the corresponding IP-VRF (i.e. a lookup based on IP-DA is performed).

A given tenant can have one or more IP-VRFs; however, without loss of generality, this document assumes one IP-VRF per tenant. In context of a given tenant's multicast traffic, the intra-subnet traffic is bridged for non-IP traffic and it is Layer-2 switched for IP traffic. Whereas, the tenant's inter-subnet multicast traffic is always routed in the corresponding IP-VRF. The difference between bridging and L2-switching for multicast traffic is that the former uses MAC-DA lookup for forwarding the multicast traffic; whereas, the latter uses IP-DA lookup for such forwarding where the forwarding states are built in the MAC-VRF using IGMP/MLD or PIM snooping.

5.1.1. IRB multicast in seamless interop mode

EVPN does not provide a Virtual LAN (VLAN) service per [IEEE802.1Q] but rather an emulated VLAN service. This VLAN service emulation is not only done for unicast traffic but also is extended for intra-subnet multicast traffic described in [I-D.ietf-bess-evpn-igmp-ml-d-proxy]. For intra-subnet multicast, an EVPN PE builds multicast forwarding states in its bridge table (BT) based on snooping of IGMP/MLD and/or PIM messages and the forwarding is performed based on destination IP multicast address of the Ethernet frame rather than destination MAC address as noted above.

In order to enable seamless integration of EVPN and MVPN PEs, this document extends the concept of an emulated VLAN service for multicast IRB applications such that the intra-subnet IP multicast traffic can get treated same as inter-subnet IP multicast traffic which means intra-subnet IP multicast traffic destined to remote PEs gets routed instead of being L2-switched - i.e., TTL value gets decremented and the Ethernet header of the L2 frame is de-capsulated and encapsulated at both ingress and egress PEs.

It should be noted that the non-IP multicast or L2 broadcast traffic still gets bridged and frames get forwarded based on their destination MAC addresses.

Link local IP multicast traffic consists IPv4 traffic with a destination address prefix of 224/8 and IPv6 traffic with a destination address prefix of FF02/16. Such IP multicast traffic along with non-IP multicast/broadcast traffic are sent per EVPN [RFC7432] BUM procedures and does not get routed via IP-VRF for multicast addresses. So, such BUM traffic will be limited to a given EVI/VLAN (e.g., a given subnet); whereas, IP multicast traffic, will be locally L2 switched for local interfaces attached on the same subnet and will be routed for local interfaces attached on a different subnet or for forwarding traffic to other EVPN PEs (refer to section 7 for data plane operation).

5.2. Operational Model for EVPN IRB PEs

Without the loss of generality, this section assumes that all EVPN PEs have IRB capability and operating in IRB mode for both unicast and multicast traffic (e.g., all EVPN PEs are homogenous in terms of their capabilities and operational modes). As it will be seen later, an EVPN network can consist of a mix of PEs where some are capable of multicast IRB and some are not and the multicast operation of such heterogeneous EVPN network will be an extension of an EVPN homogenous network. Therefore, we start with the multicast IRB solution description for the EVPN homogenous network.

The EVPN PEs terminate IGMP/MLD messages from tenant host devices or PIM messages from tenant routers on their IRB interfaces, thus avoid sending these messages over MPLS/IP core. A tenant virtual/physical router (e.g., CE) attached to an EVPN PE becomes a multicast routing adjacency of that PE. Furthermore, the PE uses MVPN BGP protocol and procedures per [RFC6513] and [RFC6514]. With respect to multicast routing protocol between tenant's virtual/physical router and the PE that it is attached to, any of the following PIM protocols is supported per [RFC6513]: PIM-SM with Any Source Multicast (ASM) mode, PIM-SM with Source Specific Multicast (SSM) mode, and PIM Bidirectional (BIDIR) mode. Support of PIM-DM (Dense Mode) is excluded in this document per [RFC6513].

The EVPN PEs use MVPN BGP routes defined in [RFC6514] to convey tenant (S,G) or (*,G) states to other MVPN or EVPN PEs and to set up overlay trees (inclusive or selective) for a given MVPN instance. The root or a leaf of such an overlay tree is terminated on an EVPN or MVPN PE. Furthermore, this inclusive or selective overlay tree is terminated on a single IP-VRF of the EVPN or MVPN PE. In case of EVPN PE, these overlay trees never get terminated on MAC-VRFs of that PE.

Overlay trees are instantiated by underlay provider tunnels (P-tunnels) - e.g., P2MP, MP2MP, or unicast tunnels per [RFC6513]. When

there are several overlay trees mapped to a single underlay P-tunnel, the tunnel is referred to as an aggregate tunnel.

Figure-1 below depicts a scenario where a tenant's MVPN spans across both EVPN and MVPN PEs; where all EVPN PEs have multicast IRB capability. An EVPN PE (with multicast IRB capability) can be modeled as a MVPN PE where the virtual IRB interface of an EVPN PE (virtual interface between a BT and IP-VRF) can be considered a routed interface for the MVPN PE.

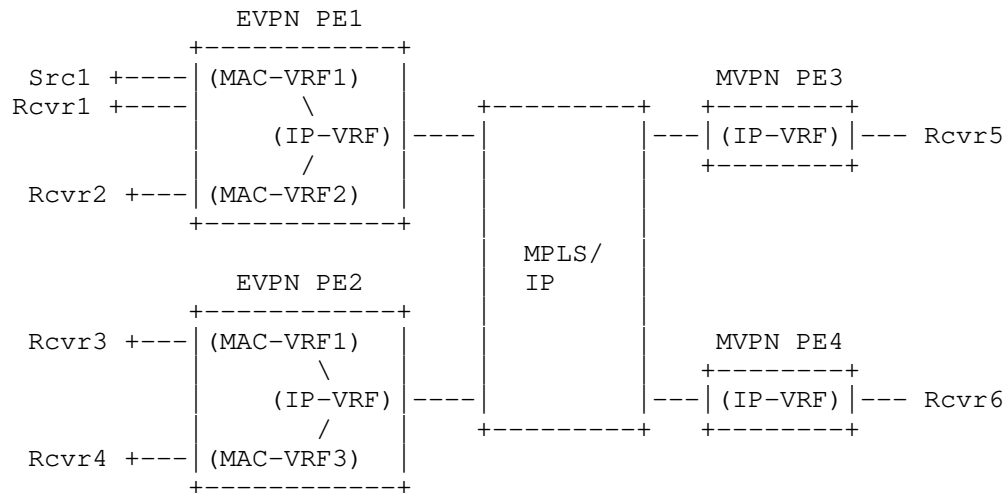


Figure-1: EVPN & MVPN PEs Seamless Interop

Figure 2 depicts the modeling of EVPN PEs based on MVPN PEs where an EVPN PE can be modeled as a PE that consists of a MVPN PE whose routed interfaces (e.g., attachment circuits) are replaced with IRB interfaces connecting each IP-VRF of the MVPN PE to a set of BTs. Similar to a MVPN PE where an attachment circuit serves as a routed multicast interface for an IP-VRF associated with a MVPN instance, an IRB interface serves as a routed multicast interface for the IP-VRF associated with the MVPN instance. Since EVPN PEs run MVPN protocols (e.g., [RFC6513] and [RFC6514]), for all practical purposes, they look just like MVPN PEs to other PE devices. Such modeling of EVPN PEs, transforms the multicast VPN operation of EVPN PEs to that of MVPN and thus simplifies the interoperability between EVPN and MVPN PEs to that of running a single unified solution based on MVPN.

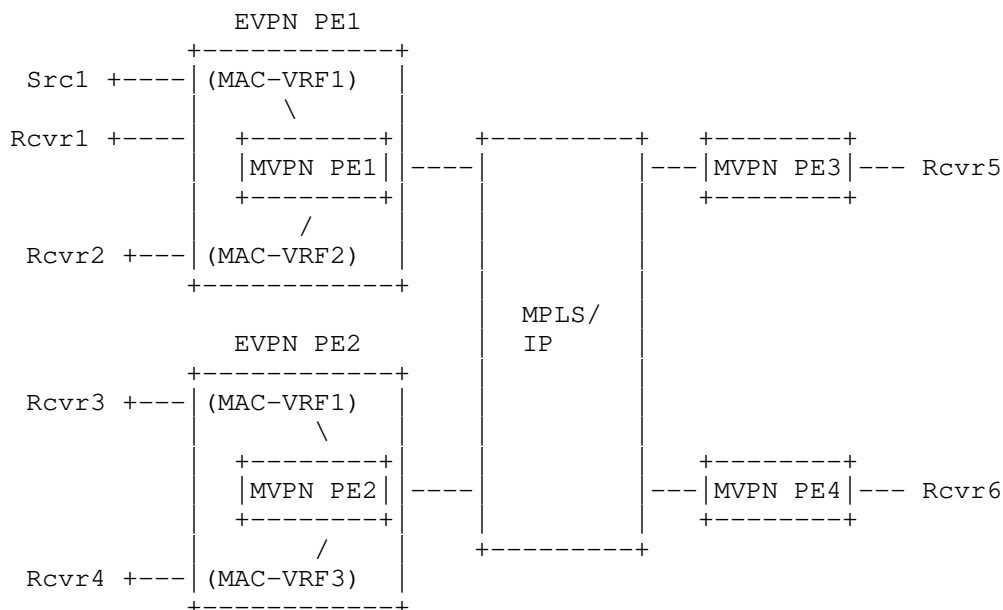


Figure-2: Modeling EVPN PEs as MVPN PEs

Although modeling an EVPN PE as a MVPN PE, conceptually simplifies the operation to that of a solution based on MVPN, the following operational aspects of EVPN need to be factored in when considering seamless integration between EVPN and MVPN PEs.

- o Unicast route advertisements for IP multicast source
- o Multi-homing of IP multicast sources and receivers
- o Mobility for Tenant's sources and receivers

5.3. Unicast Route Advertisements for IP multicast Source

When an IP multicast source is attached to an EVPN PE, the unicast route for that IP multicast source needs to be advertised. When the source is attached to a Single-Active multi-homed ES, then the EVPN DF PE is the PE that advertises a unicast route corresponding to the source IP address with VRF Route Import extended community which in turn is used as the Route Target for Join (S,G) messages sent toward the source PE by the remote PEs. The EVPN PE advertises this unicast route using EVPN route type 2 and IPVPN unicast route along with VRF Route Import extended community. EVPN route type 2 is advertised with the Route Targets corresponding to both IP-VRF and MAC-VRF/BT; whereas, IPVPN unicast route is advertised with RT corresponding to

the IP-VRF. When unicast routes are advertised by MVPN PEs, they are advertised using IPVPN unicast route along with VRF Route Import extended community per [RFC6514].

When the source is attached to an All-Active multi-homed ES, then the PE that learns the source advertises the unicast route for that source using EVPN route type 2 and IPVPN unicast route along with VRF Route Import extended community. EVPN route type 2 is advertised with the Route Targets corresponding to both IP-VRF and MAC-VRF/BT; whereas, IPVPN unicast route is advertised with RT corresponding to the IP-VRF. When the other multi-homing EVPN PEs for that ES receive this unicast EVPN route, they import the route and check to see if they have learned the route locally for that ES, if they have, then they do nothing. But if they have not, then they add the IP and MAC addresses to their IP-VRF and MAC-VRF/BT tables respectively with the local interface corresponding to that ES as the corresponding route adjacency. Furthermore, these PEs advertise an IPVPN unicast route along with VRF Route Import extended community and Route Target corresponding to IP-VRF to other remote PEs for that MVPN. Therefore, the remote PEs learn the unicast route corresponding to the source from all multi-homing PEs associated with that All-Active Ethernet Segment even though one of the multi-homing PEs may only have directly learned the IP address of the source.

EVPN-PEs advertise unicast routes as host routes using EVPN route type 2 for sources that are directly attached to a tenant BD that has been extended in the EVPN fabric. EVPN-PE may summarize sources (IP networks) behind a router that are attached to EVPN-PE or sources that are connected to a BD, which is not extended across EVPN fabric and advertises those routes with EVPN route type 5. EVPN host-routes are advertised as IPVPN host-routes to MVPN-PEs only in case of seamless interop mode.

Section 8 extends seamless interop procedures to EVPN only fabrics as an IRB solution for multicast. L3VPN provisioning is not needed between EVPN-PEs. EVPN-PEs only need to advertise unicast routes using EVPN route-type 2 or route-type 5 with VRF Route Import extended community and don't need to advertise IPVPN routes within EVPN only fabric.

Section 9 discusses DCI usecases, where EVPN and MVPN networks are connected using gateway model. In gateway model, EVPN-PE advertises unicast routes as IPVPN routes along with VRF extended community for all multicast sources attached behind EVPN-PEs. All IPVPN routes SHOULD be summarized while advertising to MVPN-PEs.

5.4. Multi-homing of IP Multicast Source and Receivers

EVPN [RFC7432] has extensive multi-homing capabilities that allows TSes to be multi-homed to two or more EVPN PEs in Single-Active or All-Active mode. In Single-Active mode, only one of the multi-homing EVPN PEs can receive/transmit traffic for a given subnet (a given BD) for that multi-homed Ethernet Segment (ES). In All-Active mode, any of the multi-homing EVPN PEs can receive/transmit unicast traffic but only one of them (the DF PE) can send BUM traffic to the multi-homed ES for a given subnet.

The multi-homing mode (Single-Active versus All-Active) of a TS source can impact the MVPN procedures as described below.

5.4.1. Single-Active Multi-Homing

When a TS source reside on an ES that is multi-homed to two or more EVPN PEs operating in Single-Active mode, only one of the EVPN PEs can be active for the source subnet on that ES. Therefore, only one of the multi-homing PE learns the unicast route of the TS source and advertises that using EVPN and IPVPN to other PEs as described previously.

A downstream PE that receives a Join/Prune message from a TS host/router, selects an Upstream Multicast Hop (UMH) which is the upstream PE that receives the IP multicast flow in case of Single-Active multi-homing. An IP multicast flow belongs to either a source-specific tree (S,G) or to a shared tree (*,G). We use the notation (X,G) to refer to either (S,G) or (*,G); where X refers to S in case of (S,G) and X refers to the Rendezvous Point (RP) for G in case of (*,G). Since the active PE (which is also the UMH PE) has advertised unicast route for X along with the VRF Route Import EC, the downstream PEs selects the UMH without any ambiguity based on MVPN procedures described in section 5.1 of [RFC6513].

The multi-homing PE that receives the IP multicast flow on its local AC, performs the following tasks:

- L2 switches the multicast traffic in its BT associated with the local AC over which it received the flow if there are any interested receivers for that subnet.
- L3 routes the multicast traffic to other BTs for other subnets if there are any interested receivers for those subnets.
- L3 routes the multicast traffic to other PEs per MVPN procedures.

The multicast traffic can be sent on Inclusive, Selective, or Aggregate-Selective tree. Regardless of what type of tree is used, only a single copy of the multicast traffic is received by the downstream PEs and the multicast traffic is forwarded optimally from the upstream PE to the downstream PEs.

5.4.2. All-Active Multi-Homing

When a TS source reside on an ES that is multi-homed to two or more EVPN PEs operating in All-Active mode, then any of the multi-homing PEs can learn the TS source's unicast route; however, that PE may not be the same PE that receives the IP multicast flow. Therefore, the procedures for Single-Active Multi-homing need to be augmented for All-Active scenario as below.

The multi-homing EVPN PE that receives the IP multicast flow on its local AC, needs to do the following task in additions to the ones listed in the previous section for Single-Active multi-homing: L2 switch the multicast traffic to other multi-homing EVPN PEs for that ES via a multicast tunnel which it is called intra-ES subnet tunnel. There will be a dedicated tunnel for this purpose which is different from inter-subnet overlay tree/tunnel setup by MVPN procedures.

When the multi-homing EVPN PEs receive the IP multicast flow via this tunnel, they treat it as if they receive the flow via their local ACs and thus perform the tasks mentioned in the previous section for Single-Active multi-homing. The tunnel type for this intra-ES subnet tunnel can be any of the supported tunnel types such as ingress-replication, P2MP tunnel, BIER, and Assisted Replication; however, given that vast majority of multi-homing ESes are just dual-homing, a simple ingress replication tunnel can serve well. For a given ES, since multicast traffic that is locally received by one multi-homing PE is sent to other multi-homing PEs via this intra-ES subnet tunnel, there is no need for sending the multicast tunnel via MVPN tunnel to these multi-homing PEs - i.e., MVPN multicast tunnels are used only for remote EVPN and MVPN PEs. Multicast traffic sent over this intra-ES subnet tunnel to other multi-homing PEs for a given ES can be either fixed or on demand basis.

By feeding IP multicast flow received on one of the EVPN multi-homing PEs to the interested EVPN PEs in the same multi-homing group, we have essentially enabled all the EVPN PEs in the multi-homing group to serve as UMH for that IP multicast flow. Each of these UMH PEs advertises unicast route for X in (X,G) along with the VRF Route Import EC to all PEs for that MVPN instance. The downstream PEs build a candidate UMH set based on procedures described in section 5.1 of [RFC6513] and pick a UMH from the set. It should be noted that both the default UMH selection procedure based on highest UMH PE

IP address and the UMH selection algorithm based on hash function specified in section 5.1.3 of [RFC6513] (which is also a MUST implement algorithm) result in the same UMH PE be selected by all downstream PEs running the same algorithm. However, in order to allow a form of "equal cost load balancing", the hash algorithm is recommended to be used among all EVPN and MVPN PEs. This hash algorithm distributes UMH selection for different IP multicast flows among the multi-homing PEs for a given ES.

Since all downstream PEs (EVPN and MVPN) use the same hash-based algorithm for UMH determination, they all choose the same upstream PE as their UMH for a given (X,G) flow and thus they all send their (X,G) join message via BGP to the same upstream PE. This results in one of the multi-homing PEs to receive the join message and thus send the IP multicast flow for (X,G) over its associated overlay tree even though all of the multi-homing PEs in the All-Active redundancy group have received the IP multicast flow (one of them directly via its local AC and the rest indirectly via the associated intra-ES subnet tunnel). Therefore, only a single copy of routed IP multicast flow is sent over the network regardless of overlay tree type supported by the PEs - i.e., the overlay tree can be of type selective or aggregate selective or inclusive tree. This gives the network operator the maximum flexibility for choosing any overlay tree type that is suitable for its network operation and still be able to deliver only a single copy of the IP multicast flows to the egress PEs. In other words, an egress PE only receives a single copy of the IP multicast flow over the network, because it either receives it via the EVPN intra-ES subnet tunnel or MVPN inter-subnet tunnel. Furthermore, if it receives it via MVPN inter-subnet tunnel, then only one of the multi-homing PEs associated with the source ES, sends the IP multicast traffic.

Since the network of interest for seamless interoperability between EVPN and MVPN PEs is MPLS, the EVPN handling of BUM traffic for MPLS network needs to be considered. EVPN [RFC7432] uses ESI MPLS label for split-horizon filtering of Broadcast/Unknown unicast/multicast (BUM) traffic from an All-Active multi-homing Ethernet Segment to ensure that BUM traffic doesn't get loop back to the same Ethernet Segment that it came from. This split-horizon filtering mechanism applies as-is for multicast IRB scenario because of using the intra-ES tunnel among multi-homing PEs. Since the multicast traffic received from a TS source on an All-Active ES by a multi-homing PE is bridged to all other multi-homing PEs in that group, the standard EVPN split-horizon filtering described in [RFC7432] applies as-is.

5.5. Mobility for Tenant's Sources and Receivers

When a tenant system (TS), source or receiver, is multi-homed behind a group of multi-homing EVPN PEs, then TS mobility SHALL be supported among EVPN PEs. Furthermore, such TS mobility SHALL only cause an temporary disruption to the related multicast service among EVPN and MVPN PEs. If a source is moved from one EVPN PE to another one, then the EVPN mobility procedure SHALL discover this move and a new unicast route advertisement (using both EVPN and IPVPN routes) is made by the EVPN PE where the source has moved to per section 5.3 above and unicast route withdraw (for both EVPN and IPVPN routes) is performed by the EVPN PE where the source has moved from.

The move of a source results in disruption of the IP multicast flow for the corresponding (S,G) flow till the new unicast route associated with the source is advertised by the new PE along with the VRF Route Import EC, the join messages sent by the egress PEs are received by the new PE, the multicast state for that flow is installed in the new PE and a new overlay tree is built for that source from the new PE to the egress PEs that are interested in receiving that IP multicast flow.

The move of a receiver results in disruption of the IP multicast flow to that receiver only till the new PE for that receiver discovers the source and joins the overlay tree for that flow.

6. Control Plane Operation

In seamless interop between EVPN and MVPN PEs, the control plane need to setup the following three types of multicast tunnels. The first two are among EVPN PEs and are associated with attached BD, but the third one is among EVPN and MVPN PEs and is associated with tenant-VRF

- 1) Intra-ES subnet tunnel
- 2) Intra-subnet BUM tunnel
- 3) Inter-subnet IP multicast tunnel

6.1. Intra-ES Subnet Tunnel

As described in section 5.4.2, when a multicast source is sitting behind an All-Active ES, then an intra-subnet multicast tunnel is needed among the multi-homing EVPN PEs for that ES to carry multicast flow received by one of the multi-homing PEs to the other PEs in that ES. We refer to this multicast tunnel as Intra-ES subnet tunnel. Vast majority of All-Active multi-homing for TOR devices in DC

networks are just dual-homing which means the multicast flow received by one of the dual-homing PE only needs to be sent to the other dual-homing PE. Therefore, a simple ingress replication tunnel is all that is needed. In case of multi-homing to three or more EVPN PEs, then other tunnel types such as P2MP, MP2MP, BIER, and Assisted Replication can be considered. It should be noted that this intra-ES subnet tunnel is only needed for All-Active multi-homing and it is not required for Single-Active multi-homing.

The EVPN PEs belonging to a given All-Active ES discover each other using EVPN Ethernet Segment route per procedures described in [RFC7432]. These EVPN PEs perform DF election per [RFC7432], [RFC8584], or other DF election algorithms to decide who is a DF for a given BD. If the BD belongs to a tenant that has IRB IP multicast enabled for it, then for fixed-mode, each PE sets up an intra-ES subnet tunnel to forward IP multicast traffic received locally on that BD to other multi-homing PE(s) for that ES. Therefore, IP multicast traffic received via a local attachment circuit is sent on this tunnel and on the associated IRB interface for that BT and other local attachment circuits if there are interested receivers for them. The other multi-homing EVPN PEs treat this intra-ES subnet tunnel just like their local ACs - i.e., the multicast traffic received over this tunnel is treated as if it is received via its local AC. Thus, the multi-homing PEs cannot receive the same IP multicast flow from an MVPN tunnel (e.g., over an IRB interface for that BD) because between a source behind a local AC versus a source behind a remote PE, the PE always chooses its local AC.

When all multihomed PE support [I-D.ietf-bess-evpn-igmp-mld-proxy], traffic may be forwarded on demand basis. Based on IGMP synchronization procedure specified in [I-D.ietf-bess-evpn-igmp-mld-proxy], join state may be synchronized between all multihomed PEs. Multihomed PE which receives the multicast traffic from its attached circuit, may send the traffic towards intra-ES subnet tunnel, only if it has received IGMP sync message from one of the multihomed PEs. Such extension is outside the scope of this document and may be covered in a separate document, if required.

If a source exists behind inter-subnet tunnel, it is possible that more than one multihomed PEs send MVPN join towards remote PE based on incoming join on their local interfaces. When the traffic is received on the inter-subnet tunnel, it is sent towards locally attached receivers. Only DF sends traffic towards multihomed ethernet segment. Traffic received on the inter-subnet tunnel, should not be sent towards Intra-ES subnet tunnel.

When ingress replication is used for intra-ES subnet tunnel, every PE in the All-Active multi-homing ES has all the information to setup these tunnels - i.e., a) each PE knows what are the other multi-homing PEs for that ES via EVPN Ethernet Segment route and can use this information to setup intra-ES subnet tunnel among themselves.

6.2. Intra-Subnet BUM Tunnel

As the name implies, this tunnel is setup to carry BUM traffic for a given subnet/BD among EVPN PEs. In [RFC7432], this overlay tunnel is used for transmission of all BUM traffic including tenant IP multicast traffic.

When an EVPN IRB PE operates in seamless interop mode, this tunnel is used for all broadcast, unknown-unicast, non-IP multicast traffic, and link-local IP multicast traffic - i.e., it is used for all BUM traffic except tenant IP multicast traffic. This tunnel is setup using IMET route for a given EVI/BD. The composition and advertisement of IMET routes are exactly per [RFC7432]. It should be noted that when an EVPN All-Active multi-homing PE uses both this tunnel as well as intra-ES subnet tunnel, there SHALL be no duplication of multicast traffic over the network because they carry different types of multicast traffic - i.e., intra-ES subnet tunnel among multi-homing PEs carries only tenant IP multicast traffic; whereas, intra-subnet BUM tunnel carries link-local IP multicast traffic and BUM traffic (w/ non-IP multicast).

6.3. Inter-Subnet IP Multicast Tunnel

As its name implies, this tunnel is setup to carry IP-only multicast traffic for a given tenant across all its subnets (BDs) among EVPN and MVPN PEs.

The following NLRIs from [RFC6514] is used for setting up this inter-subnet tunnel in the network.

Intra-AS I-PMSI A-D route is used for the setup of default underlay tunnel (also called inclusive tunnel) for a tenant IP-VRF. The tunnel attributes are indicated using PMSI attribute with this route.

S-PMSI A-D route is used for the setup of Customer flow specific underlay tunnels. This enables selective delivery of data to PEs having active receivers and optimizes fabric bandwidth utilization. The tunnel attributes are indicated using PMSI attribute with this route.

Each EVPN PE supporting a specific MVPN instance discovers the set of other PEs in its AS that are attached to sites of that MVPN using Intra-AS I-PMSI A-D route (route type 1) per [RFC6514]. It can also discover the set of other ASes that have PEs attached to sites of that MVPN using Inter-AS I-PMSI A-D route (route type 2) per [RFC6514]. After the discovery of PEs that are attached to sites of the MVPN, an inclusive overlay tree (I-PMSI) can be setup for carrying tenant multicast flows for that MVPN; however, this is not a requirement per [RFC6514] and it is possible to adopt a policy in which all tenant flows are carried on S-PMSIs.

An EVPN-IRB PE sends a tenant IP multicast flow to other EVPN and MVPN PEs over this inter-subnet tunnel that is instantiated using MVPN I-PMSI or S-PMSI. This tunnel can be considered as being originated and terminated from/to among IP-VRFs of EVPN/MVPN PEs; whereas, intra-subnet tunnel is originated/terminated among MAC-VRFs of EVPN PEs.

6.4. IGMP Hosts as TSes

IGMP messages are terminated by the EVPN-IRB PE and tenant (*,G) or (S,G) join messages are sent via MVPN Shared Tree Join route (route type 6) or Source Tree Join route (route type 7) respectively of MCAST-VPN NLRI per [RFC6514].

Here, IGMP states are terminated at IRB interfaces and local interest are expressed in the context of IP-VRF to remote PEs. Hence, If a tenant system which is an IGMP host is multi-homed to two or more EVPN PEs using All-Active multi-homing, there is no need to sync IGMP join and leave messages between these EVPN PEs using EVPN IGMP Join Synch route (route type 7) and EVPN IGMP Leave Synch route (route type 8) per [I-D.ietf-bess-evpn-igmp-mld-proxy].

In case of a network with only IGMP hosts, the preferred mode of operation is that of Shortest Path Tree(SPT) per section 14 of [RFC6514]. This mode is only supported for PIM-SM and avoids the RP configuration overhead. Such mode is chosen by provisioning/configuration.

6.5. PIM Routers as TSes

Just like a MVPN PE, an EVPN PE runs a separate tenant multicast routing instance (VPN-specific) per MVPN instance and the following tenant multicast routing instances are supported:

- PIM Sparse Mode (PIM-SM) with the ASM service model
- PIM Sparse Mode with the SSM service model
- PIM Bidirectional Mode (BIDIR-PIM), which uses bidirectional tenant-trees to support the ASM service model

A given tenant's PIM join messages for (*,G) or (S, G) are processed by the corresponding tenant multicast routing protocol and they are advertised over MPLS/IP network using Shared Tree Join route (route type 6) and Source Tree Join route (route type 7) respectively of MCAST-VPN NLRI per [RFC6514].

7. Data Plane Operation

When an EVPN-IRB PE receives an IGMP/MLD join message over one of its Attachment Circuits (ACs), it adds that AC to its Layer-2 (L2) OIF list. This L2 OIF list is associated with the MAC-VRF/BT corresponding to the subnet of the tenant device that sent the IGMP/MLD join. Therefore, tenant (S,G) or (*,G) forwarding entries are created/updated for the corresponding MAC-VRF/BT based on these source and group IP addresses. Furthermore, the IGMP/MLD join message is propagated over the corresponding IRB interface and it is processed by the tenant multicast routing instance which creates the corresponding tenant (S,G) or (*,G) Layer-3 (L3) forwarding entries. It adds this IRB interface to the L3 OIF list. An IRB is removed as a L3 OIF when all L2 tenant (S,G) or (*,G) forwarding states is removed for the MAC-VRF/BT associated with that IRB. Furthermore, tenant (S,G) or (*,G) L3 forwarding state is removed when all of its L3 OIFs are removed - i.e., all the IRB and L3 interfaces associated with that tenant (S,G) or (*,G) are removed.

When an EVPN PE receives IP multicast traffic from one of its AC, if it has any attached receivers for that subnet, it performs L2 switching of the intra-subnet traffic within the BT attached to that AC. If the multicast flow is received over an AC that belongs to an All-Active ES, then the multicast flow is also sent over the intra-ES subnet tunnel among multi-homing PEs. The EVPN PE then sends the multicast traffic over the corresponding IRB interface. The multicast traffic then gets routed in the corresponding IP-VRF and it gets forwarded to interfaces in the L3 OIF list which can include other IRB interfaces, other L3 interfaces directly connected to TSes, and the MVPN Inter-Subnet tunnel which is instantiated by an I-PMSI or S-PMSI tunnel. When the multicast packet is routed within the IP-VRF of the EVPN PE, its Ethernet header is stripped and its TTL gets decremented as the result of this IP routing. Remote multicast traffic that is received from MVPN Inter-Subnet tunnel gets routed towards all L3 OIFs. When the multicast traffic is received on an IRB interface by the BT corresponding to that interface, it gets L2 switched and sent over ACs that belong to the L2 OIF list.

7.1. Intra-Subnet L2 Switching

Rcvr1 in Figure 1 is connected to PE1 in MAC-VRF1 (same as Src1) and sends IGMP join for (C-S, C-G), IGMP snooping will record this state in local bridging entry. A routing entry will be formed as well which will point to MAC-VRF1 as RPF for Src1. We assume that Src1 is known via ARP or similar procedures. Rcvr1 will get a locally bridged copy of multicast traffic from Src1. Rcvr3 is also connected in MAC-VRF1 but to PE2 and hence would send IGMP join which will be recorded at PE2. PE2 will also form routing entry and RPF will be assumed as Tenant Tunnel "Tenant1" formed beforehand using MVPN procedures. Also this would cause multicast control plane to initiate a BGP MCAST-VPN type 7 route which would include VRI for PE1 and hence be accepted on PE1. PE1 will include Tenant1 tunnel as Outgoing Interface (OIF) in the routing entry. Now, since it has knowledge of remote receivers via MVPN control plane it will encapsulate original multicast traffic in Tenant1 tunnel towards core.

7.2. Inter-Subnet L3 Routing

Rcvr2 in Figure 1 is connected to PE1 in MAC-VRF2 and hence PE1 will record its membership in MAC-VRF2. Since MAC-VRF2 is enabled with IRB, it gets added as another OIF to routing entry formed for (C-S, C-G). Rcvr2 and Rcvr4 are also in different MAC-VRFs than multicast speaker Src1 and hence need Inter-subnet forwarding. PE2 now adds another OIF 'MAC-VRF2' to its existing routing entry. But there is no change in control plane states since it is already sent MVPN route and no further signaling is required. Traffic received on the tenant tunnel interface gets routed towards both MAC-VRF1 and MAC-VRF3. PE3 forms routing entry very similar to PE2. It is to be noted that PE3 does not have MAC-VRF1 configured locally but still can receive the multicast data traffic over Tenant1 tunnel formed due to MVPN procedures and routes traffic towards its L3 OIFs for that (C-S,C-G).

8. DCs with only EVPN PEs

As mentioned earlier, the proposed solution can be used as a routed multicast solution in data center networks with only EVPN PEs (e.g., routed multicast VPN only among EVPN PEs).

As per section 5.2, EVPN PE is modeled as a PE that consists of a MVPN PE whose routed interfaces (e.g., attachment circuits) are replaced with IRB interfaces connecting each IP-VRF of the MVPN PE to a set of BTs. Due to this, the IP multicast traffic that needs to be forwarded from the source PE to remote PEs is routed to remote PEs regardless of whether the traffic is intra-subnet or inter-subnet.

As the result, the TTL value for intra-subnet traffic that spans across two or more PEs get decremented.

However, if there are applications that require intra-subnet multicast traffic to be L2 forwarded, Appendix A discusses some options to support applications having TTL value 1. The procedure discussed in Appendix A may be used to support applications that require intra-subnet multicast traffic to be L2 forwarded.

8.1. Setup of overlay multicast delivery

It must be emphasized that this solution poses no restriction on the setup of the tenant BDs and that neither the source PE, nor the receiver PEs do not need to know/learn about the BD configuration on other PEs in the tenant VRF (Since EVPN PE is modelled as MVPN PE, source and receivers are announced to remote PE in the context of tenant VRF(MVPN) as opposed to BD context). The Reverse Path Forwarder (RPF) is selected per the tenant multicast source and the IP-VRF in compliance with the procedures in [RFC6514], using the incoming EVPN route type 2 or 5 NLRI per [RFC7432].

The VRF Route Import (VRI) extended community that is carried with the IPVPN routes in [RFC6514] MUST be carried with the EVPN unicast routes when these routes are used. The construction and processing of the VRI are consistent with [RFC6514]. The VRI MUST uniquely identify the PE which is advertising a multicast source and the IP-VRF it resides in.

VRI is constructed as following:

- The 4-octet Global Administrator field MUST be set to an IP address of the PE. This address SHOULD be common for all the IP-VRFs on the PE (e.g., this address may be the PE's loopback address or VTEP address).
- The 2-octet Local Administrator field associated with a given IP-VRF contains a number that uniquely identifies that IP-VRF within the PE that contains the IP-VRF.

EVPN PE MUST have Route Target Extended Community to import/export MVPN routes. In data center environment, it is desirable to have this RT configured using auto-generated method than static configuration.

The following is one recommended model to auto-generate MVPN RT:

- The Global Administrator field of the MVPN RT MAY be set to BGP AS Number.
- The Local Administrator field of the MVPN RT MAY be set to the VNI associated with the tenant VRF.

Every PE which detects a local receiver via a local IGMP join or a local PIM join for a specific source (overlay SSM mode) MUST terminate the IGMP/PIM signaling at the IP-VRF and generate a (C-S,C-G) via the BGP MCAST-VPN route type 7 per [RFC6514] if and only if the RPF for the source points to the fabric. If the RPF points to a local multicast source on the same MAC-VRF or a different MAC-VRF on that PE, the MCAST-VPN MUST NOT be advertised and data traffic will be locally routed/bridged to the receiver.

The VRI received with EVPN route type 2 or 5 NLRI from source PE will be appended as an export route-target extended community. The PE which has advertised the unicast route with VRI, will import the incoming MCAST-VPN NLRI in the IP-VRF with the same import route-target extended-community and other PEs SHOULD ignore it. Following such procedure the source PE learns about the existence of at least one remote receiver in the tenant overlay and programs data plane accordingly so that a single copy of multicast data is forwarded into the fabric using tenant VRF tunnel (i.e. inter-subnet tunnel/mvpn tunnel).

If the multicast source is unknown (overlay ASM mode), the MCAST-VPN route type 6 (C-*,C-G) join SHOULD be targeted towards the designated overlay Rendezvous Point (RP) by appending the received RP VRI as an export route-target extended community. Every PE which detects a local source, registers with its RP PE. That is how the RP learns about the tenant source(s) and group(s) within the MVPN. Once the overlay RP PE receives either the first remote (C-RP,C-G) join or a local IGMP/PIM join, it will trigger an MCAST-VPN route type 7 (C-S,C-G) towards the actual source PE for which it has received PIM register message in full compliance with regular PIM procedures. This involves the source PE to advertise the MCAST-VPN Source Active A-D route (MCAST-VPN route-type 5) towards all PEs. The Source Active A-D route is used to inform all PEs in a given MVPN about the active multicast source for switching from RPT to SPT when MVPNs use tenant RP-shared trees (i.e., rooted at tenant's RP) per section 13 of [RFC6514].

8.2. Handling of different encapsulations

Just as in [RFC6514] the MVPN I-PMSI and S-PMSI A-D routes are used to form the overlay multicast tunnels and signal the tunnel type

using the P-Multicast Service Interface Tunnel (PMSI Tunnel) attribute.

8.2.1. MPLS Encapsulation

The [RFC6514] assumes MPLS/IP core and there is no modification to the signaling procedures and encoding for PMSI tunnel formation therein. Also, there is no need for a gateway to inter-operate with non-EVPN PE supporting [RFC6514] based MVPN over IP/MPLS.

8.2.2. VxLAN Encapsulation

In order to signal VXLAN, the corresponding BGP encapsulation extended community [RFC9012] SHOULD be appended to the MVPN I-PMSI and S-PMSI A-D routes. The MPLS label in the PMSI Tunnel Attribute MUST be the Virtual Network Identifier (VNI) associated with the customer MVPN. The supported PMSI tunnel types with VXLAN encapsulation are: PIM-SSM Tree, PIM-SM Tree, BIDIR-PIM Tree, Ingress Replication [RFC6514]. Further details are in [RFC8365].

In this case, a gateway is needed for inter-operation between the EVPN PEs and non-EVPN MVPN PEs. The gateway should re-originate the control plane signaling with the relevant tunnel encapsulation on either side. In the data plane, the gateway terminates the tunnels formed on either side and performs the relevant stitching/re-encapsulation on data packets.

8.2.3. Other Encapsulation

In order to signal a different tunneling encapsulation such as NVGRE, GPE, or GENEVE the corresponding BGP encapsulation extended community [RFC9012] SHOULD be appended to the MVPN I-PMSI and S-PMSI A-D routes. If the Tunnel Type field in the encapsulation extended-community is set to a type which requires Virtual Network Identifier (VNI), e.g., VXLAN-GPE or NVGRE [RFC9012], then the MPLS label in the PMSI Tunnel Attribute MUST be the VNI associated with the customer MVPN. Same as in VXLAN case, a gateway is needed for inter-operation between the EVPN-IRB PEs and non-EVPN MVPN PEs.

9. DCI with MPLS in WAN and VxLAN in DCs

This section describes the inter-operation between MVPN PEs in WAN using MPLS encapsulation with EVPN PEs in a DC network using VxLAN encapsulation. Since the tunnel encapsulation between these networks are different, we must have at least one gateway in between. Usually, two or more are required for redundancy and load balancing purpose. In such scenarios, a DC network can be represented as a customer network that is multi-homed to two or more MVPN PEs via L3

interfaces and thus standard MVPN multi-homing procedures are applicable here. It should be noted that a MVPN overlay tunnel over the DC network is terminated on the IP-VRF of the gateway and not the MAC-VRF/BTs. Therefore, the considerations for loop prevention and split-horizon filtering described in [RFC9014] are not applicable here. .

9.1. Control plane inter-connect

The gateway(s) MUST be setup with the inclusive set of all the IP-VRFs that span across the two domains. On each gateway, there will be at least two BGP sessions: one towards the DC side and the other towards the WAN side. Usually for redundancy purpose, more sessions are setup on each side. The unicast route propagation follows the exact same procedures in [RFC9014]. Hence, a multicast host located in either domain, is advertised with the gateway IP address as the next-hop to the other domain. As a result, PEs view the hosts in the other domain as directly attached to the gateway and all inter-domain multicast signaling is directed towards the gateway(s). Received MVPN routes type 1-7 from either side of the gateway(s), MUST NOT be reflected back to the same side but processed locally and re-advertised (if needed) to the other side:

- o Intra-AS/Inter-AS I-PMSI A-D Route: these are distributed within each domain to form the overlay tunnels which terminate at gateway(s). They are not passed to the other side of the gateway(s).
- o C-Multicast Route: joins are imported into the corresponding IP-VRF on each gateway and advertised as a new route to the other side with the following modifications (the rest of NLRI fields and path attributes remain on-touched):
 - * Route-Distinguisher is set to that of the IP-VRF
 - * Route-target is set to the exported route-target list on IP-VRF
 - * The PMSI tunnel attribute and BGP Encapsulation extended community will be modified according to section 8
 - * Next-hop will be set to the IP address which represents the gateway on either domain
- o Source Active A-D Route: same as joins
- o S-PMSI A-D Route: these are passed to the other side to form selective PMSI tunnels per every (C-S,C-G) from the gateway to the PEs in the other domain provided it contains receivers for the

given (C-S, C-G). Similar modifications made to joins are made to the newly originated S-PMSI.

In addition, the Originating Router's IP address is set to GW's IP address. Multicast signaling from/to hosts on local ACs on the gateway(s) are generated and propagated in both domains (if needed) per the procedures in section 6 in this document and in [RFC6514] with no change. It must be noted that for a locally attached source, the gateway will program an OIF per every domain from which it receives a remote join in its forwarding plane and different encapsulation will be used on the data packets.

9.2. Data plane inter-connect

Traffic forwarding procedures on gateways are same as those described for PEs in section 5 except that, unlike a non-border leaf PE, the gateway will not only route the incoming traffic from one side to its local receivers, but will also send it to the remote receivers in the other domain after de-capsulation and appending the right encapsulation. The OIF and IIF are programmed in FIB based on the received joins from either side and the RPF calculation to the source or RP. The de-capsulation and encapsulation actions are programmed based on the received I-PMSI or S-PMSI A-D routes from either side.

The multicast traffic from local sources on each gateway may flow to the other gateway with either of the tunnel encapsulation. But, it is recommended to use VxLAN tunnel than MPLS in this case.

10. Interop with L2 EVPN PEs

A gateway device is needed to do interop between EVPN PEs that support seamless interop procedure specified in this document and L2EVPN-PEs. A tenant domain can be provisioned with one or more such gateway devices known as "Seamless interop EVPN Multicast Gateway (SEMG)". PE that is configured as SEMG must be provisioned with all BDs that are available in the tenant domain.

When advertising IMET route for a BD, PE configured as SEMG advertises EVPN Multicast Flags Extended Community with SEMG flag set. Given set of eligible PEs, one PE is selected as the SEMG designated forwarder (SEMG-DF). PE should use procedure specified in [RFC8584] for the SEMG DF election.

There are multiple possibilities that need to be considered here.

- o L2EVPN PE may or may not have support for [I-D.ietf-bess-evpn-igmp-mld-proxy]

- o Seamless interop PE may or may not support [I-D.ietf-bess-evpn-igmp-mld-proxy]
- o Network may only have L2EVPN PE and Seamless interop capable PE
- o Network may have L2EVPN PE, Seamless interop capable PE and MVPN PE.

Multicast sources and receivers can exist anywhere in the network. These usecases are discussed below.

10.1. Interaction with L2EVPN PE and Seamless interop capable PE

The following cases are considered in this section.

- o Case1: [I-D.ietf-bess-evpn-igmp-mld-proxy] is supported both at seamless interop capable PE and L2EVPN PE.
- o Case2: [I-D.ietf-bess-evpn-igmp-mld-proxy] is supported only at seamless interop capable PE.
- o Case3: [I-D.ietf-bess-evpn-igmp-mld-proxy] is not supported at interop capable PE.

[I-D.ietf-bess-evpn-igmp-mld-proxy] support is recommended for seamless interop capable PE. SEMG can group L2 EVPN PEs into two separate groups (one that supports the [I-D.ietf-bess-evpn-igmp-mld-proxy] and another that doesn't) from IMET routes that it receives from the remote peers. The interop procedure for handling these two different sets of remote L2 EVPN PEs are captured in case 1 and 2.

Case 1: [I-D.ietf-bess-evpn-igmp-mld-proxy] is supported both at seamless interop capable PE and L2EVPN PE

This may be the most common usecase.

SEMG-DF has the following special responsibilities on a BD for which it is the DF.

- o Process EVPN SMET routes from the remote L2 EVPN PEs that support [I-D.ietf-bess-evpn-igmp-mld-proxy] and creates L2 multicast state. SMET route in-turn triggers the creation of L3 multicast state similar to IGMP join received on the local AC. SEMG-DF exercises the MVPN procedures for the join.
- o It should not process IGMP control packets from L2EVPN PE that supports [I-D.ietf-bess-evpn-igmp-mld-proxy].

- o Originate SMET(*,*) route towards L2 EVPN PEs. This is to receive traffic from multicast sources that are connected behind L2 EVPN PEs.
- o When SEMG-DF receives traffic from L2 EVPN PE on the intra-subnet tunnel on BD-X, it does the following
 - * Performs FHR functionality
 - * Advertises the host route with L3 label and VRF Route-Import corresponds to the tenant domain.
 - * Sends the traffic towards the locally attached receivers.
 - * Sends the traffic towards L2EVPN receiver on BDs other than incoming BD(after multicast routing)
 - * Sends the traffic towards remote seamless interop capable PEs, where receivers are attached/connected behind that PE.
- o When SEMG-DF receives traffic from the MVPN tunnel, it does the following
 - * Sends the traffic towards the IRB interfaces, where receiver exists
 - * BD corresponding to the IRB interfaces may have local receivers or remote receivers behind L2 EVPN PE. SEMG-DF sends the traffic on the intra-subnet tunnel for remote receivers.

Case 2: [I-D.ietf-bess-evpn-igmp-mld-proxy] is not supported at L2 EVPN PE

This case only differs from case 1 in terms of the way it learns receivers behind L2 EVPN PEs and how SEMG-DF attracts traffic from sources behind L2 EVPN PE. Rest of procedures specified above is applicable for this case.

SEMG-DF has the following special responsibilities on a BD for which it is the DF

- o Process IGMP control packets from remote L2 EVPN PEs that doesn't support [I-D.ietf-bess-evpn-igmp-mld-proxy] and create L2 and L3 state.
- o When an IGMP query is received on the intra-subnet tunnel on BD-X, SEMG-DF needs to send proxy IGMP reports for all groups that it has learned from remote L2-EVPN PEs on that BD.

- o Connecting multicast router behind L2 EVPN PE is not recommended. If a multicast router is connected behind L2 EVPN PE, the BD corresponds to VRF tunnel needs to be configured in the L2 EVPN PE so that PIM router may get all joins that are received in the BD corresponds to MVPN tunnel interface at SEMG-DF.
- o SEMG-DF should get all multicast traffic from L2EVPN PEs. This may be achieved by sending IGMP query or PIM hello on the intra-subnet tunnel

Case 3: [I-D.ietf-bess-evpn-igmp-mld-proxy] is not supported at seamless interop capable PE

The procedure of handling this use case is exactly the same as case 2.

All seamless interop capable PEs other than SEMG should discard SMET routes that are coming from L2EVPN PEs and must discard all IGMP control packets, if any received on the intra-subnet tunnel. SEMG should discard incoming SMET routes and IGMP joins from L2EVPN PEs, if it is not the DF for the incoming BD.

When [I-D.ietf-bess-evpn-igmp-mld-proxy] is supported both at seamless interop capable PE and L2EVPN PE, selective forwarding is done based on receiver interest at the egress-PE, when overlay tunnel type is Ingress-replication or selective tunnel.

10.2. Network having L2EVPN PE, Seamless interop capable PE and MVPN PE

Since MVPN-PE can only interact with Seamless interop capable PEs, SEMG-DF acts as FHR and LHR for sources and receivers behind L2 EVPN PE. Only SEMG-DF advertises IPVPN unicast route along with VRF Route Import extended community for hosts behind L2 EVPN PE. No additional procedures are required, when they all co-exist.

11. Connecting external Multicast networks or PIM routers.

External multicast networks or PIM routers can be attached to any seamless interop capable EVPN-PEs or set of EVPN-PEs or MVPN-PEs. Multicast network or PIM router can also be attached to any IRB enabled interface or set of interfaces. The fabric can be used as a Transit network for connecting the external multicast networks. All PIM signaling is terminated at PE's IRB interfaces.

No additional procedures are required while connecting external multicast networks.

12. TS RP options

RP can be configured in the EVPN-PE itself in the tenant VRF or in the external multicast networks connected behind an EVPN PE or in the MVPN network. When RPF is not local to EVPN-PE, EVPN-PE operates in rpt-spt mode as PER procedures specified in section 13 of [RFC6514].

EVPN fabric without having any external multicast network/attached MVPN network, doesn't need RP configuration. A configuration option SHALL be provided to the end user to operate the fabric in RP less mode. When an EVPN-PE is operating in RP-less mode, EVPN-PE MUST advertise all attached sources to remote EVPN PEs using procedure specified in [RFC6514].

In RP less mode, (C-*,C-G) RPF may be set to NULL or may be set to wild card interface(Any interface on the tenant VRF). In RP-less mode, traffic is always forwarded based on (C-S,C-G) state.

13. IANA Considerations

IANA has allocated the codepoint for Multicast Flags Extended Community which is defined in [I-D.ietf-bess-evpn-igmp-mld-proxy].

The Multicast Flags Extended Community contains a 16-bit Flags field. The bits are numbered 0-15, from high-order to low-order. IANA is requested to assign the following new flags in the "Multicast Flags Extended Community Flags" registry.

Bit	Name	Reference
-----	-----	-----
0-11	Unassigned	
12	SEMG	This document
13	Seamless interop capable PE	This document
14	MLD Proxy Support	[I-D.ietf-bess-evpn-igmp-mld-proxy]
15	IGMP Proxy Support	[I-D.ietf-bess-evpn-igmp-mld-proxy]

14. Security Considerations

All the security considerations in [RFC7432], [RFC6513] and [RFC6514] apply directly to this document because this document leverages these RFCs control plane and their associated procedures.

15. Acknowledgements

The authors would like to thank Niloofar Fazlollahi, Aamod Vyavaharkar, Raunak Banthia, and Swadesh Agrawal for their discussions and contributions.

16. References

16.1. Normative References

- [I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Mishra, M., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-13 (work in progress), September 2021.
- [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>>.
- [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>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.
- [RFC9014] Rabadan, J., Ed., Sathappan, S., Henderickx, W., Sajassi, A., and J. Drake, "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, DOI 10.17487/RFC9014, May 2021, <<https://www.rfc-editor.org/info/rfc9014>>.

- [RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <<https://www.rfc-editor.org/info/rfc9135>>.

16.2. Informative References

- [RFC4389] Thaler, D., Talwar, M., and C. Patel, "Neighbor Discovery Proxies (ND Proxy)", RFC 4389, DOI 10.17487/RFC4389, April 2006, <<https://www.rfc-editor.org/info/rfc4389>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.
- [RFC7080] Sajassi, A., Salam, S., Bitar, N., and F. Balus, "Virtual Private LAN Service (VPLS) Interoperability with Provider Backbone Bridges", RFC 7080, DOI 10.17487/RFC7080, December 2013, <<https://www.rfc-editor.org/info/rfc7080>>.
- [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>>.

Appendix A. Supporting application with TTL value 1

It is possible that some deployments may have a host on the tenant domain that sends multicast traffic with TTL value 1. The interested receiver for that traffic flow may be attached to different PEs on the same subnet. The procedures specified in section 5 always routes the traffic between PEs for both intra and inter subnet traffic. Hence traffic with TTL value 1 is dropped due to the nature of routing.

This section discusses few possible ways to support traffic having TTL value 1 or traffic that require L2 bridging behavior. Implementation MAY support any of the following model.

A.1. Policy based model

Policies may be used to enforce EVPN BUM procedure for traffic flows with TTL value 1. Traffic flow that matches the policy is excluded from seamless interop procedure specified in this document, hence TTL decrement issue will not apply.

A.2. Exercising BUM procedure for VLAN/BD

Servers/hosts sending the traffic with TTL value 1 may be attached to a separate VLAN/BD, where multicast routing is disabled. When multicast routing is disabled, EVPN BUM procedure may be applied to all traffic ingressing on that VLAN/BD. On the Egress PE, the RPF for such traffic may be set to BD interface, where the source is attached.

A.3. Intra-subnet bridging

The procedure specified in the section enables a PE to detect an attached subnet source (i.e., source that is directly attached in the tenant BD/VLAN). By applying the following procedure for the attached source, Traffic flows having TTL value 1 can be supported.

- On the ingress PE, do the bridging on the interface towards the core interface
- On the egress side, make a decision whether to bridge or route at the outgoing interface (OIF) based on whether the source is attached to the OIF's BD/VLAN or not.

Recent ASIC supports single lookup forwarding for bridging and routing (L2+L3). The procedure mentioned here leverages this ASIC capability.

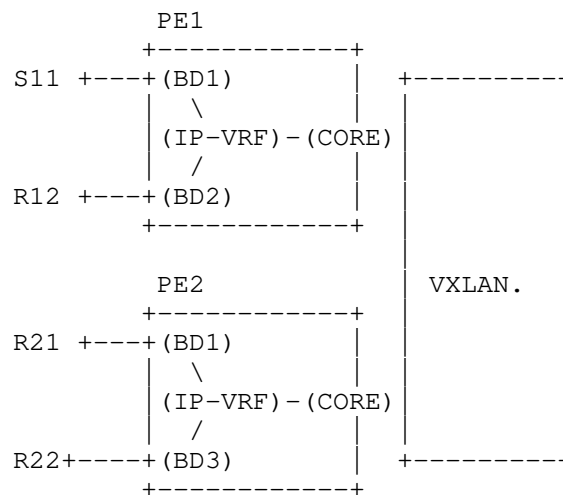


Figure 3 Intra-subnet bridging

Consider the above picture. In the picture

- PE1 and PE2 are seamless interop capable PEs
- S11 is a multicast host directly attached to PE1 in BD1
- Source S11 sends traffic to Group G11
- R21, R22 are IGMP receivers for group G11
- R21 and R22 are attached to BD1 and BD3 respectively at PE2.

When source S11 starts sending the traffic, PE1 learns the source and announces the source using MVPN procedures to the remote PEs.

At PE2, IGMP joins from R21, R22 result the creation of (*,G11) entry with outgoing OIF as IRB interface of BD1 and BD3. When PE2 learns the source information from PE1, it installs the route (S11, G11) at the tenant VRF with RPF as CORE interface.

PE2 inherits (*, G11) OIFs to (S11, G11) entry. While inheriting OIF, PE2 checks whether source is attached to OIF's subnet. OIF matching source subnet is added with flag indicating bridge only interface. In case of (S11, G11) entry, BD1 is added as the bridge only OIF, while BD3 is added as normal OIF(L3 OIF). PEs (PE2) sends MVPN join (S11, G11) towards PE1, since it has local receivers.

At Ingress PE(PE1), CORE interface is added to (S11, G11) entry as an OIF (outgoing interface) with a flag indicating that bridge only interface. With this procedure, ingress PE(PE1) bridges the traffic on CORE interface. (PE1 retains the TTL and source-MAC). The traffic is encapsulated with VNI associated with CORE interface. PE1 also routes the traffic for R12 which is attached to BD2 on the same device.

PE2 decapsulates the traffic from PE1 and does inner lookup on the tenant VRF associated with incoming VNI. Traffic lookup on the tenant VRF yields (S11, G11) entry as the matching entry. Traffic gets bridged on BD1 (PE2 retains the TTL and source-MAC) since the OIF is marked as bridge only interface. Traffic gets routed on BD2.

Authors' Addresses

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US

Email: sajassi@cisco.com

Kesavan Thiruvengkatasamy
Cisco
170 West Tasman Drive
San Jose, CA 95134, US

Email: kethiruv@cisco.com

Samir Thoria
Cisco
170 West Tasman Drive
San Jose, CA 95134, US

Email: sthoria@cisco.com

Ashutosh Gupta
VMware
3401 Hillview Ave, Palo Alto, CA 94304

Email: ashutoshgupta@vmware.com

Luay Jalil
Verizon

Email: luay.jalil@verizon.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2022

A. Sajassi, Ed.
G. Badoni
P. Warade
S. Pasupula
Cisco Systems
J. Drake, Ed.
Juniper
J. Rabadan, Ed.
Nokia
October 25, 2021

EVPN Support for L3 Fast Convergence and Aliasing/Backup Path
draft-sajassi-bess-evpn-ip-aliasing-03

Abstract

This document proposes an EVPN extension to allow several of its multihoming functions, fast convergence and aliasing/backup path, to be used in conjunction with inter-subnet forwarding.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 28, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

1. Introduction	2
1.1. Ethernet Segments for Host Routes in Symmetric IRB . . .	3
1.2. Inter-subnet Forwarding for Prefix Routes in the Interface-less IP-VRF-to-IP-VRF Model	4
1.3. Ethernet Segments for Prefix routes in IP-VRF-to-IP-VRF use-cases	5
1.4. Terminology and Conventions	6
2. Ethernet Segments for L3 Aliasing/Backup Path and Fast Convergence	8
3. IP Aliasing and Backup Path	9
3.1. Constructing the IP A-D per EVI Route	10
4. Fast Convergence for Routed Traffic	10
4.1. Constructing IP A-D per Ethernet Segment Route	11
4.1.1. IP A-D per ES Route Targets	11
4.2. Avoiding convergence issues by synchronizing IP prefixes	11
4.3. Handling Silent Host MAC/IP route for IP Aliasing	12
4.4. MAC Aging	12
5. Determining Reachability to Unicast IP Addresses	13
5.1. Local Learning	13
5.2. Remote Learning	13
5.3. Constructing the EVPN IP Routes	13
5.3.1. Route Resolution	13
6. Forwarding Unicast Packets	14
7. Load Balancing of Unicast Packets	14
8. IP Aliasing and Unequal ECMP for IP Prefix Routes	14
9. Security Considerations	15
10. IANA Considerations	15
11. Contributors	15
12. Acknowledgments	15
13. References	15
13.1. Normative References	15
13.2. Informative References	16
Authors' Addresses	16

1. Introduction

This document proposes an EVPN extension to allow several of its multihoming functions, fast convergence and aliasing/backup path, to be used in conjunction with inter-subnet forwarding. It re-uses the existing EVPN routes, the Ethernet A-D per ES and the Ethernet A-D per EVI routes, which are used for these multihoming functions. In

particular, there are three use-cases that could benefit from the use of these multihoming functions:

- a. Inter-subnet forwarding for host routes in symmetric IRB [I-D.ietf-bess-evpn-inter-subnet-forwarding].
- b. Inter-subnet forwarding for prefix routes in the interface-less IP-VRF-to-IP-VRF model [I-D.ietf-bess-evpn-prefix-advertisement].
- c. Inter-subnet forwarding for prefix routes when the ESI is used exclusively as an L3 construct [I-D.ietf-bess-evpn-prefix-advertisement].

1.1. Ethernet Segments for Host Routes in Symmetric IRB

Consider a pair of multi-homing PEs, PE1 and PE2, as illustrated in Figure 1. Let there be a host H1 attached to them. Consider PE3 and a host H3 attached to it.

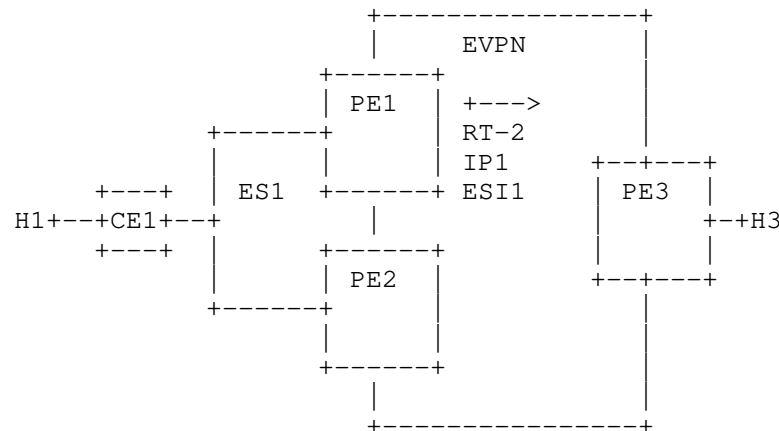


Figure 1: Inter-subnet traffic between Multihoming PEs and Remote PE

With Asymmetric IRB [I-D.ietf-bess-evpn-inter-subnet-forwarding], if H3 sends inter-subnet traffic to H1, routing will happen at PE3. PE3 will be attached to the destination IRB interface and will trigger ARP/ND requests if it does not have an ARP/ND adjacency to H1. A subsequent routing lookup will resolve the destination MAC to H1's MAC address. Furthermore, H1's MAC will point to an ECMP EVPN destination on PE1 and PE2, either due to host route advertisement from both PE1 and PE2, or due to Ethernet Segment MAC Aliasing as detailed in [RFC7432].

With Symmetric IRB [I-D.ietf-bess-evpn-inter-subnet-forwarding], if H3 sends inter-subnet traffic to H1, a routing lookup will happen at PE3's IP-VRF and this routing lookup will not yield the destination IRB interface and therefore MAC Aliasing is not possible. In order to have per-flow load balancing for H3's routed traffic to H1, an IP ECMP list (to PE1/PE2) needs to be associated to H1's host route in the IP-VRF route-table. If H1 is locally learned only at one of the multi-homing PEs, PE1 or PE2, due to LAG hashing, PE3 will not be able to build an IP ECMP list for the H1 host route.

With the extension described in this document, PE3's IP-VRF becomes Ethernet-Segment-aware and builds an IP ECMP list for H1 based on the advertisement of ES1 along with H1 in a MAC/IP route and the availability of ES1 on PE1 and PE2.

1.2. Inter-subnet Forwarding for Prefix Routes in the Interface-less IP-VRF-to-IP-VRF Model

In the Interface-less IP-VRF-to-IP-VRF model described in [I-D.ietf-bess-evpn-prefix-advertisement] there is no Overlay Index and hence no recursive resolution of the IP Prefix route to either a MAC/IP Advertisement or an Ethernet A-D per ES/EVI route, which means that the fast convergence and aliasing/backup path functions are disabled. Although the use-case is different, in a sense the recursive resolution of an IP Prefix route to an Ethernet A-D per ES/EVI route is already described in section 4.3 of [I-D.ietf-bess-evpn-prefix-advertisement], Bump-in-the-Wire Use-Case, but that section does not describe aliasing.

Although this document can be considered to be adding the aliasing/backup path function to the Bump-in-the-Wire Use-Case, the scenario illustrated in Figure 2 will be used to explain the procedures.

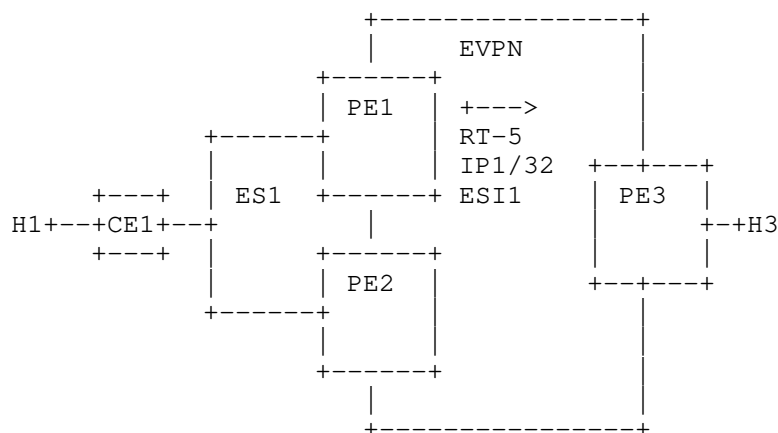


Figure 2: Inter-subnet example with IP Prefix routes

Consider PE1 and PE2 are multi-homed to CE1 (in an All-Active Ethernet Segment ES1), and PE1, PE2 and PE3 are attached to an IP-VRF of the same tenant. Suppose H1's host route is learned (via ARP or ND snooping) on PE1 only, and PE1 advertises an EVPN IP Prefix route for H1's host route. If H3 sends inter-subnet traffic to H1, a routing lookup on PE3 would normally yield a single next-hop, i.e., PE1.

This document proposes the use of the ESI in the IP Prefix route and the recursive resolution to A-D per ES/EVI routes advertised from PE1 and PE2, so that H1's host route in PE3 can be associated to an IP ECMP list (to PE1/PE2) for aliasing purposes.

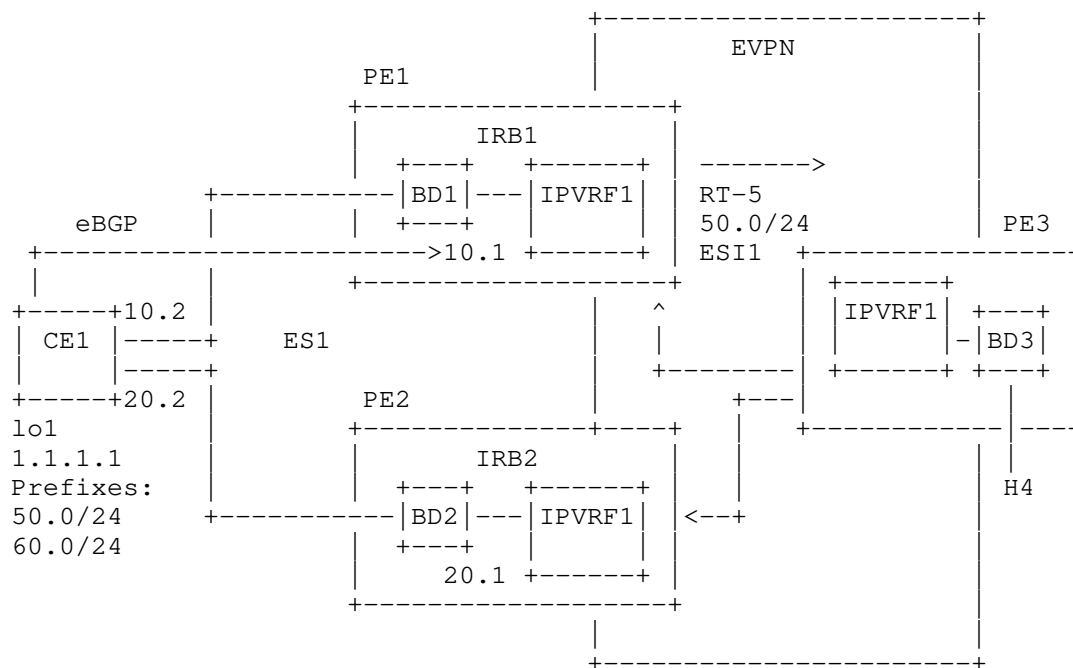
1.3. Ethernet Segments for Prefix routes in IP-VRF-to-IP-VRF use-cases

This document also enables fast convergence and aliasing/backup path to be used even when the ESI is used exclusively as an L3 construct.

As an example, consider the scenario in Figure 3 in which PE1 and PE2 are multi-homed to CE1. However, and contrary to CE1 in Figure 2, in this case the links between CE1 and PE1/PE2 are used exclusively for L3 protocols and L3 forwarding in different BDs, and a BGP session established between CE1's loopback address and PE1's IRB address.

In these use-cases, sometimes the CE supports a single BGP session to one of the PEs (through which it advertises a number of IP Prefixes seating behind itself) and yet, it is desired that remote PEs can build an IP ECMP list or backup IP list including all the PEs multi-homed to the same CE. For example, in Figure 3, CE1 has a single

eBGP neighbor, i.e., PE1. Load-balancing for traffic from CE1 to H4 can be accomplished by a default route with next-hops PE1 and PE2, however, load-balancing from H4 to any of the prefixes attached to CE1 would not be possible since only PE1 would advertise EVPN IP Prefix routes for CE1's prefixes. This document provides a solution so that PE3 considers PE2 as a next-hop in the IP ECMP list for CE1's prefixes, even if PE2 did not advertise the IP Prefix routes for those prefixes in the first place.



Note:

IP addresses expanded by adding 0s
E.g., 50.0 expands to 50.0.0.0

Figure 3: Layer-3 Multihoming PEs

1.4. Terminology and Conventions

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.

- IRB: Integrated Routing and Bridging

- IRB Interface: Integrated Bridging and Routing Interface. A virtual interface that connects the Bridge Table and the IP-VRF on an NVE.
- BD: Broadcast Domain. An EVI may be comprised of one BD (VLAN-based or VLAN Bundle services) or multiple BDs (VLAN-aware Bundle services).
- Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.
- CE: Customer Edge device, e.g., a host, router, or switch.
- EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.
- MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.
- Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.
- Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.
- IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by any routing protocol, E.g., EVPN, IP-VPN and BGP PE-CE IP address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.
- EVPN IP route: An EVPN IP Prefix route or an EVPN MAC/IP Advertisement route.
- LACP: Link Aggregation Control Protocol.
- PE: Provider Edge device.
- Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.
- All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

- RT-2: EVPN MAC/IP Advertisement route, as specified in [RFC7432].
- RT-4: EVPN Ethernet Segment route, as specified in [RFC7432].
- RT-5: EVPN IP Prefix route, as specified in [I-D.ietf-bess-evpn-prefix-advertisement].

2. Ethernet Segments for L3 Aliasing/Backup Path and Fast Convergence

The first two use cases described in Section 1 do not require any extensions to the Ethernet Segment definition and both cases support Ethernet Segments as a set of Ethernet links and specified in [RFC7432], or virtual Ethernet Segments as a set of logical links specified in [I-D.ietf-bess-evpn-virtual-eth-segment].

The third use case in Section 1 requires an extension to the way Ethernet Segments are defined and associated. In this case, the Ethernet Segment is a Layer-3 construct characterized as follows:

- The ES is defined as a set of Layer-3 links to the multi-homed CE and its state MUST be linked to the layer-3 reachability from each multi-homed PE to the CE's loopback address via a non-EVPN route in the PE's IP-VRF.
- The ESI SHOULD be of type 4 [RFC7432] and set to the router ID of the multi-homed CE.
- All-active or single-active multi-homing redundancy modes are supported, however, the redundancy mode only affects the procedures in Section 3.
- PEs attached to the same Layer-3 ES discover each other through the exchange of RT4 routes (Ethernet Segment routes). DF Election procedures [RFC8584] MAY be used for single-active multi-homing mode.
- The routes advertised from the multi-homed CE's and installed in the PE's IP-VRF table with the CE's loopback as the next-hop SHOULD be re-advertised by the PE in EVPN IP Prefix routes with the ESI of the CE. The rest of the EVPN IP Prefix routes fields are set as per the Interface-less model in [I-D.ietf-bess-evpn-prefix-advertisement].

In the example depicted in Figure 3, ES1 is defined as the set of layer-3 links that connects PE1 and PE2 to CE1. Its ESI, e.g., ESI-1, is derived as a type 4 ESI using the CE's router ID. ES-1 will be operationally up in the PE as long as CE1's loopback route is installed in the PE's IP-VRF and learned via any routing protocol

except for an EVPN route. E.g., an active static route to 1.1.1.1 via next-hop 10.0.0.2 would make the ES operationally up in PE1, and the eBGP routes received from CE1 with next-hop 1.1.1.1 will be re-advertised as RT-5 routes with ESI-1.

3. IP Aliasing and Backup Path

In order to address the use-cases described in Section 1, above, this document proposes that:

1. A PE that is attached to a given ES will advertise a set of one or more Ethernet A-D per ES routes for that ES. Each is termed an 'IP A-D per ES' route and is tagged with the route targets (RTs) for one or more of the IP-VRFs defined on it for that ES; the complete set of IP A-D per ES routes contains the RTs for all of the IP-VRFs defined on it for that ES.

A remote PE imports an IP A-D per ES route into the IP-VRFs corresponding to the RTs with which the route is tagged. When the complete set of IP A-D per ES routes has been processed, a remote PE will have imported an IP A-D per ES route into each of the IP-VRFs defined on it for that ES; this enables fast convergence for each of these IP-VRFs.

2. A PE advertises for this ES, an Ethernet A-D Per EVI route for each of the IP-VRFs defined on it. Each is termed an 'IP A-D per EVI' route and is tagged with the RT for a given IP-VRF.

A remote PE imports an IP A-D per EVI route into the IP-VRF corresponding to the RT with which the route is tagged. The label contained in the route enables aliasing/backup path for the routes in that IP-VRF.

To address the third use-case described in Section 1, where the links between a CE and its multihomed PEs are used exclusively for L3 protocols and L3 forwarding, a PE uses the procedures described in 1) and 2), above.

The processing of the IP A-D per ES and the IP A-D per EVI routes is as defined in [RFC7432] and [RFC8365] except that the fast convergence and aliasing/backup path functions apply to the routes contained in an IP-VRF. In particular, a remote PE that receives an EVPN MAC/IP Advertisement route or an IP Prefix route with a non-reserved ESI and the RT of a particular IP-VRF SHOULD consider it reachable by every PE that has advertised an IP A-D per ES and IP A-D per EVI route for that ESI and IP-VRF.

3.1. Constructing the IP A-D per EVI Route

The construction of the IP A-D per EVI route is the same as that of the Ethernet A-D per EVI route, as described in [RFC7432], with the following exceptions:

- The Route-Distinguisher is for the corresponding IP-VRF.
- The Ethernet Tag should be set to 0.
- The route SHOULD carry the RT of the corresponding IP-VRF.
- The route MUST carry the PE's MAC Extended Community if the encapsulation used between the PEs for inter-subnet forwarding is an Ethernet NVO tunnel [I-D.ietf-bess-evpn-prefix-advertisement].
- The route SHOULD carry the Layer 2 Extended Community [RFC8214]. For all-active multihoming, all PEs attached to the specified ES will advertise P=1. For backup path, the Primary PE will advertise P=1 and the Backup PE will advertise P=0, B=1.
 - o The Primary PE SHOULD be a PE with a routing adjacency to the attached CE.
 - o The Primary PE MAY be determined by policy or MAY be elected by a DF Election as in [RFC8584] as described in Section 2.

4. Fast Convergence for Routed Traffic

Host or Prefix reachability is learned via the BGP-EVPN control plane over the MPLS/NVO network. EVPN IP routes for a given ES are advertised by one or more of the PEs attached to that ES. When one of these PEs fails, a remote PE needs to quickly invalidate the EVPN IP routes received from it.

To accomplish this, EVPN defined the fast convergence function specified in [RFC7432]. This document extends fast convergence to inter-subnet forwarding by having each PE advertise a set of one or more IP A-D per ES routes for each locally attached Ethernet segment (refer to Section 4.1 below for details on how these routes are constructed). A PE may need to advertise more than one IP A-D per ES route for a given ES because the ES may be in a multiplicity of IP-VRFs and the Route-Targets for all of these IP-VRFs may not fit into a single route. Advertising a set of IP A-D per ES routes for the ES allows each route to contain a subset of the complete set of Route Targets. Each IP A-D per ES route is differentiated from the other routes in the set by a different Route Distinguisher (RD).

Upon failure in connectivity to the attached ES, the PE withdraws the corresponding set of IP A-D per ES routes. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for all IP addresses associated with the Ethernet Segment in question, across IP-VRFs. If no other PE has advertised an IP A-D per ES route for the same Ethernet Segment, then the PE that received the withdrawal simply invalidates the IP entries for that segment. Otherwise, the PE updates its next-hop adjacencies accordingly.

These routes should be processed with higher priority than EVPN IP route withdrawals upon failure. Similar priority processing is needed even on the intermediate Route Reflectors.

4.1. Constructing IP A-D per Ethernet Segment Route

This section describes the procedures used to construct the IP A-D per ES route, which is used for fast convergence (as discussed in Section 3). The usage/construction of this route remains similar to that described in section 8.2.1. of [RFC7432] with a few notable exceptions as explained in following sections.

4.1.1. IP A-D per ES Route Targets

Each IP A-D per ES route MUST carry one or more Route Targets (RTs). The set of IP A-D per ES routes MUST carry the entire set of IP-VRF RTs for all the IP-VRFs defined on that ES.

4.2. Avoiding convergence issues by synchronizing IP prefixes

Consider a pair of multi-homing PEs, PE1 and PE2. Let there be a host H1 attached to them. Consider PE3 and a host H3 attached to it.

If the host H1 is learned on both the PEs, the ECMP path list is formed on PE3 pointing to (PE1/PE2). Traffic from H3 to H1 is not impacted even if one of the PEs fails as the path list gets corrected upon receiving the withdrawal of the fast convergence route(s) (IP A-D per ES routes).

In a case where H1 is locally learned only on PE1 due to LAG hashing or a single routing protocol adjacency to PE1, at PE3, H1 has ECMP path list (PE1/PE2) using Aliasing as described in this document. Traffic from H3 can reach H1 via either PE1 or PE2.

PE2 should install local forwarding state for EVPN IP routes advertised by other PEs attached to the same ES (i.e., PE1) but not advertise them as local routes. When the traffic from H3 reaches PE2, PE2 will be able forward the traffic to H1 without any convergence delay (caused by triggering ARP/ND to H1 or to the next-

hop to reach H1). The synchronization of the EVPN IP routes across all PEs of the same Ethernet Segment is important to solve convergence issues.

4.3. Handling Silent Host MAC/IP route for IP Aliasing

Consider the example of Figure 1 for IP aliasing. If PE1 fails, PE3 will receive the withdrawal of the fast convergence route(s) and update the ECMP list for H1 to be just PE2. When the EVPN IP route for H1 is also withdrawn, neither PE2 nor PE3 will have a route to H1, and traffic from H3 to H1 is blackholed until PE2 learns H1 and advertises an EVPN IP route for it.

This blackholing can be much worse if the H1 behaves like a silent host. IP address of H1 will not be re-learned on PE2 till H1 ARP/ND messages or some traffic triggers ARP/ND for H1.

PE2 can detect the failure of PE1's reachability in different ways:

- a. When PE1 fails, the next hop tracking to PE1 in the underlay routing protocols can help detect the failure.
- b. Upon the failure of its link to CE1, PE1 will withdraw its IP A-D route(s) and PE2 can use this as a trigger to detect failure.

Thus to avoid blackholing, when PE2 detects loss of reachability to PE1, it should trigger ARP/ND requests for all remote IP prefixes received from PE1 across all affected IP-VRFs. This will force host H1 to reply to the solicited ARP/ND messages from PE2 and refresh both MAC and IP for the corresponding host in its tables.

Even in core failure scenario on PE1, PE1 must withdraw all its local layer-2 connectivity, as Layer-2 traffic should not be received by PE1. So when ARP/ND is triggered from PE2 the replies from host H1 can only be received by PE2. Thus H1 will be learned as local route and also advertised from PE2.

It is recommended to have a staggered or delayed deletion of the EVPN IP routes from PE1, so that ARP/ND refresh can happen on PE2 before the deletion.

4.4. MAC Aging

In the same example as in Section 4.3, PE1 would do ARP/ND refresh for H1 before it ages out. During this process, H1 can age out genuinely or due to the ARP/ND reply landing on PE2. PE1 must withdraw the local entry from BGP when H1 entry ages out. PE1

deletes the entry from the local forwarding only when there are no remote synced entries.

5. Determining Reachability to Unicast IP Addresses

5.1. Local Learning

The procedures for local learning do not change from [RFC7432] or [I-D.ietf-bess-evpn-prefix-advertisement].

5.2. Remote Learning

The procedures for remote learning do not change from [RFC7432] or [I-D.ietf-bess-evpn-prefix-advertisement].

5.3. Constructing the EVPN IP Routes

The procedures for constructing MAC/IP Address or IP Prefix Advertisements do not change from [RFC7432] or [I-D.ietf-bess-evpn-prefix-advertisement].

5.3.1. Route Resolution

If the ESI field is set to reserved values of 0 or MAX-ESI, the EVPN IP route resolution MUST be based on the EVPN IP route alone.

If the ESI field is set to a non-reserved ESI, the EVPN IP route resolution MUST happen only when both the EVPN IP route and the associated set of IP A-D per ES routes have been received. To illustrate this with an example, consider a pair of multi-homed PEs, PE1 and PE2, connected to an all-active Ethernet Segment. A given host with IP address H1 is learned by PE1 but not by PE2. When the EVPN IP route from PE1 and a set of IP A-D per ES and IP A-D per EVI routes from PE1 and PE2 are received, then (1) PE3 can forward traffic destined to H1 to both PE1 and PE2.

If after (1) PE1 withdraws the IP A-D per ES route, then PE3 will forward the traffic to PE2 only.

If after (1) PE2 withdraws the IP A-D per ES route, then PE3 will forward the traffic to PE1 only.

If after (1) PE1 withdraws the EVPN IP route, then PE3 will do delayed deletion of H1, as described in Section 4.3.

If after (1) PE2 advertised the EVPN IP route, but PE1 withdraws it, PE3 will continue forwarding to both PE1 and PE2 as long as it has the IP A-D per ES and the IP A-D per EVI route from both.

6. Forwarding Unicast Packets

Refer to Section 5 in [I-D.ietf-bess-evpn-inter-subnet-forwarding] and [I-D.ietf-bess-evpn-prefix-advertisement].

7. Load Balancing of Unicast Packets

The procedures for load balancing of Unicast Packets do not change from [RFC7432]

8. IP Aliasing and Unequal ECMP for IP Prefix Routes

[I-D.ietf-bess-evpn-unequal-lb] specifies the use of the EVPN Link bandwidth extended community to achieve weighted load balancing to an ES or Virtual ES for unicast traffic. The procedures in [I-D.ietf-bess-evpn-unequal-lb] MAY be used along with the procedures described in this document for any of the three cases described in Section 1, with the following considerations:

- The ES weight is signaled by the multi-homed PEs in the IP A-D per ES routes.
- The remote ingress PE learning an EVPN IP Route to prefix/host P that is associated to a weighted load balancing ES, will follow the procedures in [I-D.ietf-bess-evpn-unequal-lb] to influence the load balancing for traffic to P.
- [I-D.ietf-bess-evpn-unequal-lb] also allows the use of the Link Bandwidth Extended Community along with RT5s. If the ingress PE learns a prefix P via a non-reserved ESI RT5 route with a weight (for which IP A-D per ES routes also signal a weight) and a zero ESI RT5 that includes a weight, the ingress PE will consider all the PEs attached to the ES as a single PE when normalizing weights.

As an example, consider PE1 and PE2 are attached to ES-1 and PE1 advertises an RT-5 for prefix P with ESI-1 (and link bandwidth of 1). Consider PE3 advertises an RT-5 for P with ESI=0 and link bandwidth of 2. If PE1 and PE2 advertise a link bandwidth of 1 and 2, respectively, in the IP A-D per ES routes for ES-1, an ingress PE4 SHOULD assign a normalized weight of 1 to ES-1 and a normalized weight of 2 to PE3. When PE4 sprays the flows to P, it will send twice as many flows to PE3. For the flows sent to ES-1, the individual PE link bandwidths advertised in the IP A-D per ES routes will be considered.

9. Security Considerations

The mechanisms in this document use EVPN control plane as defined in [RFC7432]. Security considerations described in [RFC7432] are equally applicable. This document uses MPLS and IP-based tunnel technologies to support data plane transport. Security considerations described in [RFC7432] and in [RFC8365] are equally applicable.

10. IANA Considerations

No IANA considerations.

11. Contributors

12. Acknowledgments

13. References

13.1. Normative References

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <<https://www.rfc-editor.org/info/rfc8214>>.
- [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>>.
- [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>>.

[RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

13.2. Informative References

[I-D.ietf-bess-evpn-inter-subnet-forwarding]
Sajassi, A., Salam, S., Thoria, S., Drake, J. E., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", draft-ietf-bess-evpn-inter-subnet-forwarding-15 (work in progress), July 2021.

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J. E., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

[I-D.ietf-bess-evpn-virtual-eth-segment]
Sajassi, A., Brissette, P., Schell, R., Drake, J. E., and J. Rabadan, "EVPN Virtual Ethernet Segment", draft-ietf-bess-evpn-virtual-eth-segment-07 (work in progress), July 2021.

[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 Multi-Homing", draft-ietf-bess-evpn-unequal-lb-14 (work in progress), May 2021.

Authors' Addresses

A. Sajassi (editor)
Cisco Systems

Email: sajassi@cisco.com

G. Badoni
Cisco Systems

Email: gbadoni@cisco.com

P. Warade
Cisco Systems

Email: pwarade@cisco.com

S. Pasupula
Cisco Systems

Email: surpasup@cisco.com

J. Drake (editor)
Juniper

Email: jdrake@juniper.net

J. Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2022

A. Sajassi, Ed.
A. Banerjee
S. Thoria
Cisco
D. Carrel
Graphiant
B. Weis
Independent
J. Drake
Juniper Networks
October 25, 2021

Secure EVPN
draft-sajassi-bess-secure-evpn-05

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 in full conformance with the provisions of BCP 78 and BCP 79.

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

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

This Internet-Draft will expire on April 28, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Requirements Language	5
2. Terminology	5
3. Requirements	7
3.1. Tenant's Layer-2 and Layer-3 data and control traffic . .	7
3.2. Tenant's Unicast and Multicast Data Protection	7
3.3. P2MP Signaling for SA setup and Maintenance	7
3.4. Granularity of Security Association Tunnels	7
3.5. Support for Policy and DH-Group List	8
4. SA and Key Management	8
4.1. Generating Initial IPsec SAs	8
4.2. Rekey of IPsec SAs	10
4.2.1. Single IPsec Device Rekey	11
4.2.2. Multiple IPsec Device Rekey	13
5. IPsec Database Generation	16
5.1. The Security Policy Database (SPD)	16
5.2. Security Association Database (SAD)	16
5.2.1. Generating Keying Material for IPsec SAs	16
5.2.1.1. g ^{ir}	16
5.2.1.2. Nonces	17
5.2.1.3. SPIs	17
5.2.1.4. IPsec key generation	18
5.3. Peer Authorization Database (PAD)	19
6. Policy distributed through the BGP RR	19
6.1. IPsec policy negotiation	20
7. BGP Component	21
7.1. Zero Touch Bring-up (ZTB)	21
7.2. Configuration Management	21
7.3. Orchestration	22

7.4. Signaling	22
8. Solution Description	22
8.1. Inheritance of Security Policies	23
8.2. Distribution of Public Keys and Policies	24
8.2.1. Minimal DIM	24
8.2.2. Multiple Policies	25
8.2.3. Multiple DH-groups	25
8.2.4. Multiple or Single ESP SA policies	25
8.3. Initial IPsec SAs Generation	25
8.4. Re-Keying	26
8.5. IPsec Databases	26
9. Encapsulation	26
9.1. Standard ESP Encapsulation	26
9.2. ESP Encapsulation within UDP packet	27
10. BGP Encoding	28
10.1. The Base (Minimal Set) DIM Sub-TLV	29
10.2. The Key Exchange Sub-TLV	29
10.3. ESP SA Proposals Sub-TLV	30
10.3.1. Transform Substructure	30
11. Applicability	31
12. Acknowledgements	32
13. IANA Considerations	32
14. Security Considerations	32
15. References	33
15.1. Normative References	33
15.2. Informative References	34
Appendix A. Additional Stuff	35
Authors' Addresses	35

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. This method is specially recommended for large scale deployment where large meshes of IKEv2 sessions among PE devices are not appropriate.

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 information. 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 singling 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.

Many key exchange methods (such as IKEv2) use a Diffie-Hellman (DH) algorithm to derive keys. When combined with an authentication method, the key exchange method allows two network devices to generate private pair-wise keys with each other. This document presents a key exchange method making use of the PE-to-RR trust model, where an RR is used to distribute keying material and policy between PE devices, also resulting in the PEs generating private pair-wise keys with each other. DH public values are provided to controllers from IPsec devices, where the controller relays the DH public values to authorized peers of that IPsec device as defined by a centralized policy. PE devices then create and install private pair-wise IPsec session keys to be used to secure communications with their peers.

Although IKEv2 is not used in this approach, the key management interfaces between IKEv2 and IPsec defined in RFC 7296 are maintained as much as possible.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

- o AC: Attachment Circuit.
- o ARP: Address Resolution Protocol.
- o BD: Broadcast Domain. As per RFC7432 [RFC7432], an EVI consists of a single or multiple BDs. In case of VLAN-bundle and VLAN-based service models (see RFC7432 [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.
- o BD Route Target: refers to the Broadcast Domain assigned Route Target RFC4364 [RFC4364]. In case of VLAN-aware bundle service model, all the BD instances in the MAC-VRF share the same Route Target.
- o BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per RFC7432 [RFC7432].
- o DGW: Data Center Gateway.
- o Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].
- o Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload. Examples of this type of tunnels are VXLAN or GENEVE [GENEVE].
- o EVI: EVPN Instance spanning the NVE/PE devices that are participating on that EVPN, as per [RFC7432].
- o EVPN: Ethernet Virtual Private Networks, as per [RFC7432].
- o GRE: Generic Routing Encapsulation.
- o GW IP: Gateway IP Address.
- o IPL: IP Prefix Length.

- o IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload).
- o 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.
- o IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).
- o 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.
- o ML: MAC address length.
- o ND: Neighbor Discovery Protocol.
- o NVE: Network Virtualization Edge.
- o GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].
- o NVO: Network Virtualization Overlays.
- o RT-2: EVPN route type 2, i.e., MAC/IP advertisement route, as defined in [RFC7432].
- o RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [EVPN-PREFIX].
- o 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.).
- o SN: Subnet.
- o TS: Tenant System.
- o VA: Virtual Appliance.
- o 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.

- o VTEP: VXLAN Termination End Point, as in RFC 7348 [RFC7348].
- o VXLAN: Virtual Extensible LAN, as in RFC 7348 [RFC7348].

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

3. Requirements

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

3.1. Tenant's Layer-2 and Layer-3 data and 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.

3.2. Tenant's Unicast and 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.

3.3. P2MP Signaling for SA setup and Maintenance

BGP P2MP signaling must be used for IPsec SAs setup and maintenance. This reduces the number of message exchanges from $O(N^2)$ to $O(N)$ among the participating PE devices.

3.4. Granularity of Security Association Tunnels

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

- o Per PE: A single IPsec tunnel between a pair of PEs to be used for all tenants' traffic supported by the pair of PEs.
- o 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.

- o Per subnet: A single IPsec tunnel per subnet (e.g., per VLAN/EVI) of a tenant on a pair of PEs.
- o Per L3 flow: A single IPsec tunnel per pair of IP addresses of a tenant on a pair of PEs.
- o Per L2 flow: A single IPsec tunnel per pair of MAC addresses of a tenant on a pair of PEs.
- o Per AC pair: A single IPsec tunnel per pair of Attachment Circuits between a pair of PEs.

3.5. 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.

4. SA and Key Management

The BGP Route Reflector (RR) acts as a trusted third party, which relays policy and keying material between PE devices. Communications between the RR and the PEs MUST be authenticated, encrypted, and integrity-protected. All algorithms are selected by the management station associated with the RR. The combination of the RR and a set of PE devices comprises of a cooperating group of devices that make up a VPN, where each PE device is authorized to communicate with other PE devices in the group. Policies can allow a PE device to communicate with all other PE devices in the group, or may restrict it to a subset of those devices.

DH public values from each PE are distributed to other authorized peer PEs via the RR. Each PE device creates and maintains a DH pair, which it uses to communicate with other members of the VPN. This distribution of DH public values (and other related values) is intended to be embedded into the BGP protocol as described later. In particular, the RR provides a mechanism for secure key management. However, it does not provide policy information or configuration as that is assumed to be provided by the management station.

4.1. Generating Initial IPsec SAs

When an PE device (PE) begins operation, it generates a private/public DH pair, using an algorithm defined in the IKEv2 Diffie-Hellman Group Transform IDs [IKEV2-IANA]. If the device does not have any active peers it simply distributes its DH public value to the BGP RR, along with a nonce to be used during SA creation. Whenever a private/public DH pair is created, a new nonce MUST also be created. Whenever DH public values are transmitted, they are

transmitted with the corresponding nonce. Whenever a DH private or DH public value is used, it is used along with the corresponding nonce. However, in the diagrams and descriptions below, the nonces are often left out for the sake of clarity.

Upon receiving a peer's DH public value and nonce, the receiver creates IPsec SAs (as described in Section 5.2). For each peer, a pair of IPsec SAs are created by combining the PE device's own DH private value with the DH public number received from the Controller.

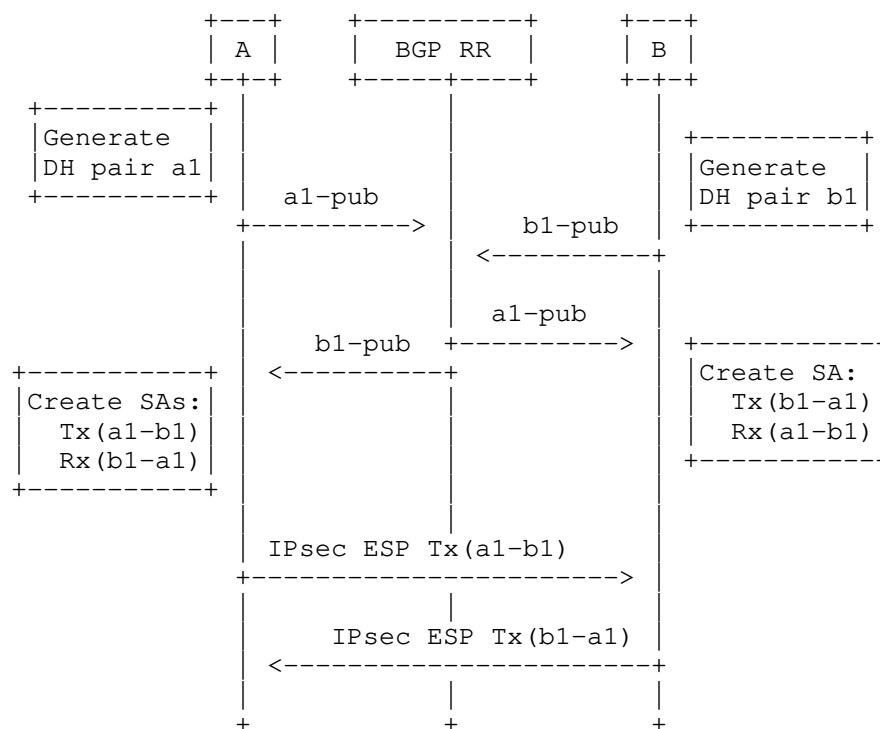


Figure 1: Generation of Initial IPsec SAs between two peers

Figure 1 shows IPsec SA generation between a pair of PE devices. Two PE devices (A and B shown in Figure 1) join the network. Each creates its own DH pair (labelled "a1" on A and "b1" on B), and distributes the DH public value (labelled a1-pub and b1-pub) to the BGP RR. The BGP RR forwards the DH public value to all authorized peers, although for simplicity of exposition the figure only shows the two IPsec devices.

When each device receives the peer's DH public value, a pair of IPsec SAs are generated: one outbound and one inbound. As shown in the

figure, A generates an outbound SA labeled Tx(a1-b1), representing that it has been generated using A's DH pair labeled a1 and B's DH pair labeled b1. B generates the same IPsec SA as an inbound SA, which is labeled Rx(a1-b1). Similarly, A generates an inbound IPsec SA labelled Rx(b1-a1), which is the same IPsec SA on B which is labelled Tx(b1-a1).

This process repeats on both A and B as they discover other PE devices with which they are authorized to communicate.

4.2. Rekey of IPsec SAs

Any IPsec device may initiate a rekey at any time. Common reasons to perform a rekey include a local time or volume based policy, or may be the result of a cipher counter mode Initialization Vector (IV) counter nearing its final value. The rekey process is performed individually for each remote peer. If rekeying is performed with multiple peers simultaneously, then the decision process and rules described in this rekey are performed independently for each peer.

A decision process choosing an outbound IPsec SA is followed when certain events occur, as described in the rules below. The same decision process is followed regardless of whether the device is performing a rekey or responding to a peer's rekey. The decision process is:

1. Determine the outbound SAs with the remote peer's most recently distributed DH public value.
2. Determine which of those outbound SAs are "live". A "live" outbound SA is one built from a DH value from the local peer for which it has observed inbound traffic using any SA based on the same local DH pair. This proves that the remote peer is prepared to receive traffic protected by that DH pair.
3. Choose the "live" outbound SA built from the local peer's most recent DH public value.

A rekey operation follows these four basic rules.

Rule 1: When an IPsec device needs to perform a rekey with a remote peer, it creates a new pair of IPsec SAs by combining the new DH private value with the peer's DH public values. If the remote peer is also in the midst of a rollover and its DH public value has already been received, then this may result in creating two sets of SAs: one pair with the remote peer's old DH public value, and one pair with the remote peer's new DH public value.

Rule 2: When an IPsec device receives a new remote peer's DH public value from the controller it creates and installs a new pair of IPsec SAs by combining the remote peer's new DH public value with its own current local DH private values. If both devices are in the midst of a rollover, this may result in creating two sets of SAs with the remote peer's new DH public value: one with the local old DH private value, and one with the local new DH private value. The outbound SA decision process is performed.

Rule 3: The first IPsec packet received by a rekeying IPsec device on an inbound SA using its new DH pair causes it to perform the outbound SA decision process. It may also shorten the lifetime of IPsec SAs using its own old DH pair that are shared with this peer, as they are no longer in use (other than the inbound SA might receive packets in transit).

Rule 4: The first IPsec packet received from a remote rekeying IPsec device using the remote peer's new DH pair allows the IPsec device to shorten the lifetime of IPsec SAs shared with this peer using unused remote DH pairs.

Two examples follow: a single IPsec device performing a rekey with its peers, and two IPsec devices performing a simultaneous rekey. The same rekey operations described above are exhibited in both cases.

4.2.1. Single IPSec Device Rekey

When a single IPsec device begins a rekey, it first generates a new DH pair and generates new IPsec SA pairs for each peer with which it is communicating. It does this by combining the new DH private value with each peer's existing DH public value. Only when the new IPsec SAs have been installed and the device is prepared to receive on those new SAs does it then distribute the new DH public value to the Controller, which forwards the new DH public value to its authorized peers. The rekeying IPsec device continues to transmit on the old SAs for each peer until it observes that peer begin to transmit on the new SAs.

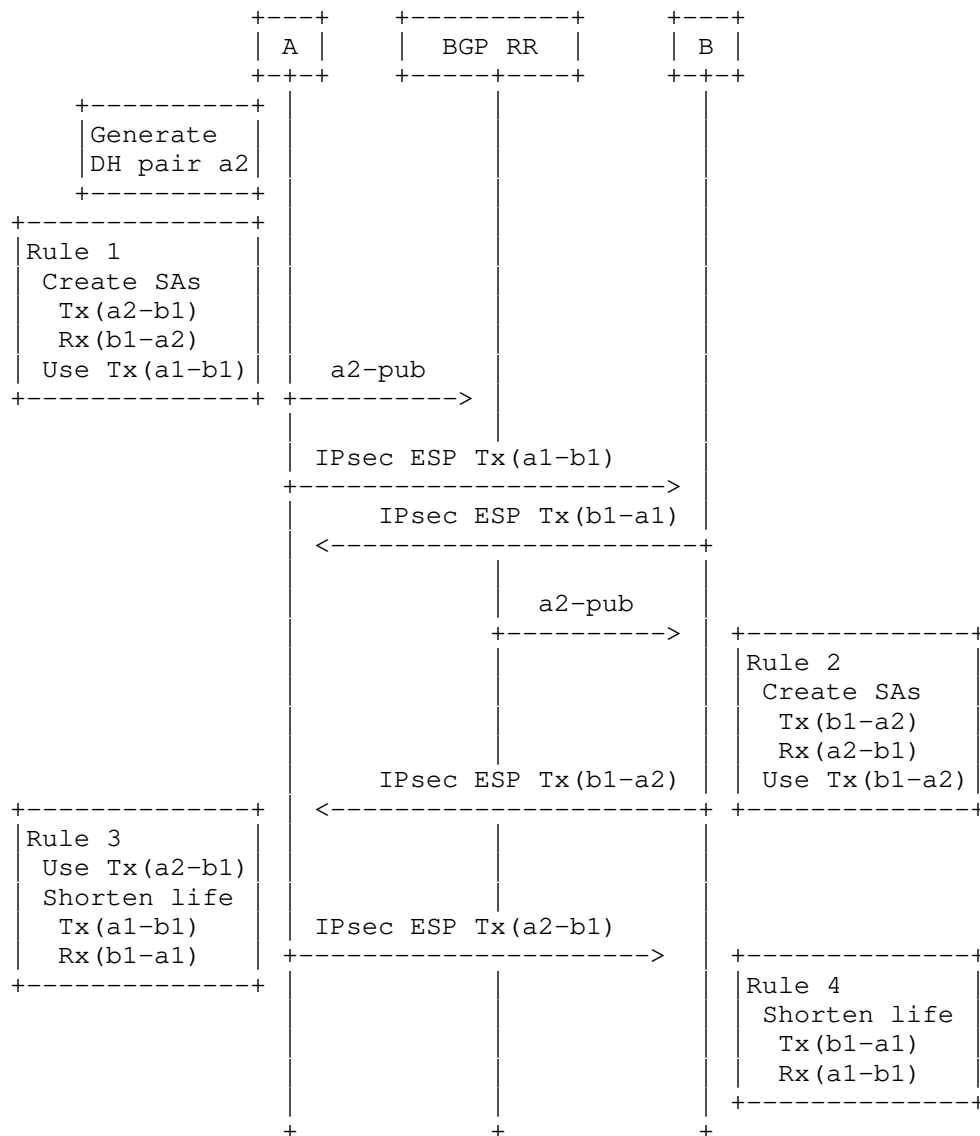


Figure 2: Single IPsec Device Rekey between two peers

In Figure 3, device A is shown as performing a rekey, and it creates a DH pair labelled "a2". The following steps are followed.

1. Rule 1 requires creating new IPsec SAs for each peer. In this example, A creates a new outbound IPsec SA to communicate with B labelled Tx(a2-b1), and a new inbound IPsec SA labelled

Rx(b1-a2). A continues to transmit on Tx(a1-b1) (generated as shown in Figure 2).

2. A distributes the new public value (a2-pub) to the Controller who forwards it to A's authorized peers, which includes B. During this time, both A and B continue to use the initial IPsec SAs setup between them using a1 and b1.
3. When B receives a2 from the controller, B follows Rule 2 by creating Tx(b1-a2), Rx(a2-b1). B also follows the outbound SA decision process, which causes it to change its outbound IPsec SA to A to Tx(b1-a2).
4. When A receives a packet protected by Rx(b1-a2), it follows Rule 3 and performs the outbound SA decision process. This causes it to change its outbound IPsec SA to Use Tx(a2-b1). It also optionally shortens the lifetime of the old IPsec SAs shared with this peer.
5. When B receives a packet protected by Tx(a2-b1), it follows Rule 4, in which it may shorten the lifetime of the old IPsec SAs shared with this peer using DH pairs that are no longer in use.

At the end of the rekey, both A and B retain a single DH pair, and a single set of IPsec SAs between them.

4.2.2. Multiple IPsec Device Rekey

When two or more IPsec device simultaneously begin a rekey, they each follow the rekeying method described in the previous section. Every rekeying IPsec device generates a new DH pair and generates new IPsec SA pairs for each peer with which it is communicating by combining their new DH private value with each peer's existing DH public value. When this completes on a particular IPsec device, it distributes the new DH public value to the Controller, which forwards it to its authorized peers. Each continues to transmit on the existing SAs for each peer until it observes that peer transmitting on the new SAs. During a simultaneous rekey up to four pairs of IPsec SAs may be temporarily created, but the four rules ensure that they converge on a single new set of IPsec SAs.

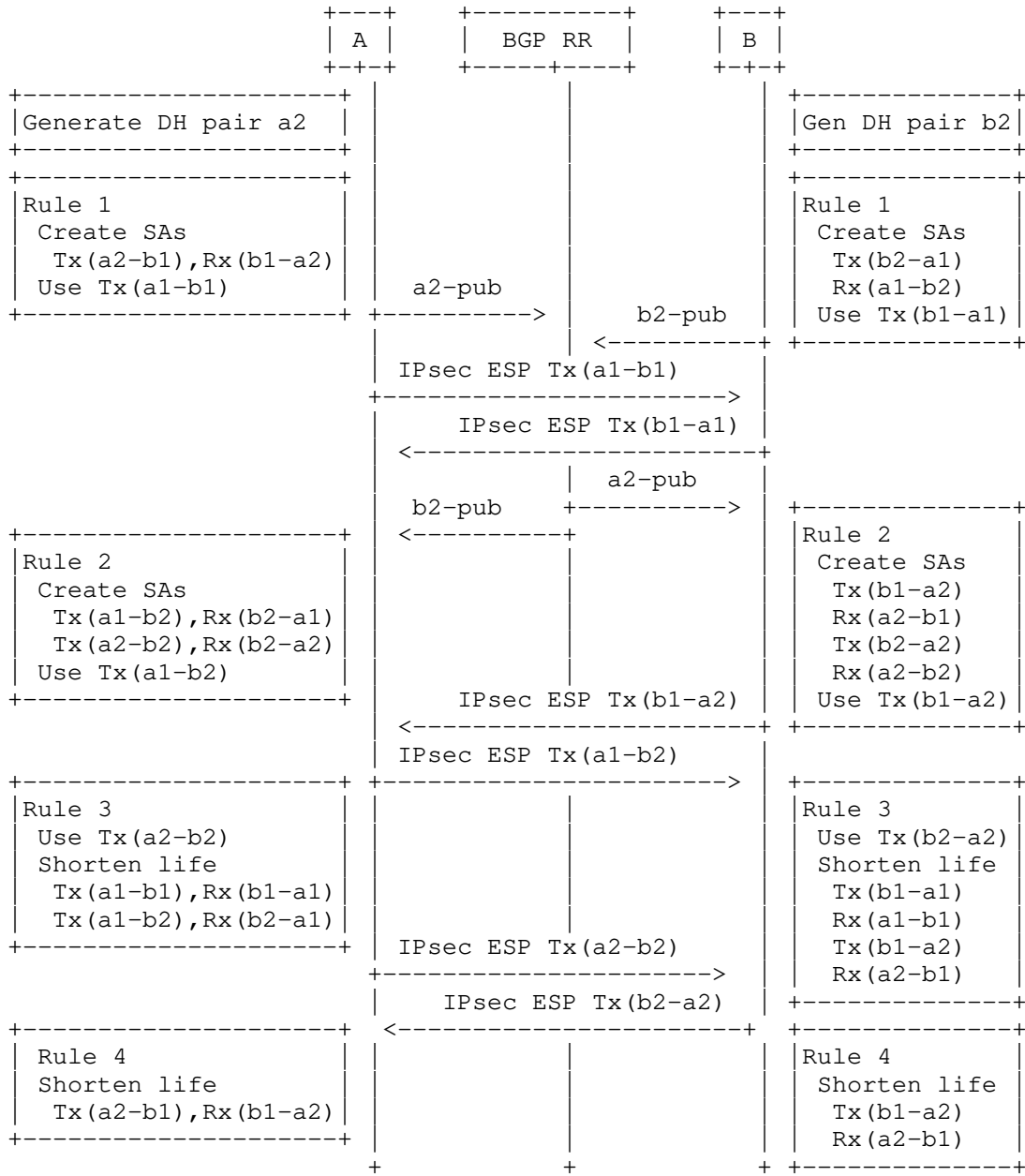


Figure 3: Simultaneous IPsec Device Rekey between two peers

In Figure 4, device A and device B are both shown as performing a rekey. Their initial state corresponds to the final state shown in

Figure 2 (i.e., they are communicating using a single pair of IPsec SAs created from DH pairs "a1" and "b1").

1. A and B follow Rule 1, which includes creating new IPsec SAs for each peer. In this example, A creates a new outbound IPsec SA to communicate with B labelled Tx(a2-b1), and a new inbound IPsec SA labelled Rx(b1-a2). B creates a new outbound IPsec SA to communicate with A labelled Tx(a1-b2), and a new inbound IPsec SA labelled Rx(b2-a1). A and B continue to transmit on IPsec SAs previously created from DH pairs "a1" and "b1".
2. A distributes the new public value (a2-pub) to the Controller who forwards it to A's authorized peers, which includes B. B also distributes the new public value (b2-pub) to the Controller who forwards it to B's authorized peers, which includes A.
3. When A and B receive each other's new peer DH public value from the controller they follow Rule 2. But because now there are four DH values that could be in use between A and B, they must be prepared to use IPsec SAs using each permutation of DH values: a1-b1, a1-b2, a2-b1, a2-b2. Prior to implementing Rule 2, each has already created sets of IPsec SAs matching two of the permutations, so just two more sets must be generated during Rule 2.
 - * One pair is created using the IPsec device's old DH pair with the peer's new DH pair. This is necessary, because the peer may transmit on this pair.
 - * One pair is created using the IPsec device's new DH pair with the peer's new DH pair. This is the set of IPsec SAs that will be used at the end of the rekey process.

Each peer begins transmitting on an IPsec SA that combines the remote peer's new DH pair and its own old DH pair, which is the most recent "live" SA on which it can transmit. I.e., A begins transmitting on Tx(a1-b2) and B begins transmitting on Tx(b1-a2).

4. When A receives a packet protected by Rx(b1-a2), it understands that the remote peer has received its new DH public value. A also understands that because of Rule 2 that B must have created IPsec SAs using a2-b2. This allows A to follow Rule 3 and change its outbound IPsec SA to Use Tx(a2-b2). Similarly, when B receives a packet protected by Rx(a1-b2), B recognizes that it can also begin to transmit using Tx(b2-a2). Note that it is also possible that A will receive a packet protected by Rx(b2-a2) or B will receive a packet protected by Rx(a2-b2), and then knows it can transmit on an IPsec SA using both of the new DH pairs.

5. Also in Rule 3, Both A and B optionally shorten the lifetime of older IPsec SAs shared with this peer derived from unused DH pairs to be cleaned up. A shortens the lifetime of SAs based on a1. B shortens the lifetime of SAs based on b1.
6. When A and B receive a packet protected by the remote peer's latest DH pair, they shortens the lifetime of SAs based on the remote peer's unused DH pair.

5. IPsec Database Generation

The PAD, SPD, and SAD all need to be setup as defined in the IPsec Security Architecture [RFC4301].

5.1. The Security Policy Database (SPD)

The SPD is implemented using methods outside the scope of this document. The SPD describes the type of traffic that will be protected between IPsec devices and the policy (e.g., ciphers) used to create SAs.

5.2. Security Association Database (SAD)

The SAD is constructed from IPsec policy (e.g., ciphers) obtained (depending on the controller protocol method) either from the controller or distributed by a peer (see Section 6).

Keying Material is generated following the method defined in IKEv2, and depends on SPIs, nonces, and the Diffie-Hellman shared secret.

The following sections describe how the necessary values are determined.

5.2.1. Generating Keying Material for IPsec SAs

5.2.1.1. g^{ir}

A DH public value is distributed from the peer.

A DH shared secret (g^{ir}) is computed using the peer's public value, and the device's private value. The DH group to be used must be known by the device. Options include distribution by an SDN controller, or distribution by the peer with the DH public value (see Section 6).

5.2.1.2. Nonces

Nonces are distributed with a DH public value, and are used only with that value. It is RECOMMENDED that nonces are generated as described in Section 2.10 of [RFC7296].

IKEv2 Key derivation specifies an initiator's nonce (N_i) and a responder's nonce (N_r). While neither peer is truly initiating a session, in order to fit the IKE key material models the roles must be assigned. The initiator is chosen as the peer with the larger nonce and the responder is the peer with the smaller. This does mean that the roles can change for each rekey and for each SA within a rekey.

5.2.1.3. SPIs

SPI values that are unique to each generation of keying material need to be determined. While each peer could distribute its own inbound SA value, the SPI value would be used by many peers. Although this is not a problem for an SA lookup (lookup can include the source and destination IP addresses), experience has shown that this is sub-optimal for some hardware SA lookup algorithms. Instead, this specification proposes generating values that are unpredictable and indistinguishable from randomly-generated SPI values.

SPI values are generated using the IKEv2 prf+ function, where nonces are used as the input to the prf. This produces a statistically random SPI value that should be unique. However, with a 32 bit value there is still a very small, but non-zero, chance of SPIs repeating for a given pair of peers. To prevent this and ensure uniqueness in the operational window, we also use the lower 2 bits from each peer's rekey counter.

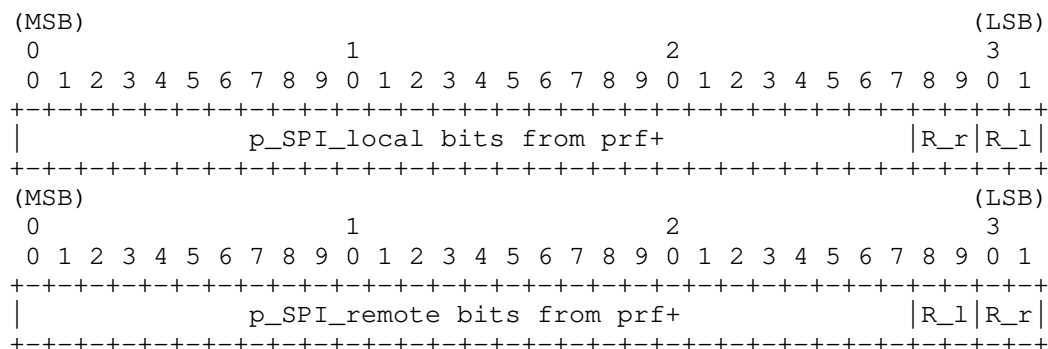
First the SPIs are taken from the prf+ function as 32 bit values and assigned based on which peer is taking the role of initiator and which is taking the role of responder. The p_SPI_i is taken by the device providing N_i , where p_SPI_r is taken by the other device.

$$\{p_SPI_i \mid p_SPI_r\} = \text{prf+}(N_i \mid N_r, \text{"SPI generation"})$$

Next p_SPI_i and p_SPI_r are mapped from initiator and responder roles to local and remote roles based on the choice of N_i and N_r in 5.2.1.2 and are renamed to p_SPI_local and p_SPI_remote .

Then, 2 2-bit Rotation Numbers (RN) are generated from the 2 least significant bits (LSB) of the 2 rekey counter values (see Section 6). These 4 bits replace the least significant bits of p_SPI_local and p_SPI_remote with the local RN bits taking the least significant

position in `p_SPI_local` and the remote RN bits taking the least significant position in `p_SPI_remote`. This shown in the following two diagrams with `RN_local` shown as `R_l` and `RN_remote` shown as `R_r`.



The reason for changing terminology from initiator/responder to local/remote is because the roles of initiator/responder can change in every rekey. The order of `RN_local` and `RN_remote` needs to remain constant. If that order was based on initiator/responder, there's a risk that if the initiator and responder roles changed and the two peers re-keyed on different frequencies, they could end up with identical RN values.

In some circumstances additional values may also need to be added to the `prf` for peers to ensure that they have implemented the same policy. Appendix A.3.1 includes a discussion of when this might be needed. In these cases, only the `prf+` inputs are modified and the Rotation Numbers MUST still be added as above.

Because a device is not choosing its inbound SPI, its SA lookup process needs to be aware that duplicates could occur across different peers. In that case, the inbound SA Lookup SHOULD include a source IP address in addition to the SPI value (see Section 4.1 of [RFC4301]).

5.2.1.4. IPsec key generation

As described in previous sections, a DH public value and a nonce are distributed by peers. These are used to generate IPsec keys following the method defined in the IKEv2. SKEYSEED is generated following Section 2.14 of [RFC7296]:

$$\text{SKEYSEED} = \text{prf}(\text{Ni} \parallel \text{Nr}, g^{\text{air}})$$

KEYMAT can be similarly derived as defined by IKEv2 (Section 2.17 of [RFC7296]), although only SK_d is required to be generated (shown in Section 2.14 of [RFC7296]).

$$\text{SK_d} = \text{prf+}(\text{SKEYSEED}, \text{Ni} \mid \text{Nr} \mid \text{SPIi} \mid \text{SPIr})$$
$$\text{KEYMAT} = \text{prf+}(\text{SK_d}, \text{Ni} \mid \text{Nr})$$

However, with the simplification where only SK_d is generated, it can be observed that the derivation of SK_d could be skipped entirely, and an optimized derivation of KEYMAT could be as follows:

$$\text{KEYMAT} = \text{prf+}(\text{SKEYSEED}, \text{Ni} \mid \text{Nr} \mid \text{SPIi} \mid \text{SPIr})$$

Note: A single specification for generating KEYMAT will be determined in a future version of this document.

5.3. Peer Authorization Database (PAD)

The PAD identifies authorized peers. PAD entries are either statically configured, or may be dynamically updated by the controller.

The PAD omits authentication data for each peer, because it has delegated authentication and authorization to the controller.

The controller protocol MUST be able to describe an identity that a receiver can match against its local PAD database, to ensure that the peer is an authorized peer.

6. Policy distributed through the BGP RR

An IPsec device distributes to a controller a DH public value and the associated information and policy needed to create IPsec SAs in a Device Information Message (DIM). The controller then distributes the DIM to all authorized peers of that device. The following data elements MUST be embedded in a DIM message:

- o DH public number (used for key computation)
- o Nonce (used for key computation and SPI generation)
- o Peer identity (used to identify a peer in the PAD)
- o An Indication whether this is the initial distributed policy
- o A rekey counter, which increases for each unique DIM sent

In cases where a single fixed IPsec policy has been pre-distributed, it is not necessary for the peer to send or receive that policy in a DIM. However, in cases where an IPsec device needs to indicate the policy it is willing to use, the following data elements SHOULD be included in a DIM:

- o An IPsec policy or policies
- o A lifetime bounding the use of the DH public number. When this DH public number is used to create an IPsec SA, the shortest lifetime is used as an SA lifetime for the pair of generated IPsec SAs. When the lifetime expires, the local version of the DIM and IPsec SAs generated from it MUST be deleted.

Appendix A suggests different ways that this policy may be included in a controller protocol. This document does not define a normative protocol format, because the DIM very likely needs to be integrated into an existing controller protocol rather than be an independent key management protocol. However, the controller protocol MUST provide a strong authentication between the device and the controller, and integrity of the messages MUST be provided. Confidentiality of the messages SHOULD also be provided. It is important that the controller protocol be protected with algorithms that are at least as strong as the algorithms used to protect the IPsec packets.

6.1. IPsec policy negotiation

In many controller based networks, there is a single IPsec policy used by all devices and there is no need to redistribute or select policy details. However, in some circumstances, there may be a need to have multiple policy options. This could happen when a controller changes the policy and wants to smoothly migrate all devices to the new policy. Or it could happen if a network supports devices with different capabilities. In these cases, devices need to be able to choose the correct policy to use for each other device, and must do this without sending additional messages and without sending individual messages to each peer. When a device supports multiple policies, it MUST include those policies within the DIM. This is done by sending multiple distinct policies, in order of preference, where the first policy is the most preferred. The policy to use is selected by taking the receiver's list of policies (i.e., the list advertised by the device that generates N_r), starting with the first policy, compare against the initiator's (device that generates N_i) list, and choosing the first one found in common. The method conforms to the IKEv2 Cryptographic Algorithm Negotiation described in Section 2.7 of [RFC7296]. (However, see additional discussion when IKEv2 payloads are used in Appendix A.3.1).

If there is no match, this indicates a controller configuration error. These devices MUST NOT establish new SAs until a DIM is received that does produce a match.

When a device supports more than one DH group, then a unique DH public number MUST be specified for each in order of preference. The selection of which DH group to use follows the same logic as Policy selection, using the receiver's list order until a match is found in the initiator's list.

7. BGP Component

The architecture that encompasses device-to-controller trust model, has several components among which is the signaling component. Secure EVPN Signaling, as defined in this document, is the BGP signaling component of the overall Architecture. We will briefly describe this Architecture here to further facilitate understanding how Secure EVPN fits into the overall architecture. The Architecture describes the components needed to create BGP based SD-WANs and how these components work together. Our intention is to list these components here along with their brief description and to describe this Architecture in details in a separate document where to specify the details for other parts of this architecture besides the BGP signaling component which is described in this document.

The Architecture consists of four components. These components are Zero Touch Bring-up, Configuration Management, Orchestration, and Signaling. In addition to these components, secure communications must be provided between the edge nodes and all servers/devices providing the architecture components.

7.1. Zero Touch Bring-up (ZTB)

The first component is a zero touch capability that allows an edge device to find and join its SD-WAN with little to no assistance other than power and network connectivity. The goal is to use existing work in this area. The requirements are that an edge device can locate its ZTB server/component of its SD-WAN controller in a secure manner and to proceed to receive its configuration.

7.2. Configuration Management

After an edge device joins its SD-WAN, it needs to be configured. Configuration covers all device configuration, not just the configuration related to Secure EVPN. The previous Zero Touch Bring-up component will have directed the edge device, either directly or indirectly, to its configuration server/component. One example of a configuration server is the I2NSF Controller. After a device has

been configured, it can engage in the next two components. Configuration may include updates over time and is not a one time only component.

7.3. Orchestration

This component is optional. It allows for more dynamic updates of configuration and statistics information. Orchestration can be more dynamic than configuration.

7.4. Signaling

Signaling is the component described in this document. The functionality of a Route Reflector is well understood. Here we describe the signaling component of BGP SD-WAN Architecture and the BGP extension/signaling for IPsec key management and policy.

8. 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 section 4 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 10 and are based on the DIM message components in section 5 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.

8.1. 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.

8.2. 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.

8.2.1. Minimal DIM

For SA establishment, at the minimum, a PE needs to advertise to other PEs, its DIM values as specified in section 5. 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.

8.2.2. 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. 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.

8.2.3. 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.

8.2.4. 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.

8.3. Initial IPsec SAs Generation

The procedure for generation of initial IPsec SAs is described in section 4. 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 4.1.

8.4. 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 4.2 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.

8.5. 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 RFC 4301 [RFC4301]. Section 5 of this document gives a summary description of how these databases are setup where key is exchanged via P2MP signaling through 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).

9. 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.

9.1. 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 and encrypted using ESP-Transport mode.

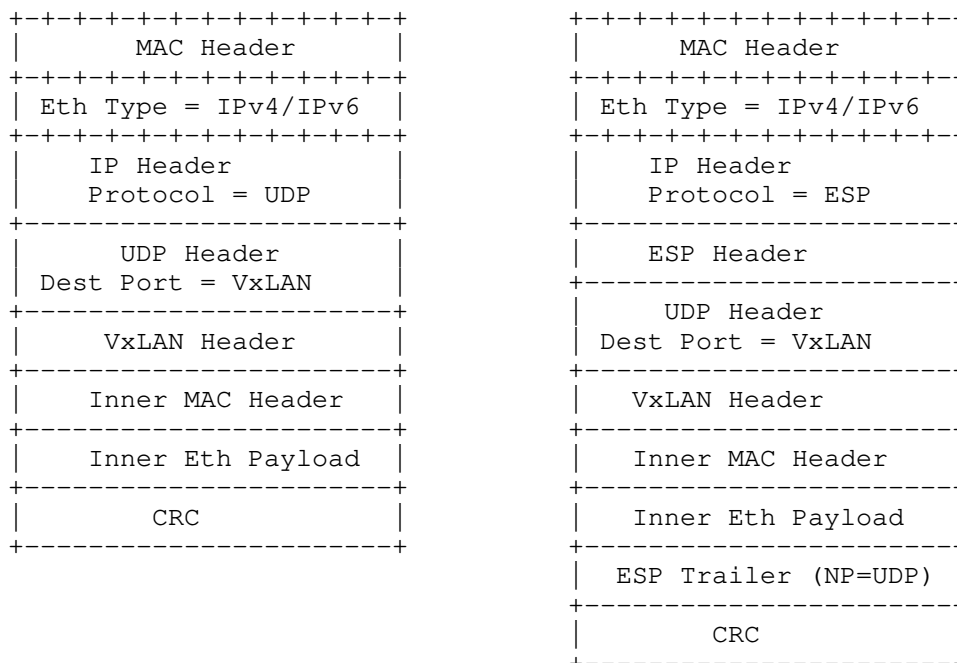


Figure 4

9.2. ESP Encapsulation within UDP packet

In scenarios where NAT traversal is required (RFC 3948 [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 and encrypted using ESP-in-UDP with Transport mode.

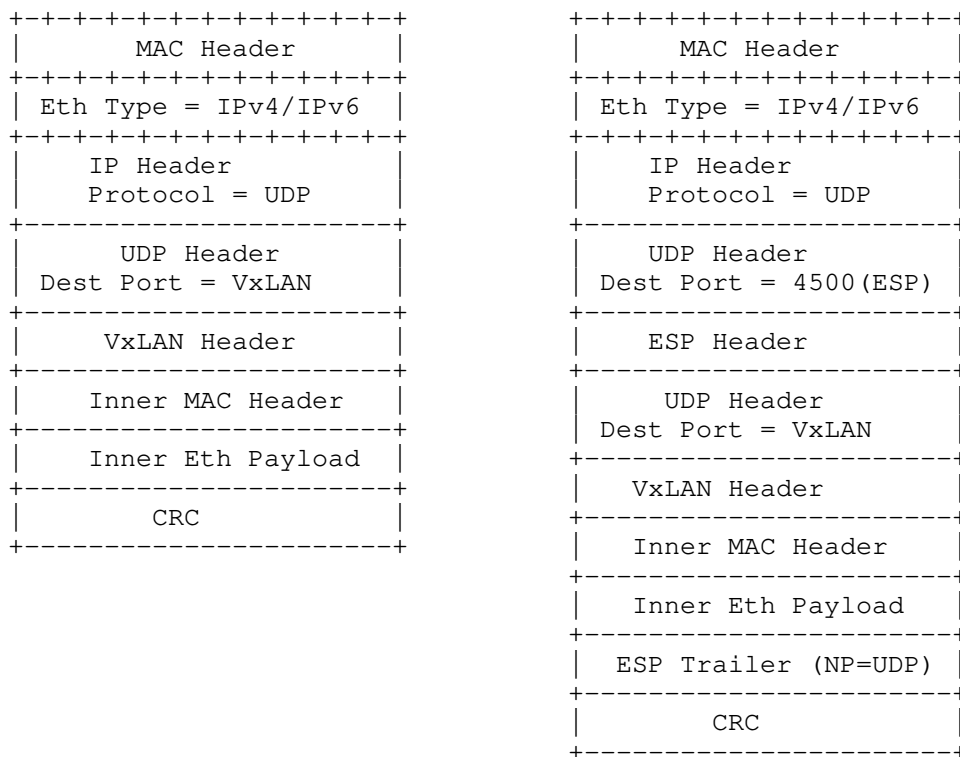


Figure 5

10. 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.

10.1. 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.

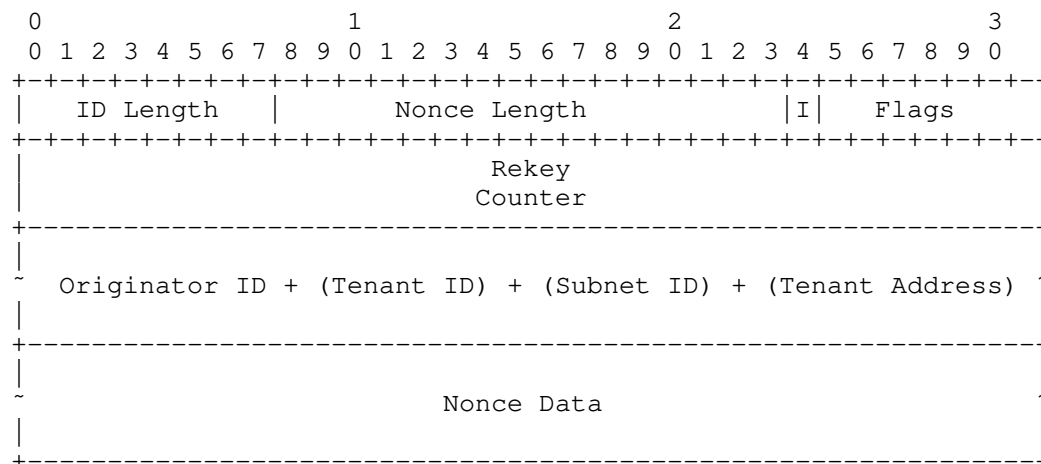


Figure 6

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 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 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. Its length is a multiple of 32 bits. Nonce lengths should be chosen to meet minimum requirements described in IKEv2 [RFC7296].

10.2. The 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.

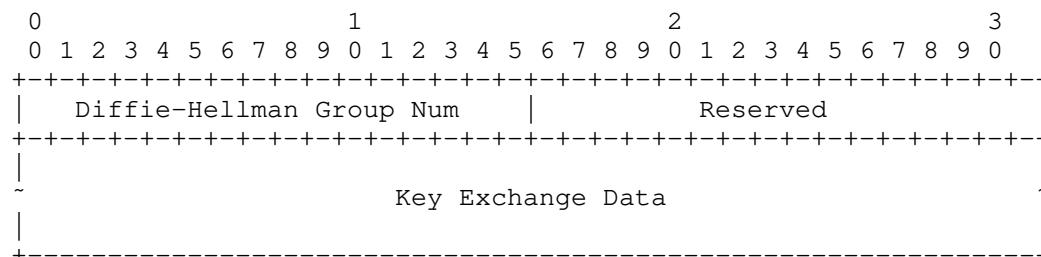


Figure 7

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 MOPT groups in [RFC7296] and for ECP groups in [RFC4753].

10.3. 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.

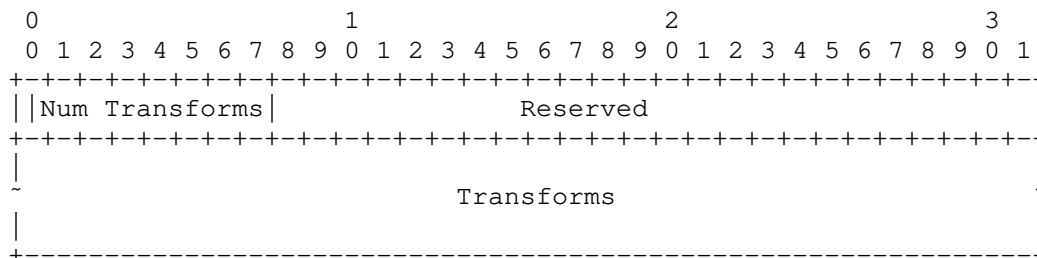


Figure 8

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.

10.3.1. Transform Substructure

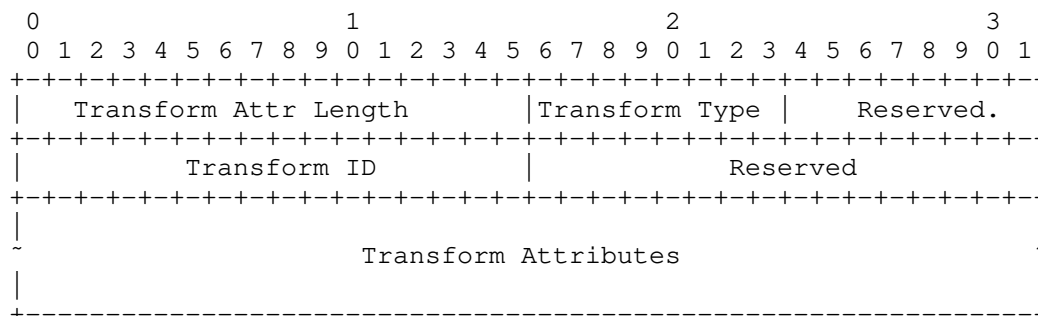


Figure 9

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].

11. Applicability

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 RFC 4364 [RFC4364], VPLS RFC 4761 [RFC4761] and RFC 4762 [RFC4762], and MVPN RFC 6513 [RFC6513] and RFC 6514 [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

Figure 10

12. Acknowledgements

TBD.

13. 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.

14. Security Considerations

This document proposes that a device re-use an ephemeral Diffie-Hellman exponential with multiple peers. There are some known potential vulnerabilities to this approach, which can be mitigated by the device first validating a peer's public value to be a safe public value before combining its own private value with it. The tests which MUST be performed are described in [RFC6989]. See [REUSE] for additional security considerations when reusing ephemeral Diffie-Hellman keys.

A controller acts as a "trusted third party", which asserts that a particular Diffie-Hellman public value is associated with a particular entity. A device receiving the public key is not required to validate the assertion.

A subverted controller can act as a "man-in-the-middle" between a pair of devices. The easiest attack would be for the attacker to adjust the routing for the desired traffic through a compromised gateway and directly observe the cleartext. It is also possible that a subverted controller could provide a device with a Diffie-Hellman public value that actually belongs to a compromised gateway rather

than the intended gateway, but doing so does not seem to be necessary. Nonetheless, the attack of a subverted controller can be mitigated by having a device sign its Diffie-Hellman public value (e.g, as a CMS Signed data object), where the receiver validates the digital signature on the object. However, this adds significant processing cost to a rekey and does not fit the controller-based network architecture model.

A subverted IPsec device whose DH pair has been compromised would be vulnerable to all of its IPsec traffic using that DH pair being compromised. Assuming the use of strong DH algorithms (including quantum resistant algorithms as they become available), the compromise would most likely be due to the device itself being compromised. Such a compromised device is also vulnerable to a direct plaintext compromise.

PFS is achieved between rekey periods, as DH pairs are required to be generated independently. However, because a device uses the same long-term key to generate session key with multiple peers, there is no PFS between sessions within the same rekey period. To reduce key exposure outside of a rekey period, when a connection is closed each endpoint MUST forget not only the keys used by the connection but also any information that could be used to recompute those keys. However, the DH private key value and the nonce distributed with it may be forgotten only once the last IPsec SA that uses the private key value is removed from the SAD and there is no chance that a new IPsec SA could be setup that requires the private key value.

If quantum resistance is considered to be an issue, the controller can distribute a PSK, which could be used to create the SK_d in the manner shown in [I-D.ietf-ipsecme-qr-ikev2].

15. References

15.1. Normative References

- [GENEVE] Gross, J., et al., "Geneve: Generic Network Virtualization Encapsulation", 2018,
<<https://tools.ietf.org/html/draft-ietf-nvo3-geneve-06>>.
- [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>>.

- [RFC3948] Huttunen, A., Swander, B., Volpe, V., DiBurro, L., and M. Stenberg, "UDP Encapsulation of IPsec ESP Packets", RFC 3948, DOI 10.17487/RFC3948, January 2005, <<https://www.rfc-editor.org/info/rfc3948>>.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<https://www.rfc-editor.org/info/rfc4301>>.
- [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>>.

15.2. Informative References

- [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>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <<https://www.rfc-editor.org/info/rfc4762>>.
- [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>>.

[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>>.

Appendix A. Additional Stuff

TBD.

Authors' Addresses

Ali Sajassi (editor)
Cisco
170 W Tasman Drive
San Jose, CA
USA

Email: sajassi@cisco.com

Ayan Banerjee
Cisco
170 W Tasman Drive
San Jose, CA
USA

Email: ayabaner@cisco.com

Sameer Thoria
Cisco
170 W Tasman Drive
San Jose, CA
USA

Email: sthoria@cisco.com

David Carrel
Graphiant
CA
USA

Email: carrel@graphiant.com

Brian Weis
Independent
CA
USA

Email: bew.stds@gmail.com

John Drake
Juniper Networks
CA
USA

Email: jdrake@juniper.net

BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2022

J. Rabadan, Ed.
S. Sathappan
Nokia
October 25, 2021

Domain Path (D-PATH) for Ethernet VPN (EVPN) Interconnect Networks
draft-sr-bess-evpn-dpath-00

Abstract

The BGP Domain PATH (D-PATH) attribute is defined for Inter-Subnet Forwarding (ISF) BGP Sub-Address Families that advertise IP prefixes. When used along with EVPN IP Prefix routes or IP-VPN routes, it identifies the domain(s) through which the routes have passed and that information can be used by the receiver BGP speakers to detect routing loops or influence the BGP best path selection. This document extends the use of D-PATH so that it can also be used along with EVPN MAC/IP Advertisement routes in EVPN Broadcast Domains (BD) and Auto-Discovery per EVPN Instance routes in Virtual Private Wire Services (VPWS).

Status of This Memo

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

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

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

This Internet-Draft will expire on April 28, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction and Problem Statement	2
2. Conventions used in this document	2
3. Terminology	3
4. Use of Domain Path Attribute (D-PATH) with EVPN MAC/IP Advertisement routes	4
4.1. Loop Detection	6
4.2. D-PATH and Best Path Selection for MAC/IP Advertisement routes	6
4.3. D-PATH and Best Path Selection for Ethernet A-D per EVI routes	7
4.4. Error Handling	8
5. Use-Case Examples	8
6. Security Considerations	11
7. IANA Considerations	11
8. Acknowledgments	11
9. Contributors	11
10. References	11
10.1. Normative References	11
10.2. Informative References	12
Authors' Addresses	12

1. Introduction and Problem Statement

The BGP Domain PATH (D-PATH) attribute [I-D.ietf-bess-evpn-ipvpn-interworking] is defined for Inter-Subnet Forwarding (ISF) BGP Sub-Address Families that advertise IP prefixes. When used along with EVPN IP Prefix routes or IP-VPN routes, it identifies the domain(s) through which the routes have passed and that information can be used by the receiver BGP speakers to detect routing loops or influence the BGP best path selection. This document extends the use of D-PATH so that it can also be used along with EVPN MAC/IP Advertisement routes in EVPN Broadcast Domains (BD) [I-D.ietf-bess-rfc7432bis] and Auto-Discovery per EVPN Instance routes in Virtual Private Wire Services (VPWS) [RFC8214].

2. Conventions used in this document

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

14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

This section summarizes the terminology that is used throughout the rest of the document.

- o MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [I-D.ietf-bess-rfc7432bis]. It is also the instantiation of an EVI (EVPN Instance) in a PE.
- o BD and BT: a Broadcast Domain and Bridge Table, as defined in [I-D.ietf-bess-rfc7432bis]. A BD is a group of PEs attached to the same EVPN layer-2 multipoint service. A BT is the instantiation of a Broadcast Domain in a PE. When there is a single Broadcast Domain in a given EVI, the MAC-VRF in each PE will contain a single BT. When there are multiple BTs within the same MAC-VRF, each BT is associated to a different Ethernet Tag. The EVPN routes specific to a BT, will indicate which Ethernet Tag the route corresponds to.
- o AC: Attachment Circuit or logical interface associated to a given BT. To determine the AC on which a packet arrived, the PE will examine the combination of a physical port and VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).
- o RT-2: Route Type 2 or MAC/IP Advertisement route, as per [I-D.ietf-bess-rfc7432bis].
- o RT-1: Route Type 1 or Ethernet Auto-Discovery route, as per [I-D.ietf-bess-rfc7432bis].
- o ES and ESI: Ethernet Segment and Ethernet Segment Identifier, as defined in [I-D.ietf-bess-rfc7432bis].
- o TS: Tenant System.
- o EVPN Layer2-Domain: two PEs are in the same Layer2-Domain if they are attached to the same tenant and the packets between them do not require a data path MAC lookup (in the BT of a MAC-VRF) in any intermediate router. A Layer2-Domain Gateway PE is always configured with multiple Layer2-Domain identifiers (Layer2-Domain-ID) in the MAC-VRF that connects those Layer2-Domains, each Layer2-Domain-ID representing a Layer2-Domain.

Example: Figure 1 illustrates an example where PE1 and PE2 belong to different Layer2-Domains since packets between them (for flows

between TS1 and TS2) require a MAC lookup in two of the gateways that are connecting the three EVPN Layer2-Domains. E.g., if frames from TS1 to TS2 go through PE1, GW1, GW3 and PE2, a MAC lookup is performed at GW1 and GW3.

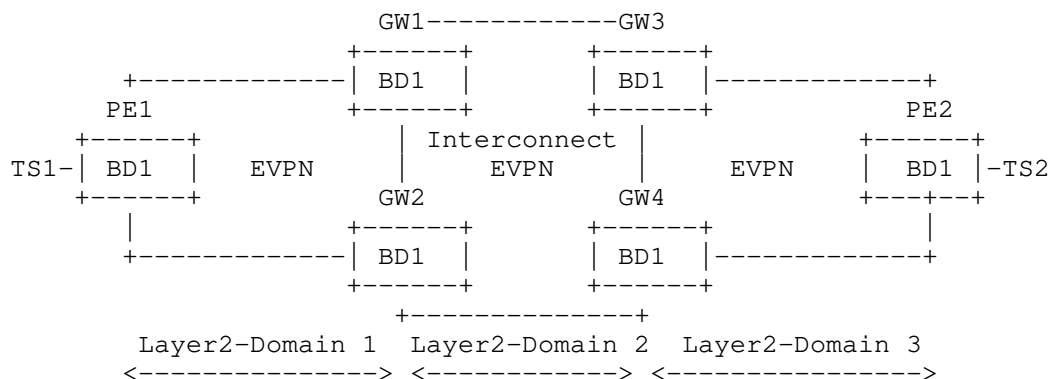


Figure 1: EVPN Sub-Domain example

- o Layer2-Domain Gateway PE: a PE that is attached to two or more EVPN Layer2-Domains. An example of Layer2-Domain Gateway PE is a PE following the procedures in section 4.4 or section 4.6 of [RFC9014]. In the example in Figure 1, GW1 and GW2 connect Layer2-Domains 1 and 2, whereas GW3 and GW4 connect Layer2-Domains 2 and 3. GW1 and GW2 import the MAC/IP Advertisement route for TS1 coming from the Layer2-Domain 1 into the MAC-VRF for BD1, and re-advertise it into Layer2-Domain 2. Likewise, GW3 and GW4 import the route into their MAC-VRF and re-advertise it into Layer2-Domain 3.
4. Use of Domain Path Attribute (D-PATH) with EVPN MAC/IP Advertisement routes

This document extends the use of the D-PATH attribute specified in [I-D.ietf-bess-evpn-ipvpn-interworking] so that D-PATH can be advertised and processed along with the following EVPN route types:

- o EVPN MAC/IP Advertisement routes that are not used for Inter-Subnet Forwarding (ISF) [RFC9135]. Note: if the EVPN MAC/IP Advertisement route is used for Inter-Subnet Forwarding, the procedures for the D-PATH advertisement and processing are described in [I-D.ietf-bess-evpn-ipvpn-interworking].
- o EVPN A-D per EVI routes that are used for EVPN-VPWS [RFC8214]. The use of D-PATH in A-D per EVI routes not used for EVPN-VPWS is out of scope of this document.

The use of D-PATH along with other EVPN route types will be described in future versions of this document.

When used along with non-ISF MAC/IP Advertisement routes or A-D per EVI routes, the D-PATH attribute is characterized as follows:

1. D-PATH is composed of a sequence of Domain segments following the format specified in [I-D.ietf-bess-evpn-ipvpn-interworking] where each Domain is represented as <DOMAIN-ID:ISF_SAFI_TYPE>. In this specification, DOMAIN-ID is a Layer2-Domain identifier configured in a MAC-VRF and ISF_SAFI_TYPE is set to either 70 (EVPN) or 0 (local route). To simplify the explanation, this document represents the domains for EVPN RT-1s and RT-2s as <Layer2-Domain-ID:TYPE>.
2. D-PATH identifies the sequence of Layer2-Domains the route has gone through, with the last <Layer2-Domain-ID:TYPE> entry (rightmost) identifying the first PE that added the D-PATH attribute.
3. D-PATH SHOULD be added/modified by a Layer2-Domain Gateway PE that re-advertises the route and MAY be added by a PE that originates the route, as follows:
 - A. A Layer2-Domain Gateway PE that connects two Layer2-Domains "X" and "Y", and receives a route on a Layer2-Domain identified by a Layer2-Domain-ID "X" SHOULD append a domain <X:EVPN> to the existing (or newly added) D-PATH attribute when re-advertising the route to Layer2-Domain "Y". The route is re-advertised if it is first imported in a MAC-VRF (or VPWS instance), the MAC (or Ethernet Tag) is active, and policy allows the re-export of the route to a BGP neighbor.
 - B. A Layer2-Domain Gateway PE MAY also add the D-PATH attribute for locally learned MACs or MAC/IP pairs. In this case, the domain added would be <A:0>, where A is the Layer2-Domain-ID configured on the Gateway PE's MAC-VRF that is specific to local routes (MAC/IP learned via local AC), and "0" is the TYPE of the Layer2-Domain and indicates that the route is locally originated and not re-advertised after receiving it from a BGP-EVPN neighbor. The Layer2-Domain-ID for local routes MAY be shared by all the redundant Layer2-Domain Gateway PEs for local routes, or each Layer2-Domain Gateway PE of the redundancy group can use its own Layer2-Domain-ID.
 - C. A PE that is connected to a single Layer2-Domain (therefore the PE is not a Layer2-Domain Gateway) MAY add the D-PATH with a domain <B:0>, where B is the Layer2-Domain-ID

configured on the PE's MAC-VRF (or VPWS) for locally learned MAC/IPs (or Ethernet Tag IDs for VPWS). "0" is the TYPE that indicates the route is not re-advertised, but originated in the PE.

4. On received EVPN routes, D-PATH is processed and used for loop detection (Section 4.1) as well as to influence the best path selection (Section 4.2 and Section 4.3).

4.1. Loop Detection

An EVPN route received by a PE with a D-PATH attribute that contains one or more of its locally associated Layer2-Domain-IDs for the MAC-VRF or VPWS instance is considered to be a looped route. A looped route MUST NOT be installed, but kept in the BGP RIB, flagged as "looped".

For instance, in the example of Figure 1, assuming PE1 advertises TS1's MAC/IP and does not add the D-PATH attribute, the Layer2-Domain Gateway GW1 receives two MAC/IP Advertisement routes for TS1's MAC/IP:

- o A RT-2 with next-hop PE1 and no D-PATH.
- o A RT-2 with next-hop GW2 and D-PATH={length=1,<6500:1:EVPN>}, assuming that the Layer2-Domain-ID for Layer2-Domain 1 is 6500:1.

In this case, Layer2-Domain Gateway GW1 flags the RT-2 with D-PATH as "looped", and does not install the MAC in the BT of the MAC-VRF, since the route includes one of GW1's Layer2-Domain-IDs.

4.2. D-PATH and Best Path Selection for MAC/IP Advertisement routes

When two (or more) MAC/IP Advertisement routes with the same route key (and same or different RDs) are received, a best path selection algorithm is used. This section summarizes the best path selection for MAC/IP Advertisement routes, which extends the rules in [I-D.ietf-bess-rfc7432bis] by including D-PATH in the tie-breaking algorithm. While the algorithm may be implemented in different ways, the selection result SHOULD be the same as the result of the rules that follow.

The tie-breaking algorithm begins by considering all EVPN MAC/IP Advertisement routes equally preferable routes to the same destination, and then selects routes to be removed from consideration. The process terminates as soon as only one route remains in consideration.

1. When the Default Gateway extended community is present in some of the routes, the MAC/IP Advertisement routes without the Default Gateway indication are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis].
2. Then the routes with the Static bit set in the MAC Mobility extended community are preferred, and the routes without the Static bit set are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis]. Note that this rule does not apply to routes with the Default Gateway extended community, since these routes SHALL NOT convey the MAC Mobility extended community [I-D.ietf-bess-rfc7432bis].
3. Then the routes with the highest MAC Mobility Sequence number are preferred, hence the routes that are not tied for having the highest Sequence number are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis].
4. Then routes with the highest Local Preference are preferred, hence routes that are not tied for having the highest Local Preference are removed from consideration, as defined in [RFC4271].
5. Then routes with the shortest D-PATH are preferred, hence routes not tied for the shortest D-PATH are removed. Routes without D-PATH are considered zero-length D-PATH.
6. Then routes with the numerically lowest left-most Sub-Domain-ID are preferred, hence routes not tied for the numerically lowest left-most Sub-Domain-ID are removed from consideration.
7. If the steps above do not produce a single route, the rest of the rules after the highest Local Preference in [RFC4271] apply after step 6.

The above selection criteria is followed irrespective of the ESI value in the routes. EVPN Multi-Homing procedures for Aliasing or Backup paths in [I-D.ietf-bess-rfc7432bis] are applied to the selected MAC/IP Advertisement route.

4.3. D-PATH and Best Path Selection for Ethernet A-D per EVI routes

When two (or more) EVPN A-D per EVI routes with the same route key (and same or different RDs) are received for a VPWS, a best path selection algorithm is used. The selection algorithm follows the same steps as in Section 4.2 except for steps 1-3 which do not apply since the Default Gateway and MAC Mobility extended community are irrelevant to the EVPN A-D per EVI routes.

The above selection is followed for A-D per EVI routes with ESI=0. For non-zero ESI routes, the EVPN Multi-Homing procedures in [RFC8214] for Aliasing and Backup path are followed to select the routes (P and B flags are considered for the selection of the routes when sending traffic to a remote Ethernet Segment). If the mentioned Multi-Homing procedures do not produce a single route to each of the remote PEs attached to the same ES, steps 4-7 in Section 4.2 are followed.

4.4. Error Handling

The error handling for the D-PATH attribute is described in [I-D.ietf-bess-evpn-ipvpn-interworking]. This document extends the use of D-PATH to non-ISF EVPN routes.

5. Use-Case Examples

This section illustrates the use of D-PATH in EVPN routes with examples.

Figure 2 and Figure 3 illustrate an integrated interconnect solution for an EVPN BD, as described in section 4.4 and section 4.6 of [RFC9014]. GW1 and GW2 are Layer2-Domain Gateway PEs connecting two L2-Domains identified by D-PATH domain {1:1:EVPN} and {1:2:EVPN}, respectively. Received Ethernet A-D routes, ES routes, and Inclusive Multicast routes from the routers in one Layer2-Domain are consumed and processed by GW1 and GW2, but not re-advertised to the other Layer2-Domain. However, MAC/IP Advertisement routes received by GW1 and GW2 in one Layer2-Domain are processed and, if installed, re-advertised into the other Layer2-Domain.

Consider the example of Figure 2, where PE1 advertises a MAC/IP Advertisement route for M1/IP1. The route is processed and installed by GW1 and GW2 in BD1, and both will re-advertise the routes into the Layer2-Domain-2. By using D-PATH in GW1 and GW2, when the route is received on PE2, PE2 has the visibility of the Layer2-Domains through which the route has gone, and can also use the D-PATH for best path selection in case PE2 receives a MAC/IP Advertisement route for M1/IP1 by some other means. In addition, GW1 and GW2 can compare the D-PATH of the incoming routes with their local list of Layer2-Domain-IDs, and detect a loop if any of the local Layer2-Domain-IDs matches a domain in the received D-PATH. This procedure prevents the re-advertisement of the route back into Layer2-Domain-1.

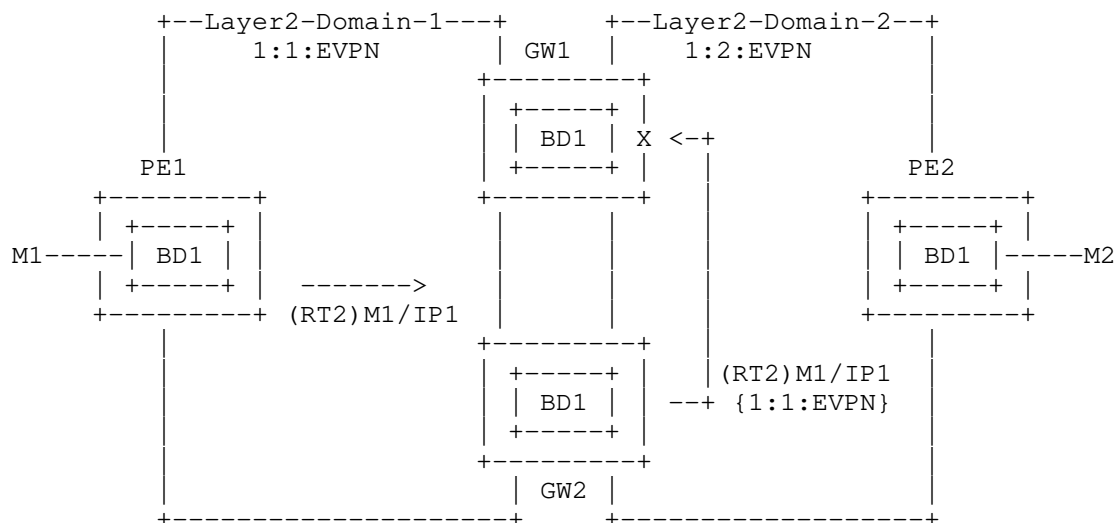


Figure 2: EVPN Interconnect

The example of Figure 3 illustrates how GW1 and GW2 can also have local ACs in BD1 and learn local MAC (or MAC/IP) addresses from devices connected to the ACs. Assuming GW2 learns M3/IP3 via local AC, GW2 advertises a MAC/IP Advertisement route for M3/IP3 into each of the Layer2-Domains that GW2 is connected to. As described in Section 4, GW2 can advertise these two MAC/IP Advertisement routes with a configured Layer2-Domain-ID for local MAC/IPs routes that can be shared with GW1. Consider this Layer2-Domain-ID is 1:3 and it is configured on both, GW1 and GW2. When GW2 advertises the route into each Layer2-Domain, it adds the D-PATH attribute with a domain {1:3:0}. These routes will be flagged by GW1 as "looped" since 1:3 is configured as a local Layer2-Domain-ID in GW1. In addition, PE1 and PE2 will receive the routes with the D-PATH and they will have the visibility of the origin of the routes, in this case local Layer2-Domain Gateway routes. This information can be used to influence the best path selection in case of multiple routes for M3/IP3 are received on PE1 or PE2 for BD1.

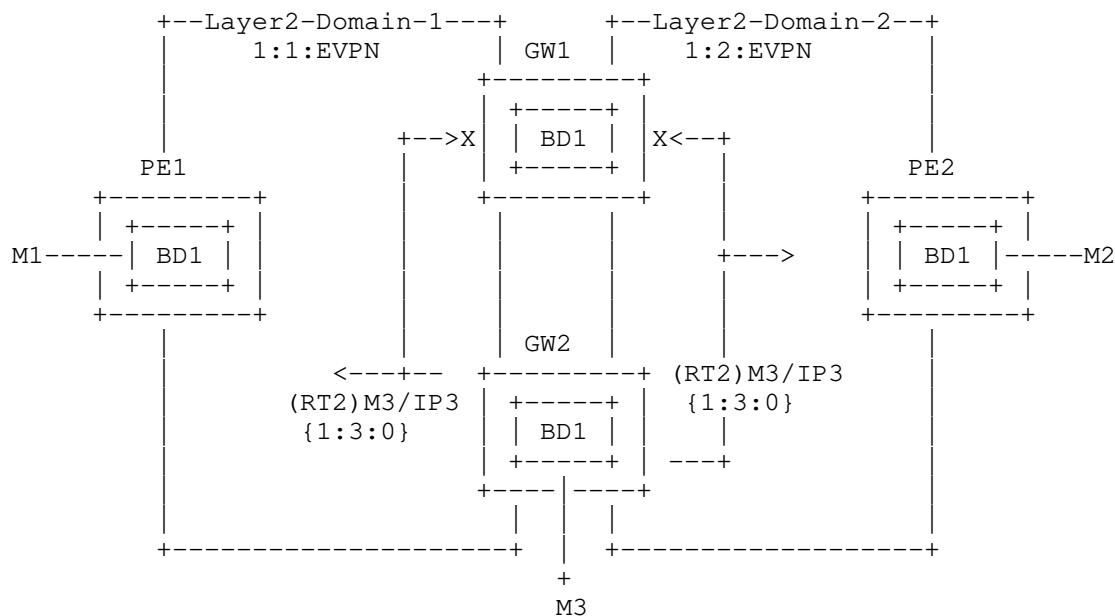


Figure 3: EVPN Interconnect local AC

As an alternative solution to configuring the same Layer2-Domain-ID for local routes on both Layer2-Domain Gateways, GW2 can be configured with Layer2-Domain-ID 1:3 for local routes, and GW1 can use a different Layer2-Domain-ID, e.g., 1:4. In this case, GW2 advertises the route for M3/IP3 into each Layer2-Domain as before, but now GW1 will not flag the route as "looped" since 1:3 is not on the list of GW1's local Layer2-Domain-IDs. GW1 receives the routes from both Layer2-Domains, and GW1 selects the route from e.g., Layer2-Domain-1. GW1 then installs the route in its BT and re-advertises the route into Layer2-Domain-2 with D-PATH {1:1:EVPN, 1:3:0}. When PE2 receives two routes for M3/IP3, one from GW2 with D-PATH {1:3:0} and another from GW1 with D-PATH {1:1:EVPN, 1:3:0}, PE2 will use best path selection and choose to send its traffic to GW2. Also GW2 will receive the route for M3/IP3 from GW1 and mark it as "looped" since that route conveys its own Layer2-Domain-IDs 1:1 and 1:3.

In a nutshell, the use of D-PATH in MAC/IP Advertisement routes helps prevent loops and influences the best path selection so that PEs choose the shortest paths to the destination PEs.

6. Security Considerations

To be added.

7. IANA Considerations

None.

8. Acknowledgments

9. Contributors

10. References

10.1. Normative References

- [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>>.
- [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>>.
- [I-D.ietf-bess-evpn-ipvpn-interworking] Rabadan, J., Sajassi, A., Rosen, E., Drake, J., Lin, W., Uttaro, J., and A. Simpson, "EVPN Interworking with IPVPN", draft-ietf-bess-evpn-ipvpn-interworking-06 (work in progress), September 2021.
- [I-D.ietf-bess-rfc7432bis] Sajassi, A., Burdet, L. A., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", draft-ietf-bess-rfc7432bis-01 (work in progress), July 2021.
- [RFC9014] Rabadan, J., Ed., Sathappan, S., Henderickx, W., Sajassi, A., and J. Drake, "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, DOI 10.17487/RFC9014, May 2021, <<https://www.rfc-editor.org/info/rfc9014>>.

[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>>.

10.2. Informative References

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <<https://www.rfc-editor.org/info/rfc9135>>.

Authors' Addresses

J. Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

S. Sathappan
Nokia
701 Middlefield Road
Mountain View, CA 94043
USA

Email: senthil.sathappan@nokia.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: 9 January 2022

P. Thubert, Ed.
Cisco Systems
A. Przygienda
Juniper Networks, Inc
J. Tantsura
Microsoft
8 July 2021

Secure EVPN MAC Signaling
draft-thubert-bess-secure-evpn-mac-signaling-00

Abstract

This specification adds attributes to EVPN to carry IPv6 address metadata learned from RFC 8505 and RFC 8928 so as to maintain a synchronized copy of the 6LoWPAN ND registrar at each EVPN router and perform locally a unicast IPv6 ND service for address lookup and duplicate address detection.

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 9 January 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Terminology	4
2.1. Requirements Language	4
2.2. Glossary	4
2.3. References	5
3. 6LoWPAN Neighbor Discovery	6
3.1. RFC 6775 Address Registration	6
3.2. RFC 8505 Extended Address Registration	7
3.2.1. R Flag	8
3.2.2. TID, "I" Field and Opaque Fields	8
3.2.3. Status	8
3.2.4. Route Ownership Verifier	9
3.3. RFC 8505 Extended DAR/DAC	9
3.4. RFC 7400 Capability Indication Option	10
4. Extending 6LoWPAN ND	11
4.1. Use of the R flag in NA	11
4.2. Distributing the 6LBR	11
4.3. Unicast Address Lookup with the 6LBR	15
5. Requirements on the EVPN-Unaware Host	21
5.1. Support of 6LoWPAN ND	21
6. Enhancements to EVPN	22
6.1. ROVR MAC Mobility Extended Community	24
6.2. Extended ROVR MAC Procedures	25
7. Protocol Operations	26
8. Security Considerations	36
9. IANA Considerations	37
10. Acknowledgments	37
11. Normative References	37
12. Informative References	38
Authors' Addresses	39

1. Introduction

"Registration Extensions for IPv6 over 6LoWPAN Neighbor Discovery" [RFC8505] (ND) provides a zeroconf routing-agnostic Host-to-Router Link-Local interface for Stateful Address Autoconfiguration. "Address-Protected Neighbor Discovery for Low-Power and Lossy Networks" [RFC8928] (AP-ND) adds a zeroconf anti-theft protection that protects the ownership of the autoconfigured address with autoconfigured proof of ownership called a Registration Ownership Verifier (ROVR).

[RFC8505] enables the host to claim an IPv6 address and obtain reachability services for that address. It is already used to inject host routes in RPL [RFC9010] and RIFT "Routing in Fat Trees" [RIFT], and to maintain a proxy-ND state in a backbone router [RFC8929]; this specification extends its applicability to the case of Ethernet Virtual Private Network (eVPN).

[RFC8505] specifies a unicast address registration mechanism that enables the host called a 6LowPAN Node (6LN) to install a ND binding state in the 6LowPAN Router (6LR) that can serve as Neighbor Cache Entry (NCE), though it is not operated as a cache. The protocol provides the means to reject the registration in case of address duplication. It also enables to discriminate mobility from multihoming. [RFC8928] adds the capability to verify the ownership of the address and prevent an attacker from stealing and/or impersonating an address.

[RFC8505] defines the 6LoWPAN Border Router (6LBR) as an abstract address registrar that provides authoritative service for Address Registration and duplicate detection. The 6LBR stores address metadata that is obtained during the Address Registration, including an owner ID and a sequence counter. As part of the process of a new Address Registration, the 6LR queries the 6LBR for existing metadata related to the address being registered. This enables in particular to detect a duplication and reject the registration. This specification extends the 6LBR abstract data model to store the Link Layer Address (LLA) of the Registering Node. This enables the 6LBR to perform locally, and using unicast communication, the IPv6 ND services of address lookup and duplicate address detection.

The [RFC8505] address registrar can be centralized, but it can also be distributed and maintained synchronized using a routing protocol. This specification adds attributes to EVPN to carry the IPv6 address metadata learned from [RFC8505] so as to maintain a synchronized copy of the 6LBR abstract data at each EVPN router.

2. Terminology

2.1. 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.

2.2. Glossary

This document uses the following acronyms:

6CIO Capability Indication Option [RFC7400]
6LN: 6LoWPAN Node (the Host) [RFC6775]
6LR: 6LoWPAN router (the router) [RFC6775]
6LBR: 6LoWPAN Border router [RFC6775]
AMC: Address Mapping Confirmation [UNICAST-LOOKUP]
AMR: Address Mapping Request [UNICAST-LOOKUP]
ARO Address Registration Option [RFC6775]
CIPO: Crypto-ID Parameters Option
DAD: Duplicate Address Detection [RFC4862]
ICMPv6: Internet Control Message Protocol for IPv6
DAC Duplicate Address Confirmation [RFC6775]
DAR Duplicate Address Request [RFC6775]
EDAC Extended Duplicate Address Confirmation [RFC8505]
EDAR Extended Duplicate Address Request [RFC8505]
EARO: Extended Address Registration Option [RFC8505]
EVPN: Ethernet VPN [RFC7432]
LLA: Link-Layer Address (the MAC address on Ethernet)
LLN Low-Power and Lossy Network [RFC6550]
NA: Neighbor Advertisement [RFC4861]
NCE: Neighbor Cache Entry [RFC4861]
ND: Neighbor Discovery [RFC4861]
NDPSO: Neighbor Discovery Protocol Signature Option
NS: Neighbor Solicitation [RFC4861]
RA: Router Advertisement [RFC4861]
ROVR: Registration Ownership Verifier [RFC8505]
TID: Transaction ID (a sequence counter in the EARO) [RFC8505]
SLAAC: Stateless Address Autoconfiguration [RFC4862]
SLLAO: Source Link-Layer Address Option [RFC4861]
TLLAO: Target Link-Layer Address Option [RFC4861]
ROVR MAC: MAC obtained from a host meeting requirements in Section 5
Validated ROVR MAC: ROVR MAC validated by procedures specified in [RFC8928]
ROVR Node: EVPN node capable of advertising ROVR MACs
non-ROVR Node: EVPN node not supporting extensions defined in this

document.
VPN: Virtual Private Network

2.3. References

This document uses the terms Clos fabric and Fat Tree interchangeably, to refer to a folded spine-and-leaf topology as defined in the terminology section of "RIFT: Routing in Fat Trees" [RIFT].

The term "leaf" represents the access switch that connects the servers to the Fat Tree. The leaf is typically a Top-of-Rack (ToR) switch.

This specification uses the terms 6LN, 6LR and 6LBR to refer specifically to nodes that implement the said roles in [RFC8505] and does not expect other functionality such as 6LoWPAN Header Compression:

- * In the context of this document, the 6LN is a server that advertises an address mapping using [RFC8505], and optionally protects its ownership with [RFC8928].
- * The 6LR and 6LBR function are collapsed at the leaf and its state is synchronized with that of the EVPN functional support using an internal interface that is out of scope. That interface could be "pull" meaning that the 6LBR fetches the EVPN information when it needs it, or "push", meaning that any information that EVPN distributes is immediately fed in all the 6LBRs in all the leaves. Note that this is pure control plane and is not subject to abbreviating optimization as the FIB may be.

In this document, readers will encounter terms and concepts that are discussed in the following documents:

EVPN: "BGP MPLS-Based Ethernet VPN" [RFC7432] and "Network Virtualization Overlay Solution" [RFC8365],

Classical IPv6 ND: "Neighbor Discovery for IP version 6" [RFC4861] and "IPv6 Stateless Address Autoconfiguration" [RFC4862],

6LoWPAN ND: Neighbor Discovery Optimization for Low-Power and Lossy Networks [RFC6775], "Registration Extensions for 6LoWPAN Neighbor Discovery" [RFC8505], "Address Protected Neighbor Discovery for Low-power and Lossy Networks" [RFC8928], and "IPv6 Backbone Router" [RFC8929].

3. 6LoWPAN Neighbor Discovery

6LoWPAN Neighbor Discovery defines a stateful address autoconfiguration mechanism for IPv6. 6LoWPAN ND enables to divorce the L3 abstractions for link and subnet from the characteristics of the L2 link and broadcast domain. It is applicable beyond its original field of IoT to any environment where the broadcast nature of the underlaying network should not be exploited, e.g., in the case of a wireless link where broadcast uses an excessive amount of spectrum, and a distributed cloud, where it may span too widely.

In contrast to Stateless Address Autoconfiguration (SLAAC) [RFC4862] which relies on broadcast for duplicate address detection (DAD) and address lookup, 6LoWPAN ND installs and maintains a state in the neighbors for the duration of their interaction. Though it is also called a Neighbor Cache Entry (BCE) in [RFC6775], and in contrast with the the BCE in SLAAC, that state is not a cache that can be casually flushed and rebuilt. It must be installed proactively and refreshed periodically to maintain the connectivity and enable unicast-only operations.

The typical abstraction for an IP Link with 6LoWPAN ND is point-to-point (P2P) between a node and a router. An IP interface bundles multiple links between this node and peers in the same subnet, aka point-to-multipoint (P2MP). The subnet is a not-on-link L3-connected collection of such nodes and links, which means that the any-to-any connectivity across the subnet is ensured through L3 routing as opposed to transitive (any-to-any) reachability from L2.

This section goes through the 6LoWPAN ND mechanisms that this specification leverages, as a non-normative reference to the reader. The relevant normative text is to be found in [RFC6775], [RFC8505], and [RFC8928].

3.1. RFC 6775 Address Registration

The classical "IPv6 Neighbor Discovery (IPv6 ND) Protocol" [RFC4861] [RFC4862] was defined for serial links and transit media such as Ethernet. It is a reactive protocol that relies heavily on multicast operations for Address Discovery (aka Lookup) and Duplicate Address Detection (DAD).

"Neighbor Discovery Optimizations for 6LoWPAN networks" [RFC6775] adapts IPv6 ND for operations over energy-constrained LLNs. The main functions of [RFC6775] are to proactively establish the Neighbor Cache Entry (NCE) in the 6LR and to prevent address duplication. To that effect, [RFC6775] introduces a new unicast Address Registration mechanism that contributes to reducing the use of multicast messages compared to the classical IPv6 ND protocol.

[RFC6775] defines a new Address Registration Option (ARO) that is carried in the unicast Neighbor Solicitation (NS) and Neighbor Advertisement (NA) messages between the 6LoWPAN Node (6LN) and the 6LoWPAN router (6LR). It also defines the Duplicate Address Request (DAR) and Duplicate Address Confirmation (DAC) messages between the 6LR and the 6LBR. In a Low-Power and Lossy Network (LLN), the 6LBR is the central repository of all the Registered Addresses in its domain and the authoritative source of truth for uniqueness and ownership.

3.2. RFC 8505 Extended Address Registration

"Registration Extensions for 6LoWPAN Neighbor Discovery" [RFC8505] updates RFC 6775 into a generic Address Registration mechanism that can be used to access services such as routing and ND proxy. To that effect, [RFC8505] defines the Extended Address Registration Option (EARO), shown in Figure 1:

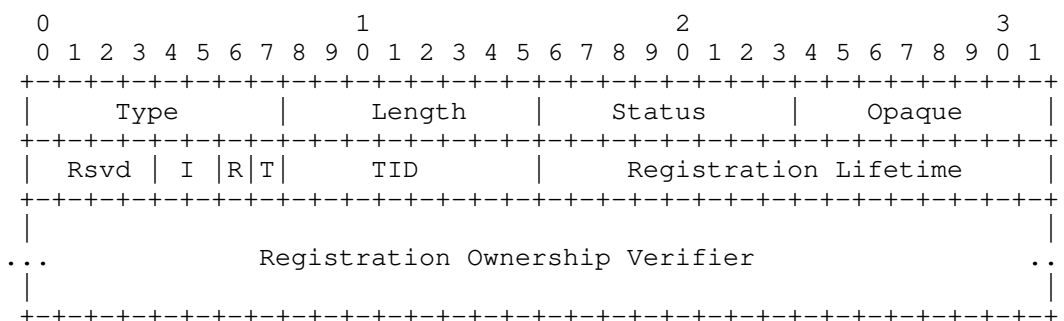


Figure 1: EARO Option Format

3.2.1. R Flag

[RFC8505] introduces the R Flag in the EARO. The Registering Node sets the R Flag to indicate whether the 6LR should ensure reachability for the Registered Address. If the R Flag is set to 0, then the Registering Node handles the reachability of the Registered Address by other means. In an EVPN network, this means that either it is a RAN that injects the route by itself or that it uses another EVPN router for reachability services.

This document specifies how the R Flag is used in the context of EVPN. An EVPN Host that implements the 6LN functionality from [RFC8505] requires reachability services for an IPv6 address if and only if it sets the R Flag in the NS(EARO) used to register the address to a 6LR acting as an EVPN border router. Upon receiving the NS(EARO), the EVPN router generates a BGP advertisement for the Registered Address if and only if the R flag is set to 1.

[RFC9010] specifies that the 'R' flags is set in the responded NA messages if and only if the route was installed. This specification echoes that behavior.

3.2.2. TID, "I" Field and Opaque Fields

When the T Flag is set to 1, the EARO includes a sequence counter called Transaction ID (TID), that is needed to format the MAC Mobility Extended Community. This is the reason why the support of [RFC8505] by the Host, as opposed to only [RFC6775], is a prerequisite for this specification); this requirement is fully explained in Section 5.1. The EARO also transports an Opaque field and an associated "I" field that describes what the Opaque field transports and how to use it.

This document specifies the use of the "I" field and the Opaque field by a Host.

3.2.3. Status

The values of the EARO status are maintained by IANA in the Address Registration Option Status Values subregistry [IANA-EARO-STATUS] of the Internet Control Message Protocol version 6 (ICMPv6) Parameters registry.

[RFC6775] and [RFC8505] defined the original values whereas [RFC9010] reduced range to 64 values and reformatted the octet field to enable to transport an external error, e.g., coming from a routing protocol.

This specification uses the format expressed in [RFC9010]. The value of 0 denotes an unqualified success, 1 indicates an address duplication, 3 a TID value that is outdated, and 4 is used in an asynchronous NA to indicate that 6LN should remove that address and possibly form new ones.

3.2.4. Route Ownership Verifier

Section 5.3 of [RFC8505] introduces the Registration Ownership Verifier (ROVR) field of variable length from 64 to 256 bits. The ROVR is a replacement of the EUI-64 in the ARO [RFC6775] that was used to identify uniquely an Address Registration with the Link-Layer address of the owner but provided no protection against spoofing.

"Address Protected Neighbor Discovery for Low-power and Lossy Networks" [RFC8928] leverages the ROVR field as a cryptographic proof of ownership to prevent a rogue third party from registering an address that is already owned. The use of ROVR field enables the 6LR to block traffic that is not sourced at an owned address.

This specification does not address how the protection by [RFC8928] could be extended for use in EVPN. On the other hand, it adds the ROVR to the BGP advertisement to share the state with the other routers via the Reflector (see Section 6.1), which means that the routers that are aware of the Host route are also aware of the ROVR associated to the Target Address, whether it is cryptographic and should be verified.

3.3. RFC 8505 Extended DAR/DAC

[RFC8505] updates the DAR/DAC messages to EDAR/EDAC messages to carry the ROVR field. The EDAR/EDAC exchange takes place between the 6LR to which the node registers an address, and the abstract 6LBR that stores the reference value for the ROVR and the TID associated to that address. It is triggered by an NS(EARO) message from a 6LN to the 6LR, to create, refresh, compare and delete the corresponding state in the 6LBR.

In the status returned with the EDAC message, the 6LBR indicates if the registration is accepted, should be challenged, or is duplicate. The status of 0 (success) indicates that the address is either new or that the current registration matches, and in particular that the ROVR at the 6LBR and the one in the EDAR message are identical.

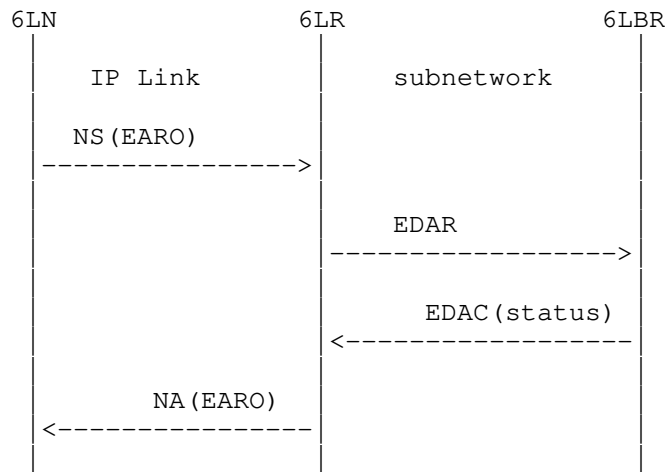


Figure 2: EDAR/EDAC flow

The EDAR/EDAC exchange is protected by the retry mechanism specified in Section 8.2.6 of [RFC6775], though in a data center, a duration significantly shorter than the default value of the Retransmission Timer [RFC4861] of 1 second may be sufficient to cover the round-trip delay between the 6R and the 6LBR.

With this specification, the 6LBR is distributed across the leaves, and all the leaves where an address is currently registered maintain a full 6LBR state for the address, aka local state in the following text. The specification leverages the EDAR/EDAC exchange to ensure that a leaf (acting as a 6LR) that needs to create a 6LBR state for a new registration has the same value for the ROVR as any 6LBR already serving that address on another leaf. At the same time, the specification avoids placing full ROVR information in BGP so 1) it is not observable by a potential attacker and 2) the new attributes remain reasonably small.

3.4. RFC 7400 Capability Indication Option

"6LoWPAN-GHC: Generic Header Compression for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)" [RFC7400] defines the 6LoWPAN Capability Indication Option (6CIO) that enables a node to expose its capabilities in router Advertisement (RA) messages.

[RFC8505] defines a number of bits in the 6CIO, in particular:

L: Node is a 6LR.

E: Node is an IPv6 ND Registrar -- i.e., it supports registrations

based on EARO.

P: Node is a Routing Registrar, -- i.e., an IPv6 ND Registrar that also provides reachability services for the Registered Address.

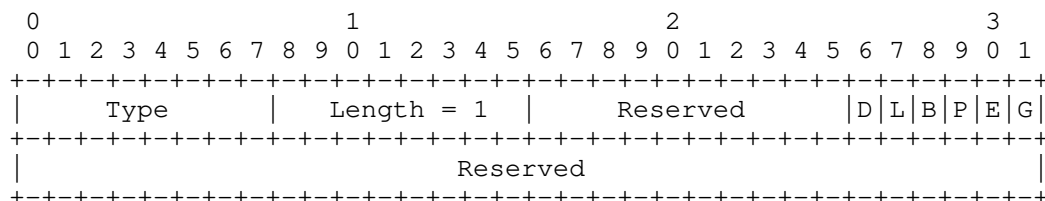


Figure 3: 6CIO flags

A 6LR that provides reachability services for a Host in an EVPN network as specified in this document includes a 6CIO in its RA messages and set the L, P and E flags to 1 as prescribed by [RFC8505].

4. Extending 6LoWPAN ND

4.1. Use of the R flag in NA

This document extends [RFC8928] and [RFC8505] as follows

This document also updates the behavior of a 6LR acting as EVPN router and of a 6LN acting as Host in the 6LoWPAN ND Address Registration as follows:

- * The use of the R Flag is extended to the NA(EARO) to confirm whether the route was installed.

4.2. Distributing the 6LBR

This specification enables to distribute the 6LBR at the edge of the EVPN network and collapse the 6LBR function with that of the EVPN support. In that model, the EVPN to 6LBR interaction becomes an internal interface, where each side informs the other in case of new information concerning an IP to Link-Layer Address (LLA) mapping. Since this is an internal interface, this specification makes no assumption on whether the 6LBR stores its own representation of the full EVPN state, which means that the EVPN support informs the 6LBR in case of any change on the EVPN side (this is called the push model, see Figure 10), or if the 6LBR queries the EVPN support when it does not have a mapping to satisfy a request (pull model, see Figure 9).

This specification leverages [RFC8929] that augments the abstract data model of the 6LBR to store the LLA associated with the registered address. Based on that additional state, the 6LBR in a leaf can communicate the mapping to the collocated EVPN function and respond to unicast address mapping lookups from the server side.

In an environment where the server ranges from a classical host to a more complex platform that runs a collection of virtual hosts interconnected by a virtual switch, but where the host-to-leaf interface remains at layer 2, the 6LR and the 6LBR functions can be collapsed in the leaf. The 6LR to 6LBR interaction also becomes an internal interface, and there is no need for EDAR/EDAC messages.

In that case, the MAC address associated to the Registered Address is indicated in the Target Link-Layer Address Option (TLLAO) in the NS message used for the registration, as shown in Figure 4. In the case of a pull model, if the 6LBR does not have a local state for the mapping, it queries the EVPN support to obtain the EVPN state if any. If a mapping is known then the 6LR/6LBR evaluates the registration for address duplication and other possible issues per [RFC8505]. Else (this is for a new mapping), if the registration is accepted, then the 6LBR notifies the EVPN support to inject a route type 2 in the fabric.

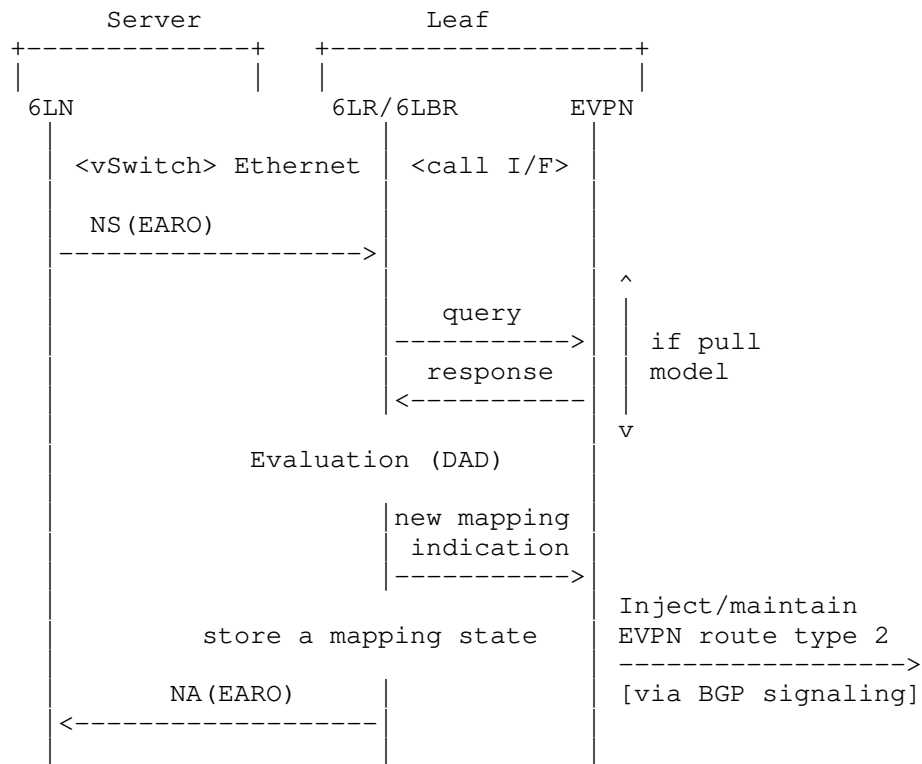


Figure 4: Direct Registration

In another type of deployment, the 6LR may be a virtual router in the server whereas the 6LBR runs in the leaf node. To address that case, the EDAR/EDAC may be used to communicate as shown in figure 5 of [RFC8505]. This draft leverages the capability to insert IPv6 ND options in the EDAR and EDAC messages introduced in [RFC8929] to place a TLLAO that carries the MAC address associated to the Registered address in the EDAR and EDAC messages as shown in Figure 5:

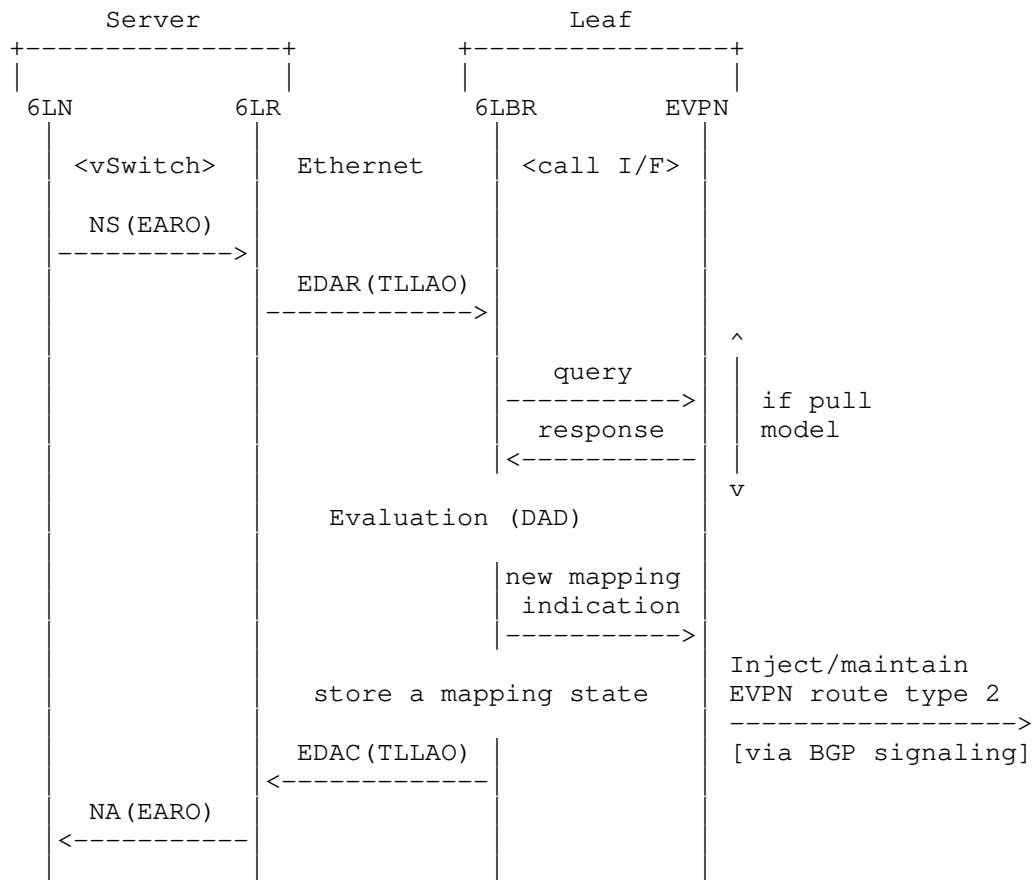


Figure 5: leveraging EDAR

[RFC8505] updates the DAR/DAC messages into the Extended DAR/DAC to carry the ROVR field. With this specification, the abstract 6LBR is distributed in all the Leaf nodes and synchronized with EVPN. When a server successfully registers an address to a leaf, the 6LR on that leaf becomes 6LBR for that address. It stores the full state for that address including the ROVR and the TID. When the address registration moves to another leaf, an EDAR/EDAC flow between the 6LR in the new leaf and the 6LBR in the old leaf confirms that the ROVR in the NS(EARO) received at the new leaf is correct, in which case the 6LR in the new leaf becomes 6LBR.

When the address is already registered to the local leaf, the EDAR/EDAC exchange is either local between a virtual router in the server and the leaf, or internal to the leaf between a collapsed 6LR and

6LBR. Based on its local state, the 6LBR in the leaf checks whether the proposed address/route is new and legit, and can reject it otherwise.

It results that duplicate addresses and address impersonation attacks can be filtered at the level of IPv6 ND by the 6LBR before the information reaches EVPN.

4.3. Unicast Address Lookup with the 6LBR

A classical IPv6 ND stack in the server that treats the subnet prefix as on-link (more in section 4.6.2. of [RFC4861]), will resolve an unknown LLA mapping with a multicast NS(lookup) message addressed to the solicited node multicast address (SNMA) associated with the destination address being resolved. The RECOMMENDED operation in that case is for the 6LBR that has a mapping state to forward the packet as a unicast MAC to the LLA that is stored for the IPv6 address as expected by [RFC6085]. The actual owner of the address can then answer unicast with a NA message, setting the override (O) bit to 1, as shown in Figure 6.

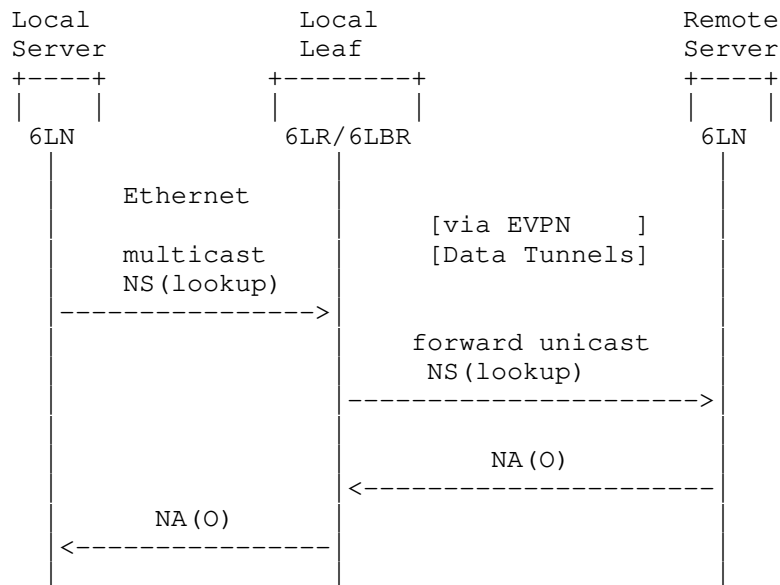


Figure 6: Forwarding legacy NS (Lookup)

Section 3.1. of [RFC8929] adds the capability to insert IPv6 ND options in the EDAR and EDAC messages. This enables the 6LBR to store the link-layer address associated with the Registered Address and to serve as a mapping server. [UNICAST-LOOKUP] leverages that state to define a new unicast address lookup operation, extending the EDAR and EDAC messages as the Address Mapping Request (AMR) and Confirmation (AMC) with a different Code Prefix [RFC8505].

In that model, the router advertises the subnet prefix as not on-link by setting the L flag to 0 in the Prefix Information Option (PIO), more in section 4.6.2. of [RFC4861]. The expected behavior is that the host that communicates with a peer in the same subnet refrains from resolving the address mapping and passes the packets directly to the router.

In the case where the router is a virtual 6LR running in the server, and the source and destination are in the same subnet served by EVPN, the router then resolves the address mapping on behalf of the host. To that effect, the router sends a unicast AMR message to the 6LBR. The message contains the SLLAO of the router to which the 6LBR will reply. If the binding is found, the 6LBR replies with an AMC message that contains the TLLOA with the requested MAC address, as shown in Figure 7.

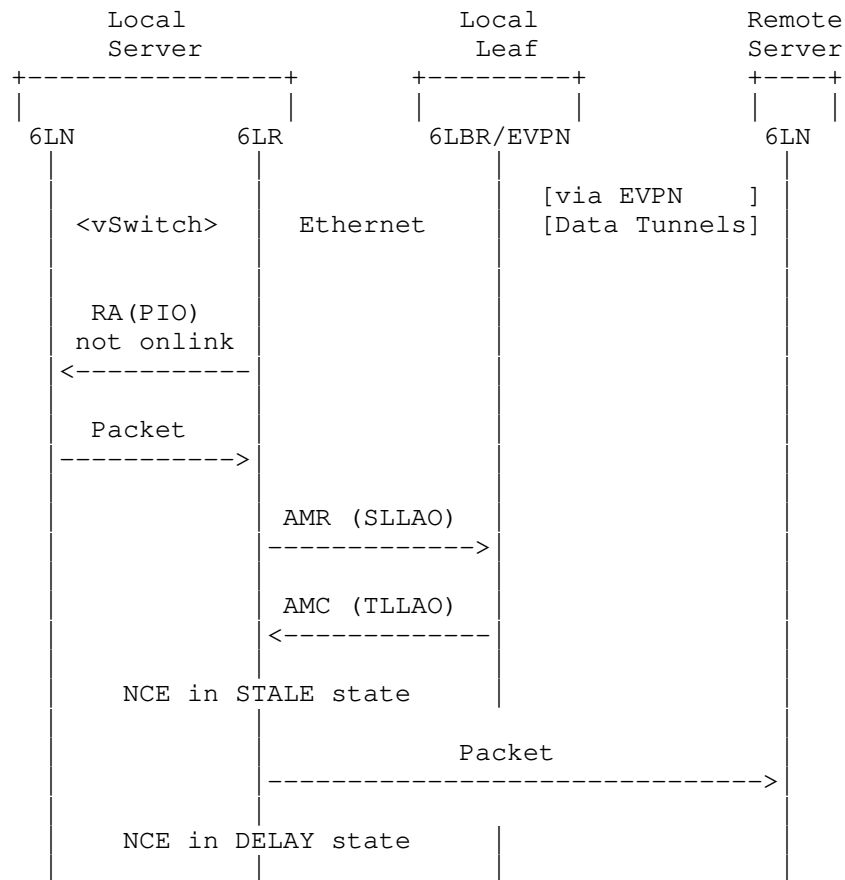


Figure 7: Unicast Lookup from the virtual Host

If it is not found, [UNICAST-LOOKUP] provides the capability to indicate immediately that the mapping is not known with a "not found" status in the AMC, as opposed to waiting for an NS(lookup) and retries to time out per [RFC4861].

In a fully stateful subnet where all nodes register all their addresses with [RFC8505], this means that the looked up address is not present in the network; in that case the packet is dropped and an ICMP error type 1 "Destination Unreachable" code 3 "Address unreachable" [RFC4443] is returned as shown in Figure 8.

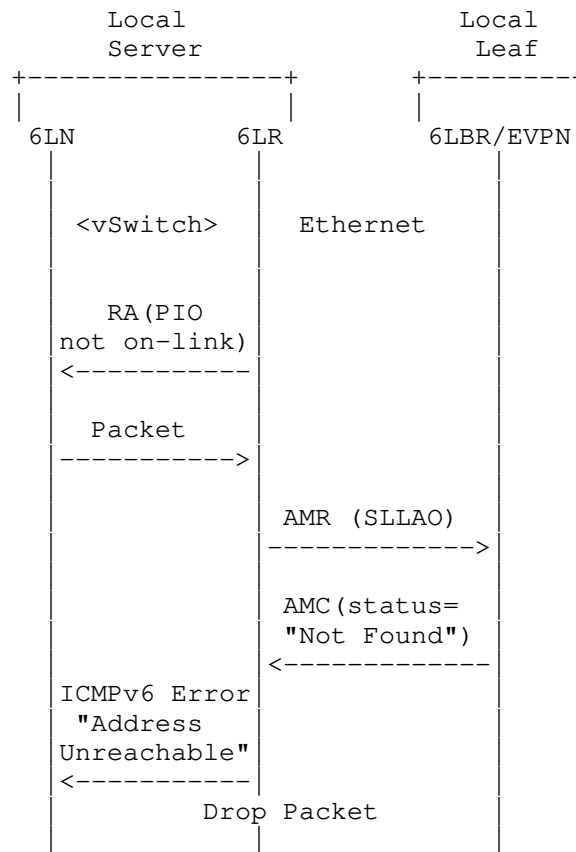


Figure 8: Unicast Lookup failure

Note that the figures above make no assumption on the pull vs. push model. In the case of pull model, the 6LBR queries the EVPN support when it does not have the mapping information to satisfy a request. Figure 9 illustrates a successful pull model lookup flow, when the route type 2 for the mapping is already known on the EVPN side.

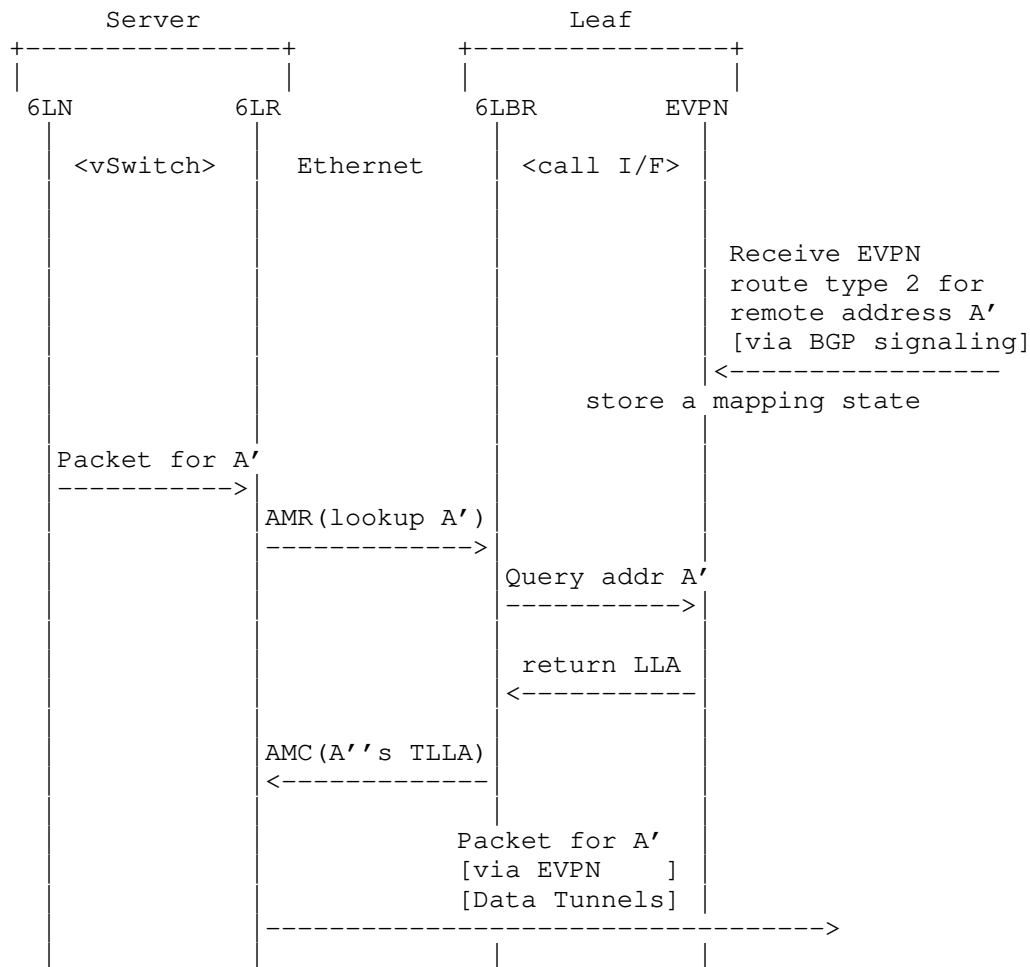


Figure 9: Pull model

In the case of push model, the EVPN support synchronizes its state upon a route type 2 with the 6LBR, and the 6LBR maintains an abstract data structure for all information known to EVPN. This way, the 6LBR already has the mapping information to satisfy any request for an existing mapping and it can answer right away. Figure 10 illustrates a successful push model lookup flow, when the 6LBR is already in possession of the mapping.

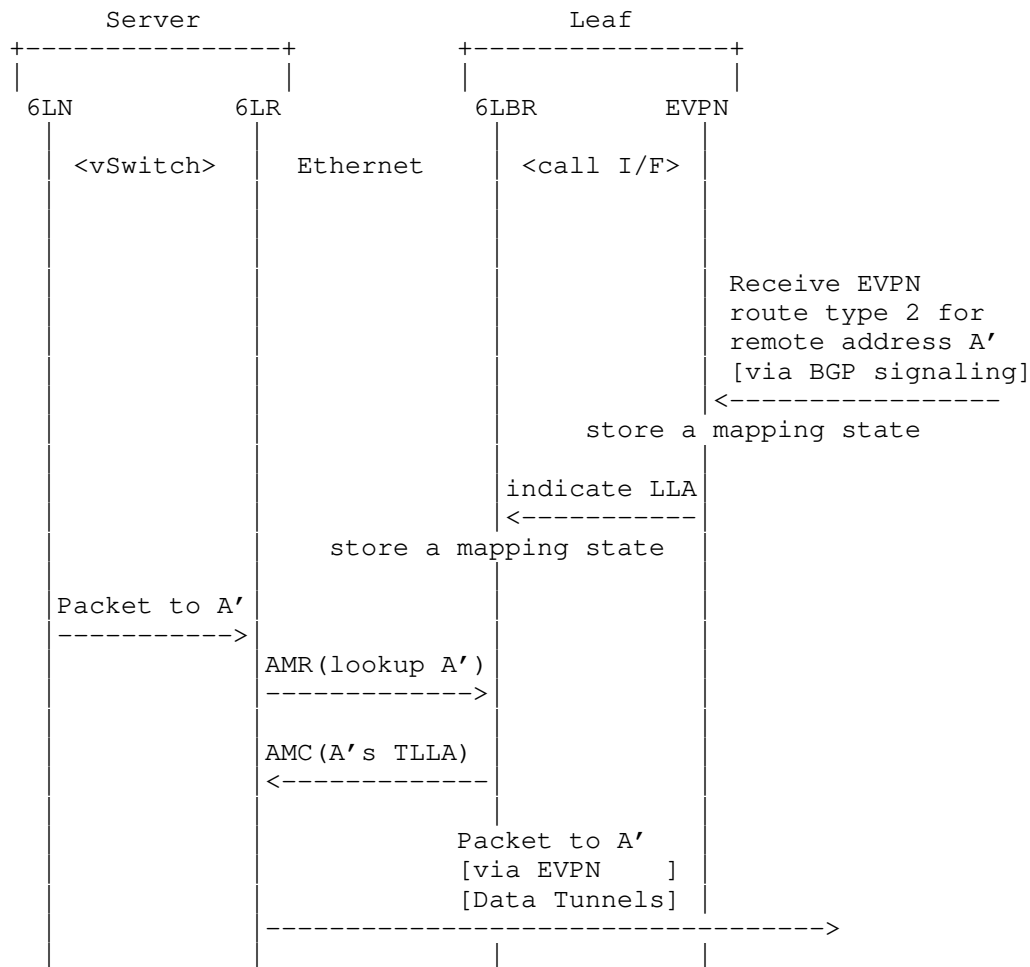


Figure 10: Push model

In a mixed environment, a lookup failure (the mapping is not found though the address is present in the network) may be caused by a legacy node that was node discovered (aka a silent node). In that case, it is an administrative decision for the 6LR to broadcast an NS(lookup) or to return an error as shown in Figure 8.

5. Requirements on the EVPN-Unaware Host

This document describes how EVPN routing can be extended to reach a Host. This section specifies the minimal EVPN-independent functionality that the Host needs to implement to obtain routing services for its addresses.

5.1. Support of 6LoWPAN ND

A host sees a prefix as not on-link (e.g., it learned that prefix in a PIO in a RA with the L flag not set) should not attempt to resolve an address within that prefix using a multicast NS(lookup). Instead, it must pass its packets to a router, preferably one that advertises that prefix in a PIO; it must register the address that it uses as source to that router to enable source address validation using [RFC8505]. It is recommended that the Host also implements [RFC8928] to prove its ownership of its addresses.

The Host is expected to request routing services from a router only if that router originates RA messages with a 6CIO that has the L, P, and E flags all set to 1 as discussed in Section 3.4, unless configured to do so. To obtain routing services for one of its addresses, the host must register the address to a router that advertises the prefix, setting the "R" and "T" flags in the EARO to 1 as discussed in Section 3.2.1 and Section 3.2.2, respectively.

This document echoes the behavior specified in [RFC9010] whereby, when the R Flag set to 1 in a NS(EARO) is not echoed in the NA(EARO), the host must understand that the route injection failed, and if the R flag is reset later in an asynchronous NA(EARO), the host must understand that routing service has failed.

The host may attach to multiple 6LRs and is expected to prefer those that provide routing services. The abstract model for this is a P2MP interface that wraps together as many P2P IP Links the host has adjacencies to 6LRs over that interface. The IPv6 address and the subnet are associated to that interface. The interface may be virtual and it may bundle multiple physical Ethernet interfaces that connect to the individual 6LRs over point to point wires, possibly via a software switch. It can also be associated to one physical interface to an external switch, either way the PI Links can be associated to sub-interface of the interface.

The Host needs to register to all the 6LRs from which it desires routing services. The multiple Address Registrations to several 6LRs should be performed in a rapid sequence, using the same EARO for the same Address. Gaps between the Address Registrations will invalidate some of the routes till the Address Registration finally shows on

those routes. The routers recognize the same (ROVR, TID) as the signal of a multihomed address and maintain all the routes. In the case of EVPN, the Ethernet Segment must also be the same. The flow for a successful multihomed registration is illustrated in Figure 13.

[RFC8505] introduces error Status values in the NA(EARO) which can be received synchronously upon an NS(EARO) or asynchronously. The Host needs to support both cases and refrain from using the address when the Status value indicates a rejection.

6. Enhancements to EVPN

This section addresses the necessary changes to EVPN formats and behavior to support address registration security per [RFC8928] and mobility per [RFC8505] while retaining interoperability with traditional nodes. With 6LR injecting not only MACs via packet sources and TLLAO options but also ROVR into mobility extended community their semantics will be somewhat extended. Specifically following issues have to be addressed:

- * The ROVR extends the semantics of the type-2 MAC advertisement via changes in MAC Mobility Extended Community in the sense that the MAC must be aligned with the ROVR and under normal circumstances only the validity of ROVR guarantees that the type-2 MAC can be allocated to the requester. A MAC validated by ROVR should take precedence over MAC addresses allocated without using it given it presents a much more trustworthy topological information (it will be called ROVR MAC in further text). EVPN nodes not supporting extensions introduced by this document will need to be led to believe that a ROVR MAC is to be preferred over any advertisement they see as long a ROVR MAC route is present. Nevertheless, primary key of NRLI is still the IP/MAC/ESI combination as defined in [RFC7432], Section 7.2 and 7.7. This implies that the same MAC (and consequently ROVR MAC) can be assigned multiple IP addresses and those represent independent NLRIs.
- * The TID field in the EARO is smaller than the mobility sequence number in [RFC7432]. To allow a ROVR MAC mobility to "win" over legacy MACs in every circumstance, signaling must be introduced that enables to distinguish a TID-generated sequence number from a legacy sequence number.
- * [RFC8505] supports IP multihoming, but does not differentiate multihoming from anycast, e.g., using the MAC address, to enable MAC address rotation. If an anycast IP address is registered with a different ROVR it will be rejected as duplicate. If it is registered with a different TID, the older sequence will be withdrawn. So the basic expectation with [RFC8505] is that the

advertisement of an anycast address is coordinated, with the same keypair known to all parties, and the same value of the TID used by all nodes (and possibly never increasing), in other words, with no concept of mobility. This specification adds a flag in EVPN that signals that the IP address is anycast and requires the local 6LBR to ignore the duplication if the same IP address is registered locally, and then to inject the NLRI with the A flag set on mobility extended community as well.

- * [RFC8928] needs the full ROVR to validate the address ownership, but the full ROVR can be too large to advertise through BGP. When an IP address is advertised through EVPN, it is REQUIRED that the EVPN Next Hop represents the address of the 6LBR of the leaf where the address was registered as well. This way, if the address is registered later on a second leaf, the 6LR in second leaf can leverage an out-of-band, i.e. via EVPN traffic carrying tunnels, EDAR/EDAC exchange with that 6LBR to validate that the ROVR in the registration is indeed the same. When that is the case, it can continue with the registration procedure and if successful, become a 6LBR for that IP address, either as a mobility event or as a multihomed registration.
- * [RFC8928] expects nodes to autoconfigure the keypair that is used to form the ROVR, in which case the IPv6 address can be locally autoconfigured with no central coordination; in that case, the ROVR only protects the ownership and enforce first-come first-serve and source address validation. But it is also possible to pre-provision the ROVR in the 6LBR and then provision the keypair in the node, e.g., in the case of a trusted server. To enable that capability in EVPN, this specification adds a flag to signal that the 6LBR that injects the address in EVPN does not provide reachability to the address. When that flag is set, the value of the TID is ignored in the mobility computation, the mapping to the MAC address is ignored, and the route to the IP address is not injected in the RIB on ROVR nodes. Non-ROVR nodes will consider the node a "honey-pot". Once the address is registered by a 6LN in the network and the according validation with the node advertising the U-bit version of the route performed, the owner will inject the route without the U-bit. A node advertising the NLRI with U-bit in its mobility extended community MUST withdraw the U-bit route once it sees a validated NLRI without the U-bit and it MAY reinject the route with the U-bit once all routes without the U-bit are withdrawn to protect the address again.

EVPN signaling is not used to carry ROVR since without challenge per [RFC8928] they do not represent any difference over using the IP/MAC combination. Instead, the full ROVR is verified upon a movement or a multi-homed advertisement using an EDAR/EDAC exchange. Additionally,

backwards compatibility could not be preserved given comparing routes based on ROVR would present a change in primary key of NLRIs which non-ROVR routers do not implement. An indication from a ROVR node that a MAC has been validated by proof of ownership is enough to convey the necessary information. Only a small hash of the ROVR is carried to speed up the identification of an address duplication.

6.1. ROVR MAC Mobility Extended Community

Extending MAC Mobility Extended Community allows to design a solution that, while backwards compatible, allows to introduce ROVR MAC as "more trusted" entities. Figure 11 presents the according extensions that will however necessitate some further explanation.

To introduce a "precedence" of ROVR MACs over normal EVPN MACs ROVR MACs are advertised to look like "sticky" MACs for non-ROVR nodes. As defined in the glossary, for simplicity reasons such nodes will be called non-ROVR nodes vs. ROVR nodes. The "sticky" bit will force non-ROVR nodes to disregard the sequence number and accept any IP/MAC route provided.

ROVR nodes MUST set the "R" flag in Mobility Extended Community to indicate that the advertisement is a ROVR MAC in case the host followed the according procedures. ROVR MACs use (instead of increasing the normal sequence number) the TID in the high bits of the sequence number field to "override" any normal MAC advertisement (further considerations will be provided in Section 6.2).

ROVR nodes MUST set the "V" flag if the address assignment passed proof of ownership per [RFC8928]. Such "validated" ROVR MAC addresses will be preferred by ROVR nodes over non validated ROVR MACs.

In case a ROVR node configures the address as "sticky" (since the sticky bit semantics have been changed to the point a ROVR cannot tell whether address is really sticky unless advertised as such by non-ROVR node) a new "X" flag called "super sticky" is introduced.

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Type=0x06      | Sub-Type=0x00 | rsv|U|A|X|V|R|S|  Reserved = 0 |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|          TID          | Reserved = 0 |          ROVR Hash          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Figure 11: Modified MAC Mobility Extended Community

Flags:

- U: Unreachable, indicating that the IP address is not reachable via that EVPN next hop, but is advertised for the purpose of protecting the value of the ROVR until a first 6LBR that can reach the address becomes available.
- A: Anycast, indicating that the IP address duplication should be ignored. When this bit is set, TID should be ignored in comparison of EVPN advertisements, i.e. all ROVR MACs at same level of validation MUST be considered having same TID.
- S: Sticky as defined in [RFC7432].
- R: ROVR Capable indicates that the advertisement is originated after processing signaling from host meeting the requirements in Section 5. This indicates a ROVR MAC.
- V: ROVR Validated indicates that the MAC passed proof of ownership per [RFC8928]. Presence of this bit implies the "R" bit being set irregardless of its value.
- X: Super Sticky indicates that the ROVR MAC is sticky and should follow procedures of sticky per [RFC7432].

Sequence Number Field:

TID: contains the ROVR MAC TID per [RFC8505]. This MUST NOT be zero, i.e. a ROVR

ROVR Hash: Hash of ROVR used to generate the according ROVR MAC. Hash is built by XOR'ing ROVR bytes in network order into the least significant byte and rotating the two bytes result after every byte by one bit to the left.

6.2. Extended ROVR MAC Procedures

In case a non-ROVR node advertises a sticky MAC by setting the "S" bit and a ROVR node sees an ROVR address registration for the same MAC it MUST follow procedures per [RFC7432].

In case a non-ROVR node advertises a sequence number larger than the one generated by TID on a ROVR node, the ROVR node SHOULD advertise a Sequence Number consisting of all bits being set to force a "roll-over" on all nodes and then fall back to advertising the TID generated sequence number again. In case a non-ROVR node persists in increasing the sequence number after that it is indication of violation of [RFC7432] on its part.

A ROVR node advertising a ROVR MAC that has not been validated and receiving same type-2 route that has been validated MUST immediately withdraw its advertisement.

A ROVR node advertising a ROVR MAC and receiving an equivalent ROVR MAC from other node with a higher TID MUST immediately withdraw its advertisement. This will allow the non-ROVR nodes to correctly interpret the sequence as MAC move despite ignoring the sequence number due to presence of "S" bit.

A ROVR node that receives a ROVR MAC with "super sticky" indication and seeing the MAC locally MUST follow analogous procedures to [RFC7432].

Multi-homing a MAC on mix of ROVR and non-ROVR nodes will lead to operational notifications since per [RFC7432] the non-ROVR node will interpret the situation as a sticky MAC that has shown up on its local interface unless an implementation is somewhat clever and understands that the presence of the same ESI on all the routes indicates that this situation does not represent a sticky MAC being moved.

7. Protocol Operations

Following section illustrates several situations and resulting signaling in EVPN from the point of view of a ROVR node.

Figure 12 illustrates the registration flow of a new address protected by [RFC8928]. The ROVR in the EARO is a Crypto-ID that derives from a public address through hashing with some other terms. The router challenges the host with a status of 5 (validation requested).

The host performs the NS again, passing the parameters that enable to build the Crypto-ID in a Crypto-ID Parameters Option (CIPO), and signing that set of parameters together with a pair of Nonce values, one from each side, in a resulting Neighbor Discovery Protocol Signature Option (NDPSO). The 6LR first verifies that the Crypto-ID can be rebuilt based on the public key, then verifies that the signature in the NDPSO was effectively performed with the associated public key. When that is the case, the registration flow can continue, else the registration is rejected with a status of 10 (Validation Failed) in the NA(EARO).

With this specification, the 6LBR communicates internally with the collocated eVPN router to inject the route in eVPN. Since the [RFC8928] validation was performed, the V flag is set. Once this is done, the local 6LBR installs a local state associated to the NCE and

becomes owner of the registration, whereas the remote leaves optionally install a remote state for the address with the indication of the 6LBR that owns the registration. The local 6LBR MUST be signalled as EVPN Next Hop for the route.

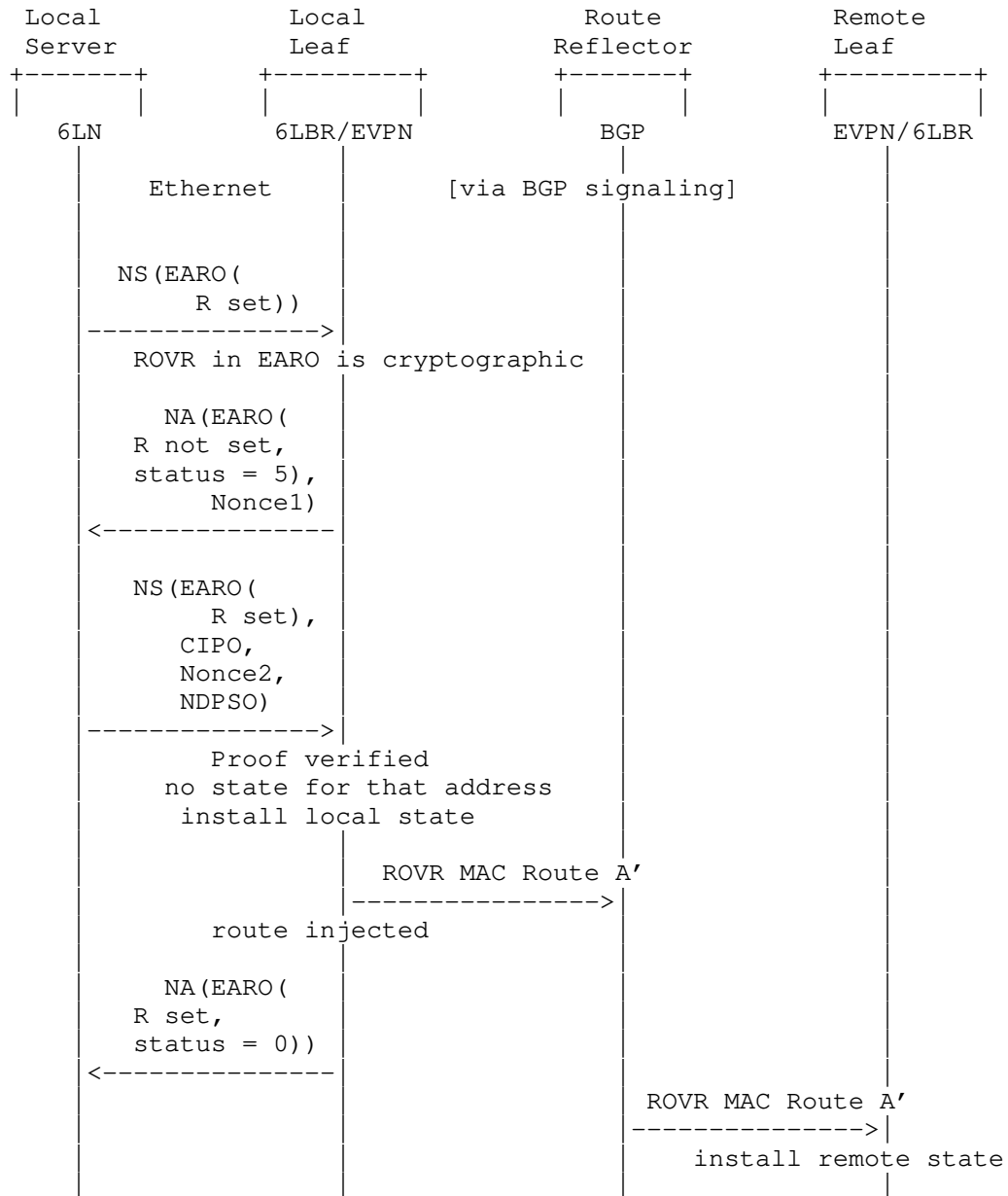


Figure 12: Host Registration

Figure 13 presents the same flow but for a multihomed address; here and in the following flows, the proof of ownership section is not shown, but its use is RECOMMENDED. The interesting piece is that when the node registers to the second 6LBR, that second 6LBR find that there is a first 6LBR that already own the registration. Using and EDAR / EDAC flow, the second 6LBR validates that the ROVR and TID are identical, in which case it accepts the registration and becomes another 6LBR owner of the registration. The result is that the 2 6LBRs are synchronized and any of the 2 can now be used, e.g., if the address is registered a third time.

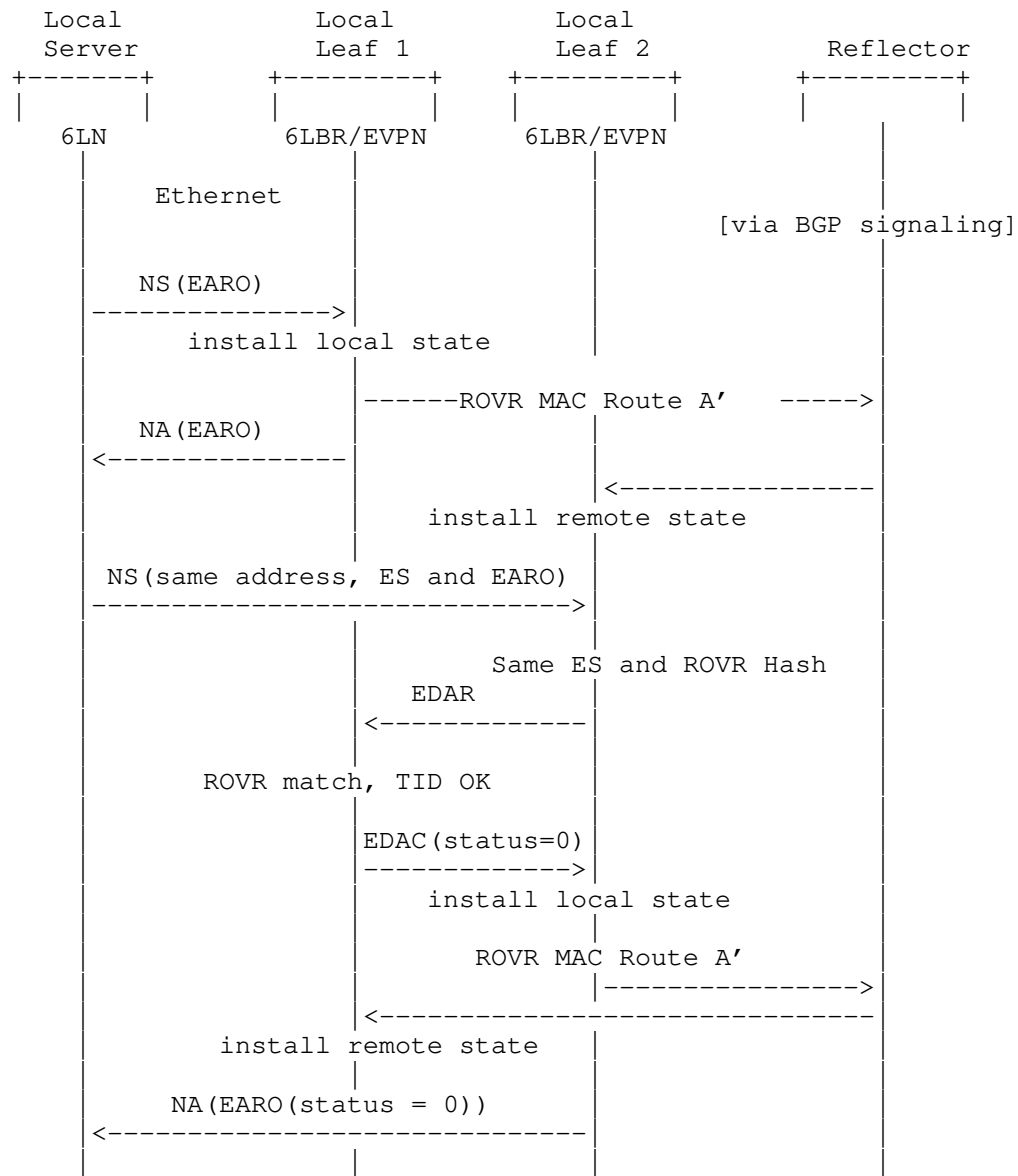


Figure 13: Multihoming

The registration is associated with a lifetime, and it must be renewed with an incremented TID. The new TID is propagated in eVPN as illustrated in Figure 14.

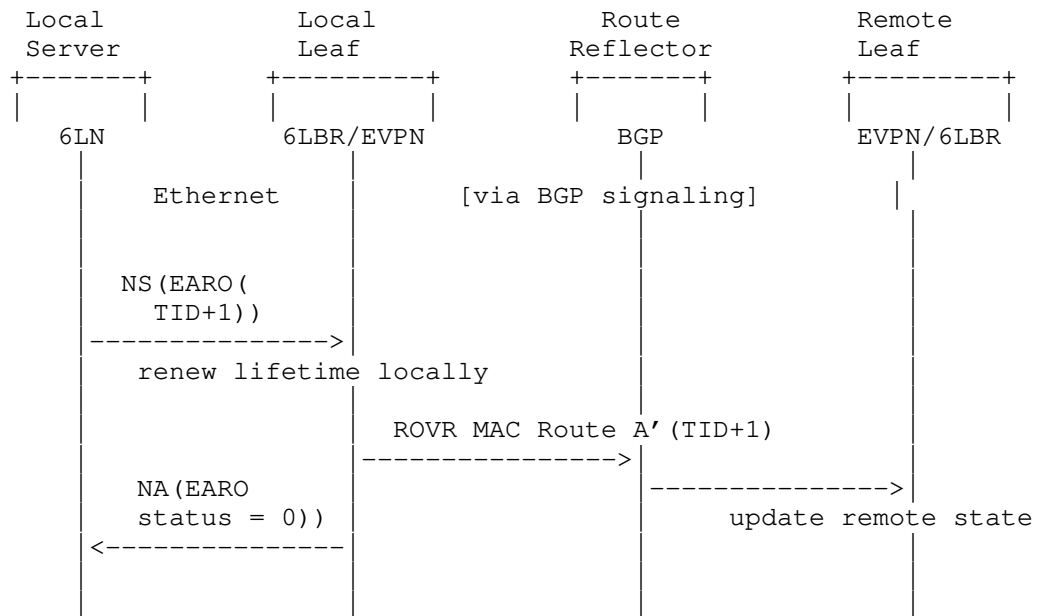


Figure 14: Host Registration Renewal

Figure 15 illustrates the case where a second host registers the same address, creating a potential address duplication situation. In most cases, the ROVR hash will be different, and the local 6LBR can reject the registration if the status is 1 (duplicate) right away.

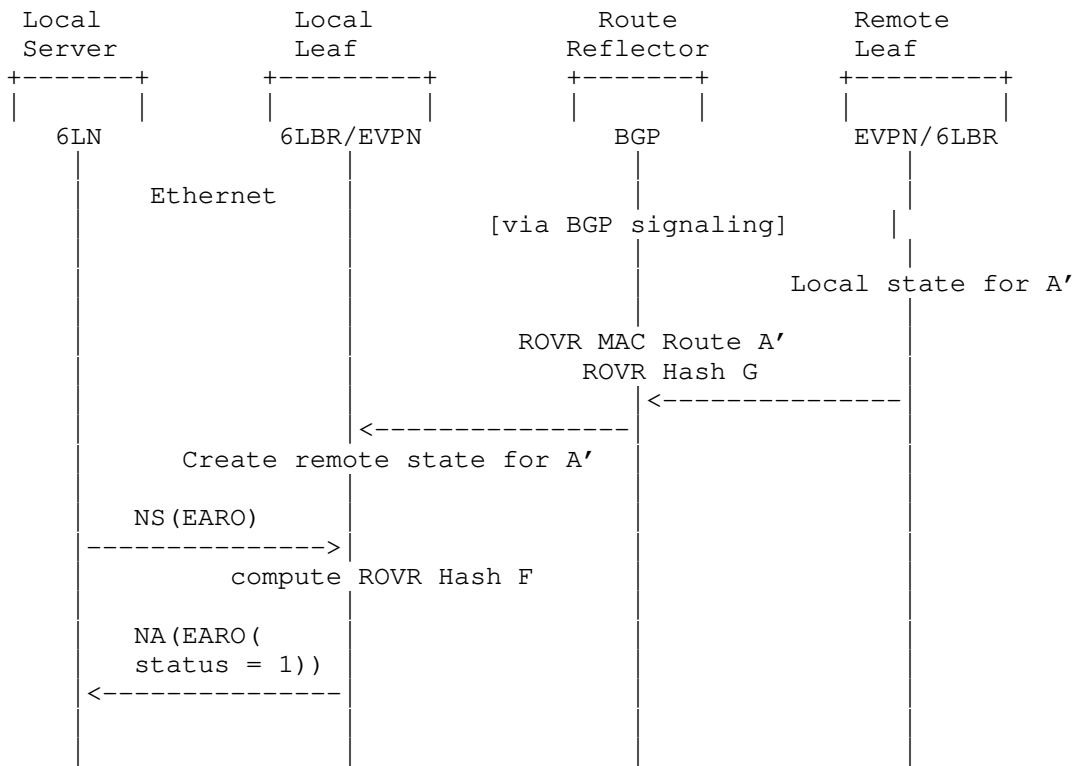


Figure 15: Duplicate Addresses

Figure 16 illustrates the case of an address duplication situation where by chance, the ROVR hashes are the same. In that case, the local 6LR checks with the 6LBR that owns the registration using an EDAR/EDAC message exchange. As opposed to the ROVR hash, the full ROVRs do not collide and the registration is also rejected.

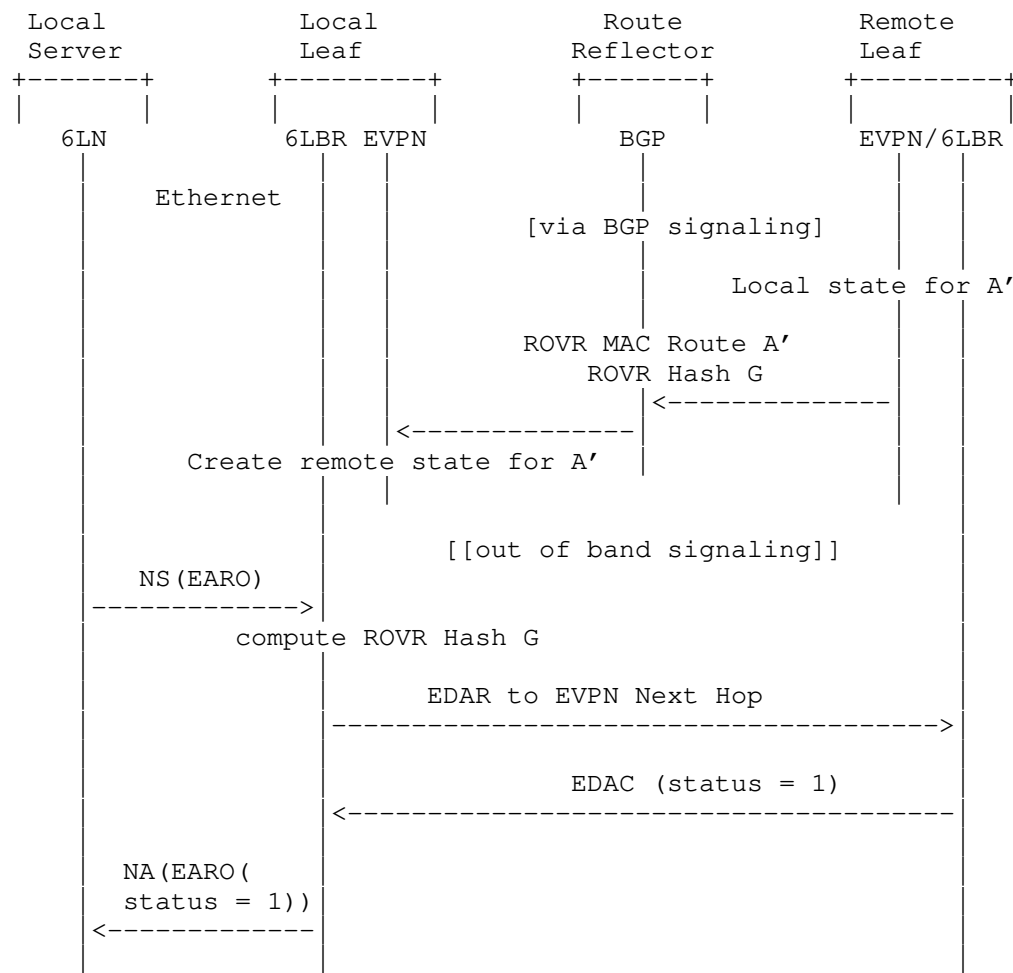


Figure 16: Duplicate Addresses, ROVR Hash Collision

Figure 17 shows a rare case where the registration has already moved elsewhere with an incremented TID when the local registration is received after being delayed in the network. In that case, the registration is rejected with a status of 3 (moved).

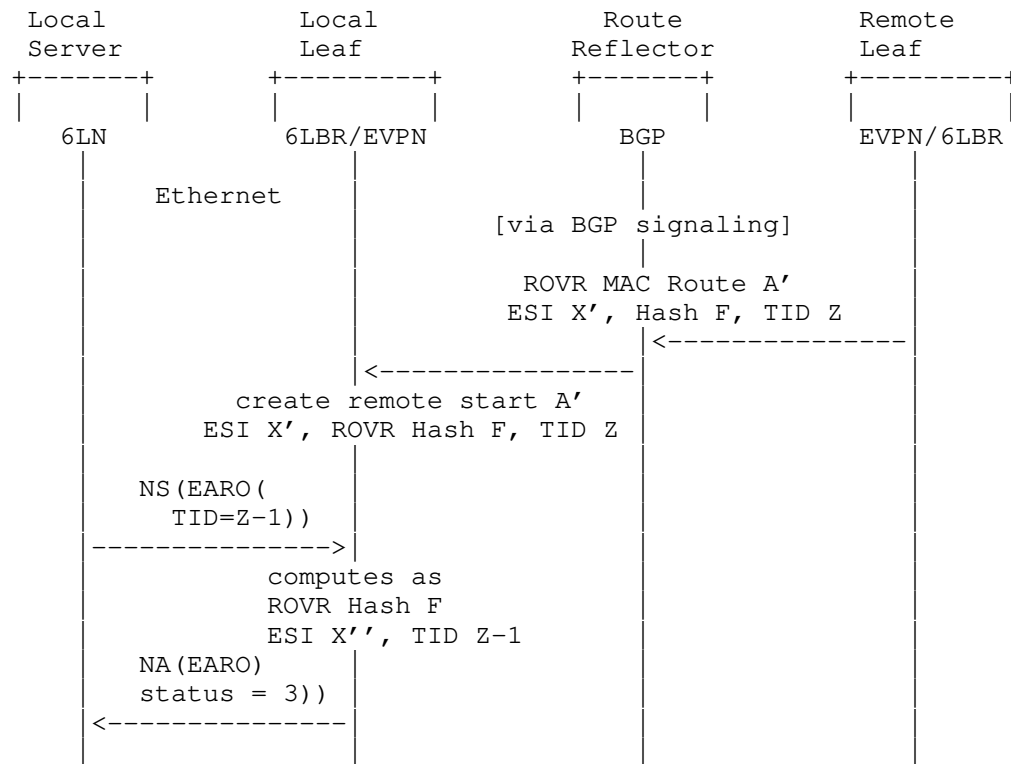


Figure 17: Address Already Moved

Address move differs from multi-homing by the ESI being different as visualized by Figure 18. In case of different ESI BGP signalling happens immediately, in case of multi-homing we can reasonably expect for the signalling to catch up on the other leg with a new, higher TID. However, since ESI matches TID doesn't matter strictly speaking and the new remote state can be installed as is. However, if 6LN is not refreshing its registration we can expect elapsed lifetime to create scenario Figure 21 over time.

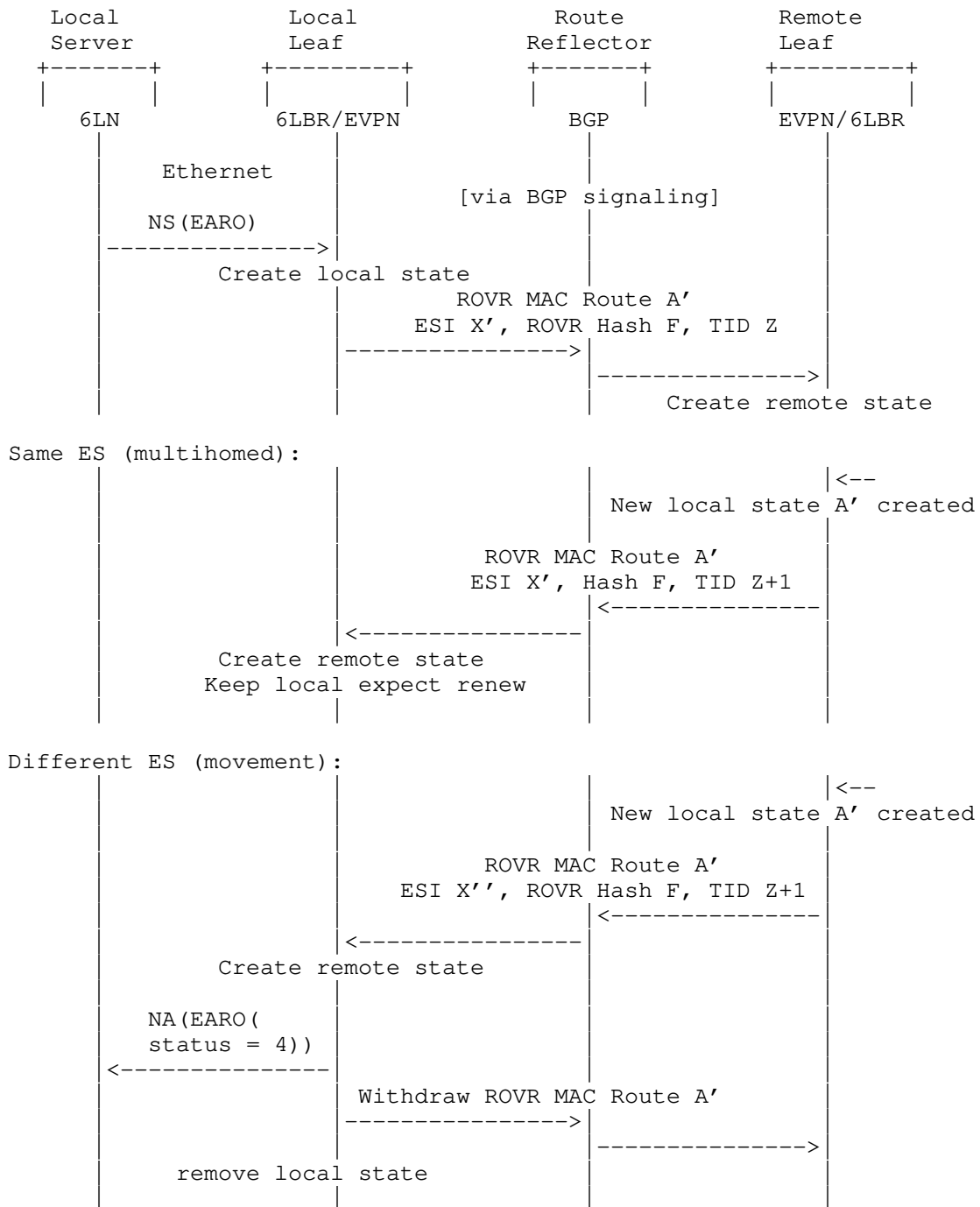


Figure 18: Address Move

The host that registered the address may cancel the registration at any time, e.g., if the address is removed from its own interface. This is done by registering with a lifetime of 0 as shown in Figure 19. The Leaf may respond with a status of 0 to indicate success, but a status of 4 (removed) is preferred for this situation.

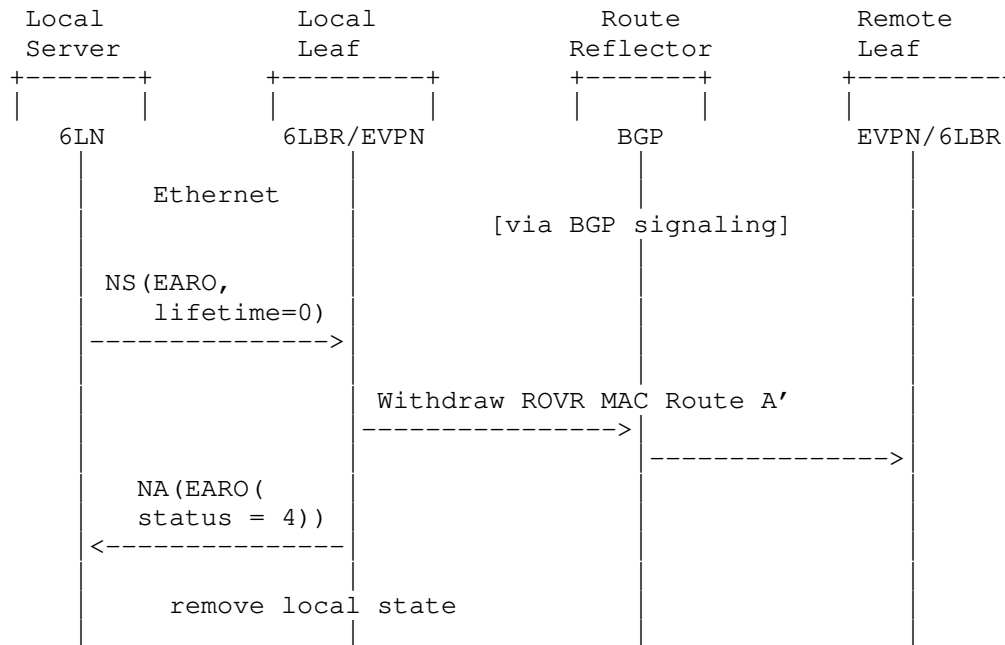


Figure 19: Address Removal

The host that registered the address may withdraw the route but maintain the NCE, e.g., in the case where it is multihomed but does not want to use one interface for the traffic back as this time. This is done by registering with the R flag set to 0 as shown in Figure 20.

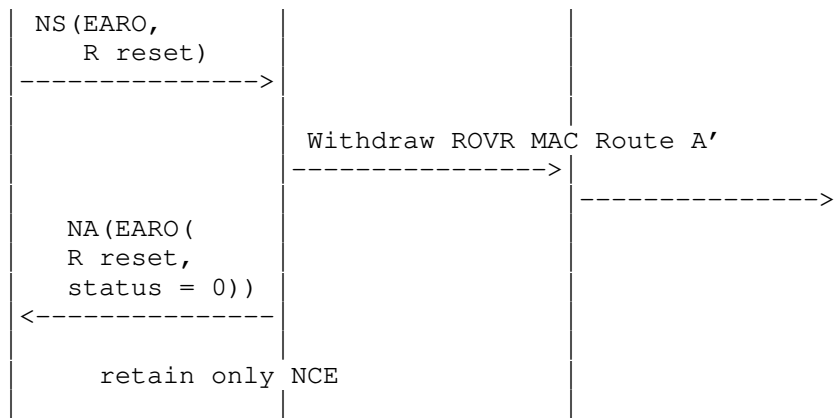


Figure 20: Route Type 2 Removal

When the lifetime elapses, the 6LBR requires the collocated eVPN router to withdraw the route.

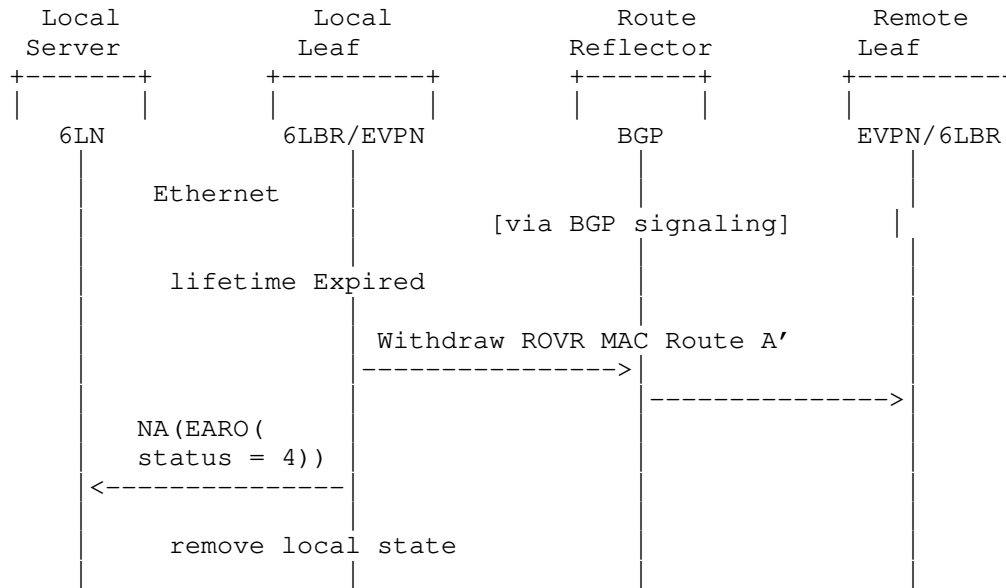


Figure 21: Lifetime Elapse

8. Security Considerations

TBD

9. IANA Considerations

10. Acknowledgments

The authors wish to thank you for reading that far. We acknowledge and express gratitude to Wen Lin, Stephane Litkowski, Eric Levy-Abegnoli, Lukas Krattiger, Jerome Tollet, and Ali Sajassi, for their help and support.

11. 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>>.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <<https://www.rfc-editor.org/info/rfc4443>>.
- [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>>.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, DOI 10.17487/RFC4862, September 2007, <<https://www.rfc-editor.org/info/rfc4862>>.
- [RFC6775] Shelby, Z., Ed., Chakrabarti, S., Nordmark, E., and C. Bormann, "Neighbor Discovery Optimization for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)", RFC 6775, DOI 10.17487/RFC6775, November 2012, <<https://www.rfc-editor.org/info/rfc6775>>.
- [RFC6085] Gundavelli, S., Townsley, M., Troan, O., and W. Dec, "Address Mapping of IPv6 Multicast Packets on Ethernet", RFC 6085, DOI 10.17487/RFC6085, January 2011, <<https://www.rfc-editor.org/info/rfc6085>>.
- [RFC7400] Bormann, C., "6LoWPAN-GHC: Generic Header Compression for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)", RFC 7400, DOI 10.17487/RFC7400, November 2014, <<https://www.rfc-editor.org/info/rfc7400>>.

- [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>>.
- [RFC8505] Thubert, P., Ed., Nordmark, E., Chakrabarti, S., and C. Perkins, "Registration Extensions for IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Neighbor Discovery", RFC 8505, DOI 10.17487/RFC8505, November 2018, <<https://www.rfc-editor.org/info/rfc8505>>.
- [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>>.
- [RFC8928] Thubert, P., Ed., Sarikaya, B., Sethi, M., and R. Struik, "Address-Protected Neighbor Discovery for Low-Power and Lossy Networks", RFC 8928, DOI 10.17487/RFC8928, November 2020, <<https://www.rfc-editor.org/info/rfc8928>>.
- [UNICAST-LOOKUP]
Thubert, P. and E. Levy-Abegnoli, "IPv6 Neighbor Discovery Unicast Lookup", Work in Progress, Internet-Draft, draft-thubert-6lo-unicast-lookup-00, 25 January 2019, <<https://datatracker.ietf.org/doc/html/draft-thubert-6lo-unicast-lookup-00>>.

12. Informative References

- [RFC6550] Winter, T., Ed., Thubert, P., Ed., Brandt, A., Hui, J., Kelsey, R., Levis, P., Pister, K., Struik, R., Vasseur, JP., and R. Alexander, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks", RFC 6550, DOI 10.17487/RFC6550, March 2012, <<https://www.rfc-editor.org/info/rfc6550>>.
- [RFC8929] Thubert, P., Ed., Perkins, C.E., and E. Levy-Abegnoli, "IPv6 Backbone Router", RFC 8929, DOI 10.17487/RFC8929, November 2020, <<https://www.rfc-editor.org/info/rfc8929>>.

- [RFC9010] Thubert, P., Ed. and M. Richardson, "Routing for RPL (Routing Protocol for Low-Power and Lossy Networks) Leaves", RFC 9010, DOI 10.17487/RFC9010, April 2021, <<https://www.rfc-editor.org/info/rfc9010>>.
- [RIFT] Przygienda, T., Sharma, A., Thubert, P., Rijsman, B., and D. Afanasiev, "RIFT: Routing in Fat Trees", Work in Progress, Internet-Draft, draft-ietf-rift-rift-12, 26 May 2020, <<https://datatracker.ietf.org/doc/html/draft-ietf-rift-rift-12>>.
- [IANA-EARO-STATUS] IANA, "Address Registration Option Status Values", <<https://www.iana.org/assignments/icmpv6-parameters/icmpv6-parameters.xhtml#address-registration>>.

Authors' Addresses

Pascal Thubert (editor)
Cisco Systems, Inc
France

Phone: +33 497 23 26 34
Email: pthubert@cisco.com

Tony Przygienda
Juniper Networks, Inc
Switzerland

Email: prz@juniper.net

Jeff Tantsura
Microsoft

Email: jefftant.ietf@gmail.com

BESS WG
Internet-Draft
Intended status: Standards Track
Expires: 27 April 2022

Y. Wang
Q. Niu
ZTE Corporation
24 October 2021

Distributed Bump-in-the-wire Use Case
draft-wang-bess-evpn-distributed-bump-in-the-wire-01

Abstract

The Bump-in-the-wire use-case of Section 4.3 of [RFC9136] is a centerlized inter-subnet forwarding solution. The centerlized inter-subnet forwarding burdens the DGWs with the L3 traffics among different subnets inside the same DC.

This draft extends the Bump-in-the-wire use-case of Section 4.3 of [RFC9136] in order to achieve a distributed inter-subnet forwarding solution.

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 27 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
1.1. Terminology and Acronyms	4
2. Problem Statement	5
2.1. Centerlized Inter-subnet Forwarding	5
2.2. RT-1 Confliction among Multiple Bump-in-the-wires	6
3. Solutions	8
3.1. Supplementary BD for Bump-in-the-wire	8
3.2. Constructing IP Prefix Advertisement Route	9
3.3. ACI-specific Supplementary Overlay Index Extended Community	11
3.4. Determining the Aliasing Pathes for RT-5E	13
3.5. Other Considerations	13
4. IANA Considerations	14
5. Security Considerations	14
6. References	14
6.1. Normative References	14
6.2. Informative References	15
Authors' Addresses	15

1. Introduction

As shown in Figure 1, the Bump-in-the-wire use-case of Section 4.3 of [RFC9136] is a centerlized inter-subnet forwarding solution. The centerlized inter-subnet forwarding burdens the DGWs with the L3 traffics among different subnets (e.g. SN1 and H3 of Figure 2) inside the same DC.

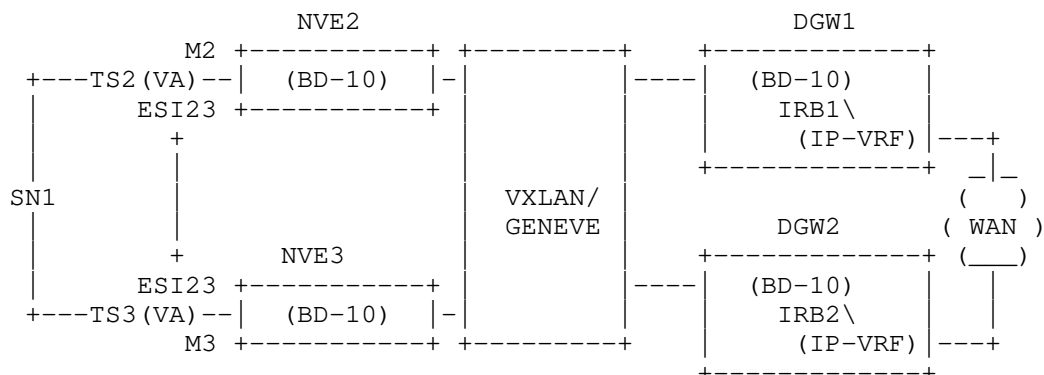


Figure 1: RFC9136's Figure 7

When a SBD is added (see Figure 4) for the IP-VRF instance, using this SBD and its SBD IRB, we can extend the Bump-in-the-wire use case to form a distributed inter-subnet forwarding solution which will not burden the DGWs with the L3 traffics among different subnets inside the same DC.

But when multiple Bump-in-the-wires are integrated into the same IP-VRF (as shown in Figure 3), the above extension is not enough, the details are described in Section 2.2, thus some further extensions are introduced to solve that problem.

The RT-5 route that specifies an ESI as overlay index is first defined in Section 4.3 of [RFC9136], where the Bump-in-the-wire use case (which is called the first type RT-5E usage) is also defined there.

Note that the RT-5E routes (which are called the second type RT-5E usage) of Section 4.3.2 of [I-D.wang-bess-evpn-arp-nd-synch-without-irb] and Section 1.3 of [I-D.sajassi-bess-evpn-ip-aliasing] are different from these RT-5E routes of Bump-in-the-wire use case in the following factors:

- * Source MAC - The ethernet header can not be absent in the first type usage even if the data plane is MPLS. The source MAC MUST be set to the MAC address of the IRB interface of BD-10 in Bump-in-the-wire usecase. But in the second type usage the ethernet header can be absent if the data plane is MPLS.
- * Recursive Resolution - The recursive resolution of the first type usage are done in the context of a BD, But the recursive resolution of the second type usage are done in the context of a IP-VRF.

- * EVPN label - The EVPN label of the corresponding RT-1 per EVI route of the first type usage is a MPLS label which identifies a BD, But the EVPN label of the corresponding RT-1 per EVI route of the second type usage is a MPLS label which identifies an IP-VRF.
- * ESI - The ESI of the first type usage is attached to a BD, But ESIs of the second type usage are attached to IP-VRFs.

The Bump-in-the-wire use case is a special form of EVPN IRB use case, that's why its corresponding RT-1 per EVI routes are resolved in BD context.

1.1. Terminology and Acronyms

Most of the acronyms and terms used in this documents comes from [RFC9136] and [I-D.wang-bess-evpn-ether-tag-id-usage] except for the following:

- * VRF AC - An Attachment Circuit (AC) that attaches a CE to an IP-VRF but is not an IRB interface.
- * VRF Interface - An IRB interface or a VRF-AC or an IRC interface. Note that a VRF interface will be bound to the routing space of an IP-VRF.
- * L3 EVI - An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN which contains VRF ACs and maybe contains IRB interfaces or IRC interfaces.
- * RT-1 per EVI - Ethernet Auto-Discovery route per EVI, and the EVI here is an IP-VRF. Note that the Ethernet Tag ID of an RT-1 per EVI route may be not zero.
- * IP-AD/ES - Ethernet Auto-Discovery route per ES, and the EVI for one of its route targets is an IP-VRF.
- * RMAC - Router's MAC, which is signaled in the Router's MAC extended community.
- * ESI Overlay Index - ESI as overlay index.
- * ET-ID - Ethernet Tag ID, it is also called ETI for short in this document.
- * RT-5E - An EVPN Prefix Advertisement Route with a non-reserved ESI as its overlay index (the ESI-as-Overlay-Index-style RT-5).

- * CE-BGP - The BGP session between PE and CE. Note that CE-BGP route doesn't have a RD or Route-Target.
- * CE-Prefix - An IP Prefixes behind a CE is called as that CE's CE-Prefix.
- * ETI-Agnostic BD - A Broadcast Domain (BD) whose data packets can be received along with any Ethernet Tag ID (ETI). Note that a broadcast domain of an L2 EVI of VLAN-aware bundle service interface is a good example of an ETI-Specific BD.
- * ETI-Specific BD - A Broadcast Domain (BD) whose data packets are expected to be received along with a normalized Ethernet Tag ID (ETI). Note that a broadcast domain of an L2 EVI of VLAN-bundle or VLAN-based service interface is a good example of an ETI-Agnostic BD.
- * BDI-Specific EADR - When the <ESI, BD> uses BDI-Specific Ethernet Auto-discovery mode, the only Ethernet A-D per EVI route of that <ESI, BD> is called as a BDI-Specific EADR in this draft.
- * ACI-Specific EADR - When the <ESI, BD> uses ACI-Specific Ethernet Auto-discovery mode, the Ethernet A-D per EVI routes of that <ESI, BD> are called as ACI-Specific EADRs in this draft.

2. Problem Statement

2.1. Centerlized Inter-subnet Forwarding

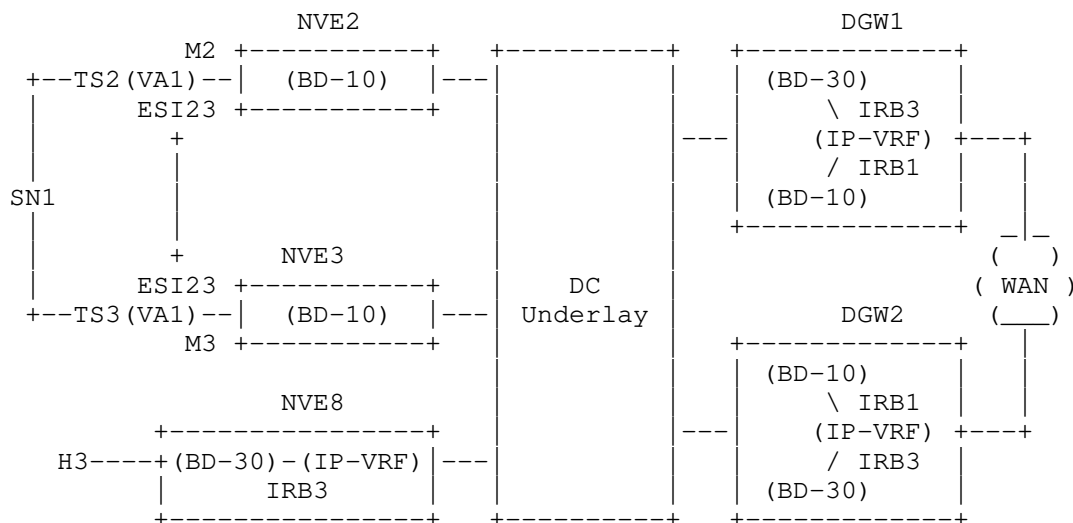


Figure 2: Centerlized Bump-in-the-wire Use Case

As shown in Figure 2, SN1 and H3 are both internal hosts of the same DC. But the communication between them have to pass through a DGW, that's why the DGWs will be burdened with inter-subnet forwarding of the internal hosts.

The Section 4.3 of [RFC9136] defined the Bump-in-the-wire use-case, where a style (which is called as RT-5E in this draft) of RT-5 routes (whose overlay index is a non-zero ESI), is used to advertise the IP prefix of subnet SN1 (see Figure 3). The RT-5E routes (whose IP prefix is SN1, and ESI is ESI23) of Section 4.3 of [RFC9136] is called as RT5E_SN1 in this draft. And the RT-1 routes (whose ESI is ESI23) corresponding to the RT5E_SN1 is called as RT1_ESI23 in this draft.

Note that when DGW1 or DGW2 receives RT5E_SN1, it should know (before the recursive resolution) that RT5E_SN1's ESI (ESI23) should be resolved in the context of BD-10, not in BD-30 (whether BD-30 is another Bump-in-the-wire BD or not). Because of RT5E_SN1's Route target (which identifies BD-10), DGW1 can know that before the recursive resolution.

2.2. RT-1 Confliction among Multiple Bump-in-the-wires

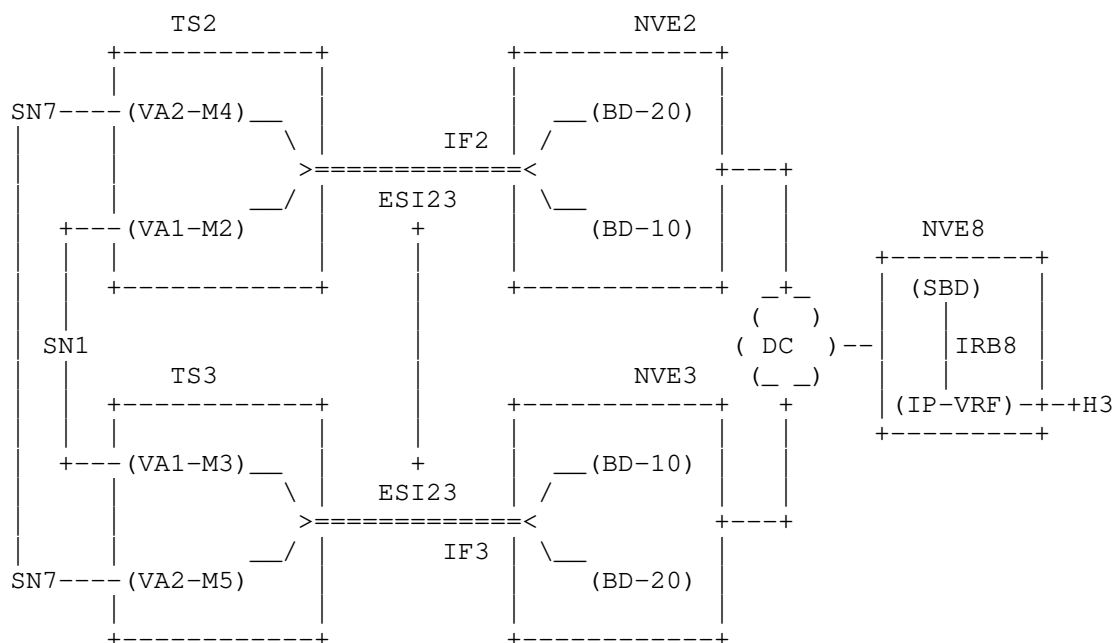


Figure 3: ET-ID Confliction of Bump-in-the-wire

This network is another view of a part of Figure 4, and it is similar to Section 4.3 of [RFC9136] with a few notable exceptions as below:

The NVE2, NVE3, BD-10, ESI23, TS2, TS3 and SN1 here is the NVE2, NVE3, BD-10, ESI23, TS2, TS3 and SN1 there (Section 4.3 of [RFC9136]). The VA1 here is the Virtual Appliance (whose VA-MAC is M2/M3 on TS2/TS3) there. The NVE8 here is the DGW1 there. The IRB8 here takes the place of the IRB1 there.

But here we have another Bump-in-the-wire instance for Virtual Appliance VA2, which are attached to another Broadcast Domain BD-20. Both BD-10 and BD-20 are integrated into the same IP-VRF by DGW1. But the subnet SN1 can only be reached through BD-10, while the subnet SN7 can only be reached through BD-20.

RT5E_SN1 (whose route-target identifying BD-10) is imported into the BD-10 at first, although it can be imported into the IP-VRF following BD-10's IRB interface, RT5E_SN1 will not be imported into the IP-VRF on other PEs which don't have an instance of BD-10. Thus such PEs are precluded from connecting to the hosts of SN1 by such rules.

Note that both BD-10 and BD-20 are L2 EVIs of VLAN-based Service Interfaces.

The solution for this problem is described in Section 3.5.

3. Solutions

3.1. Supplementary BD for Bump-in-the-wire

As shown in Figure 4, the SN1, BD-10, IP-VRF are the same as Figure 2, except that the TS2, TS3 and ESI23 are not shown in Figure 4, but they are still there unchanged. Then we add a SBD for the IP-VRF instance, and each SBD will be configured with an IRB interface (which is called its SBD IRB). Using this SBD and its SBD IRB, we can extend the Bump-in-the-wire use case to form a distributed inter-subnet forwarding solution which will not burden the DGWs with the L3 traffics among different subnets inside the same DC.

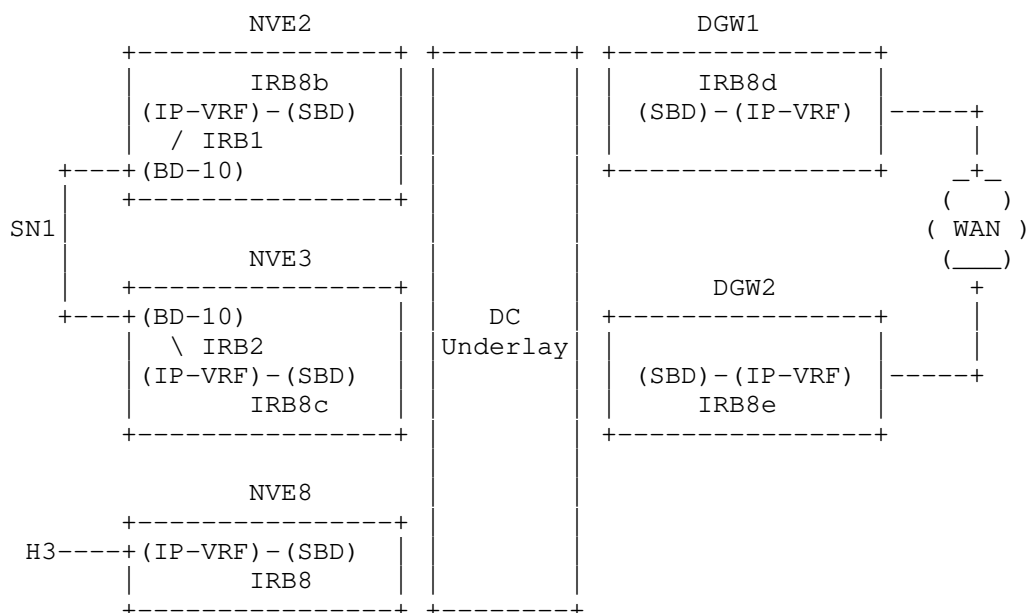


Figure 4: Distributed Bump-in-the-wire Use Case

The RT-5 route (say RT5E_SN1) advertised by NVE2/NVE3 for SN1 is the same as Section 4.3 of [RFC9136] except for the following notable differences:

- * The route-targets of RT5E_SN1 is set to the export-RT of the SBD.
- * The RT-1 route of ESI23 MUST be advertised both for BD-10 and the

SBD, when they are advertised for the SBD, the EVPN label of the RT-1 per EVI route should be set to the EVPN label of the BD-10, as if it is advertised for BD-10.

Note that when it is advertised for the SBD, it may use different RD than it is advertised for BD-10.

- * In order to process the RT5E_SN1 properly, the DGW1 and DGW2 don't have to change its behavior of Section 4.3 of [RFC9136]. But the configurations of DGW1 and DGW2 must be changed, because that the BD-10 is removed and the SBD takes its place.

Note that to the RT5E_SN1 route, the NVE8 is actually no different from DGW1 and DGW2. NVE8 is not a DC gateway, but whether NVE8 is a DC gateway is not aware by NVE1 and NVE2.

3.2. Constructing IP Prefix Advertisement Route

The RT5E_SN1 is constructed following Section 4.3 of [RFC9136] except for the following differences:

- * Route target and RD
The route target of RT5E_SN1 MUST be set to the route-target which identifies the SBD. In other words, RT5E_SN1 is advertise for the SBD, or we can see RT5E_SN1 is advertised in the context of the SBD.

The RD of RT5E_SN1 can be set to the RD of SBD too.

- * ESI and ET-ID

No matter whether BD-10 is an ETI-agnostic BD or ETI-specific BD, it will be enough to configure the SBD as an ETI-agnostic BD. But the Ethernet Tag ID of the Ethernet A-D per EVI routes of the SBD may be set to non-reserved ET-IDs.

When an CE-prefix of a Bump-in-the-wire instance is advertised by a RT-5E route, The RT-5E route is advertised in the SBD's context. The RT-5E route's ESI MUST be determined by the CE-prefix's VA MAC (which will be known by policy). Take SN1 of Figure 4 for example, by policy, we can know that the VA MAC M1 is in BD-10, then we can know that VA MAC M1 is learnt over <ESI23, BD-10>, so the ESI of RT5E_SN1 should be set to ESI23.

If BD-10 is an ETI-agnostic BD (e.g. BD-10 is of VLAN-based service interface), the ET-ID of RT5E_SN1 MUST be set to 0. If BD-10 is an ETI-specific BD (e.g. BD-10 is of VLAN-aware bundle service interface), the ET-ID of RT5E_SN1 MUST be set to the BD-ID of BD-10 (even if the SBD is ETI-agnostic).

Note that the ET-ID of RT5E_SN1 is not used to resolve (as described in Section 3.4) RT5E_SN1's ESI overlay index to a proper Ethernet A-D per EVI route.

* ACI-Specific Supplementary Overlay Index

When an IP Prefix Advertisement is advertised, The ACI-Specific Supplementary Overlay Index (SOI) extended community is always recommended to be carried along with it, if it is not clear that whether there will be conflictions among Ethernet A-D per EVI routes inside the SBD in the future.

Note that the ACI-Specific SOI here is not used to isolate IP address spaces. It is just used to resolve (as described in Section 3.4) RT5E_SN1's ESI overlay index to a proper Ethernet A-D per EVI route.

ACI-specific Overlay Index extended community should be advertised along with the RT-5E routes. Thus the ET-ID of these RT-5E routes can be set to zero if BD-10 and BD-20 are ETI-agnostic BDs.

Note that the combination of <ESI, SOI> will be used to select the corresponding RT-1 per EVI routes (in SBD) for these RT-5E routes on other PEs.

Note that in the data plane, the EVPN label that is encapsulated by NVE8 for NVE2 or NVE3 will be a label that identifies BD-10. So when BD-10 is an ETI-Specific BD, the ET-ID of RT5E_SN1 MUST be encapsulated into the ethernet header of the data packets. Otherwise such data packets won't be received by BD-10 (of NVE2 or NVE3).

3.3. ACI-specific Supplementary Overlay Index Extended Community

A new EVPN BGP Extended Community called Supplementary Overlay Index is introduced. This new extended community is a transitive extended community with the Type field of 0x06 (EVPN) and the Sub-Type of TBD. It is advertised along with EVPN MAC/IP Advertisement Route (Route Type 2) per [RFC7432] in ACI-Sepecific Ethernet Auto-Discovery mode. It may also be advertised along with EVPN Prefix Advertisement Route (Route Type 5) as per [RFC9136]. Generically speaking, the new extended community must be attached to any routes which are leant over an <ESI, EVI> of ACI-specific Ethernet Auto-Discovery.

The Supplementary Overlay Index Extended Community is encoded as an 8-octet value as follows:

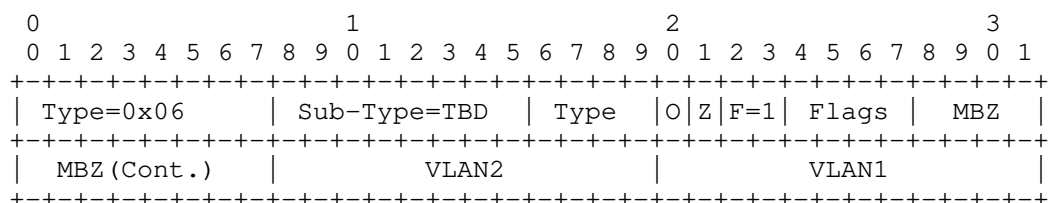


Figure 5: Supplementary Overlay Index Extended Community

- o F: Format Indicator, its value is always 1 in this draft. Other values are reserved.
- o Type: .
- * 0: VLAN-based AC-ID.

No.	Use Cases	Type	VLAN2	VLAN1	MBZ
1	untag	type 0	0	0	0
2	default	type 0	0	FFF	0
3	dot1q	type 0	0	E	0
4	QinQ	type 0	E	I	0

Table 1: VLAN-based AOIs

Notes:

E : That field is the External VLAN of the AC.

I : That field is the Internal VLAN of the AC.

0 : The tag corresponding to that field is absent.

FFF : The AC is the default subinterface (Section 3.3) of the corresponding ES.
untag : An untagged subinterface should be matched by that format.
default : A default subinterface should be matched by that format. When the AC is a default subinterface, it will match all the remaining VLAN-tags (which are left over by other subinterfaces) on its main-interface.
dot1q : A dot1q subinterface should be matched by that format.
QinQ : A QinQ subinterface should be matched by that format.
* 1-15: Reserved.

- o O Flag: Overlay Index Flag, this extended community is used as overlay index.

When type field is 0-1: For ACI-Specific Ethernet auto-discovery mode, when it is carried along with a RT-2 route, the O Flag should be set to 1, For BDI-Specific Ethernet auto-discovery, when it is carried along with a RT-2 route, the O Flag should be set to 0.

When the O Flag is set to 1, this AC-ID is also called as AOI (ACI-Specific Overlay Index), and the <ESI, AOI> of that RT-2R or RT-5E should be used to determine ECMP pathes. At the same time, the AOI should also be used like Attachment Circuit ID Extended Community too.

Note that only the lowest 8 bits of MBZ field should be used to select RT-1 per EVI routes. <lowest 8 bits of MBZ, VLAN2, VLAN1> of a type-0 AOI forms an Ethernet Tag ID of an ACI-Specific EADR.

- o Z Flag: Must be zero. Reserved for future use, the receiver should ignore this extended community if Z flag is not zero at now.
- o Flags: Reserved for future use. it is set to 0 on advertising, and ignored on receiving.

Note that although this extended community is similar to the AC-ID extended community (as per [I-D.sajassi-bess-evpn-ac-aware-bundling]), we can assume that they may be of different Sub-Types because that they have different behaviors.

3.4. Determining the Aliasing Pathes for RT-5E

No matter whether a RT-5 route is constructed following Section 4.3 of [RFC9136] or Section 3.2 of this draft, the RT-1 per EVI routes corresponding to that RT-5E route will be resolved in the context of a BD, not in an IP-VRF.

When resolving corresponding RT-1 per EVI routes for a RT-5E route, the AOI (ACI-specific SOI) Extended Community of the RT-5E route can be used.

Note that when the RT-5E's AOI is Y ($Y \neq 0$), the ET-IDs of the selected Ethernet A-D per EVI routes (of that RT-5E) should be all Y.

Note that when the RT-5E's ET-ID is not 0, and an AOI is advertised along with the RT-5E, the Ethernet A-D per EVI routes of that RT-5E should be selected according to the <ESI,AOI>.

Note that when a data packet is load-balanced according to <ESI,AOI>, in Bump-in-the-wire use case, it is the RT-5E's ET-ID which should be encapsulated into the data packet (as 802.1q Tag), not the AOI.

Note that [I-D.sajassi-bess-evpn-ac-aware-bundling] requires the Presence of Attachment Circuit ID Extended Community MUST be ignored by non multihoming PEs. It requires the remote PE (non-multihome PE, e.g. PE3) MUST process MAC route as defined in [RFC7432]. But the AOI of this case should be used to select ETI-Specific EADRs. This is non-compatible with the Attachment Circuit Extended Community, thus the new ACI-Specific Overlay Index Extended Community is defined.

3.5. Other Considerations

We can assume that maybe neither BD-10 nor BD-20 will be configured on NVE8, as illustrated in Figure 4. In such case, we assume that a SBD (Supplementary BD) can be provisioned on NVE8.

The SBD is similar to the combination of the SBD of Section 4.4.3 of [RFC9136] and the BD-10 of Section 4.3 of [RFC9136], except for the following factors:

The RT-1 per EVI routes advertised for SBD is originated from the BD-10. and the SBD don't have to advertise any EVPN routes (e.g. IMET route) of its own. because there are no hosts (even the IP address of SBD IRB will not be provisioned in this case) in the SBD.

Note that DGWs will advertise their own IP prefixes using their own L3 EVPN label and route-targets. They don't have to expect any data packets to be received from such SBD.

The route advertisement behavior of NVE2 and NVE3 should also be changed:

- * When BD-10 advertised a RT-1 per EVI route RT1a, another RT-1 per EVI route RT1b (which is the mirroring of RT1a) should be advertised for the SBD. Although RT1b is advertised for the SBD, RT1b's EVPN label should be set to BD-10's EVPN label, not the SBD's EVPN label. RT1b's ET-ID MUST be set to the AC-ID of the AC corresponding to RT1a.

Otherwise the RT-1 per EVI routes for BD-10 and BD-20 will conflict with each other, because that both BD-10 and BD-20 are of VLAN-based Service Interface.

- * The MAC addresses of IRB interface of each Bump-in-the-wire BD (e.g. BD-10 and BD-20) should be the same as the SBD IRB interface of the same L3 EVI, otherwise the source MAC may be not expected to be learnt by the CE-side L2 switches.

4. IANA Considerations

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

5. Security Considerations

TBD.

6. References

6.1. Normative References

[I-D.sajassi-bess-evpn-ac-aware-bundling]

Sajassi, A., Brissette, P., Mishra, M., Thoria, S., Rabadan, J., and J. Drake, "AC-Aware Bundling Service Interface in EVPN", Work in Progress, Internet-Draft, draft-sajassi-bess-evpn-ac-aware-bundling-04, 11 July 2021, <<https://datatracker.ietf.org/doc/html/draft-sajassi-bess-evpn-ac-aware-bundling-04>>.

- [I-D.sajassi-bess-evpn-ip-aliasing]
Sajassi, A., Badoni, G., Warade, P., Pasupula, S., Drake, J., and J. Rabadan, "EVPN Support for L3 Fast Convergence and Aliasing/Backup Path", Work in Progress, Internet-Draft, draft-sajassi-bess-evpn-ip-aliasing-02, 8 June 2021, <<https://datatracker.ietf.org/doc/html/draft-sajassi-bess-evpn-ip-aliasing-02>>.
- [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>>.
- [RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <<https://www.rfc-editor.org/info/rfc9135>>.
- [RFC9136] Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021, <<https://www.rfc-editor.org/info/rfc9136>>.

6.2. Informative References

- [I-D.wang-bess-evpn-arp-nd-synch-without-irb]
Wang, Y. and Z. Zhang, "ARP/ND Synching And IP Aliasing without IRB", Work in Progress, Internet-Draft, draft-wang-bess-evpn-arp-nd-synch-without-irb-08, 1 September 2021, <<https://datatracker.ietf.org/doc/html/draft-wang-bess-evpn-arp-nd-synch-without-irb-08>>.
- [I-D.wang-bess-evpn-ether-tag-id-usage]
Wang, Y., "Ethernet Tag ID Usage Update for Ethernet A-D per EVI Route", Work in Progress, Internet-Draft, draft-wang-bess-evpn-ether-tag-id-usage-03, 26 August 2021, <<https://datatracker.ietf.org/doc/html/draft-wang-bess-evpn-ether-tag-id-usage-03>>.
- [I-D.wz-bess-evpn-vpws-as-vrf-ac]
Wang, Y. and Z. Zhang, "EVPN VPWS as VRF Attachment Circuit", Work in Progress, Internet-Draft, draft-wz-bess-evpn-vpws-as-vrf-ac-02, 28 August 2021, <<https://datatracker.ietf.org/doc/html/draft-wz-bess-evpn-vpws-as-vrf-ac-02>>.

Authors' Addresses

Yubao Wang
ZTE Corporation
No.68 of Zijinghua Road, Yuhuatai Distinct
Nanjing
China

Email: wang.yubao2@zte.com.cn

Qibo Niu
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: niu.qibo@zte.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 18 October 2022

H. Wang
J. Dong
Huawei
G. Mirsky
Ericsson
Y. Huang
Huawei
16 April 2022

Advertising S-BFD Discriminators in BGP
draft-wang-bess-sbfd-discriminator-02

Abstract

Seamless Bidirectional Forwarding Detection (S-BFD) is a simplified BFD mechanism. It eliminates most negotiation aspects and provides advantages such as fast configuration injection. S-BFD is especially useful in multi-homing PE scenarios and reduces resource overheads on the dual-homing PEs. Although S-BFD is simpler than BFD, a large number of manual configurations are required when there are a large number of PEs.

This document provides the mechanism of distributing S-BFD discriminators with VPN service routes, which simplifies S-BFD deployment for VPN services.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 18 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Terminology	3
3. Scenarios	4
3.1. EVPN Layer 3 Service Over SRv6 BE Use Case	4
3.2. EVPN Layer 3 Service Over SPv6 Policy Use Case	5
4. Procedure	6
4.1. BGP Encoding	6
4.2. Router Procedure	8
4.2.1. Egress Node Process	8
4.2.2. Transit Node Process	9
4.2.3. Ingress Node Process	9
5. Error handling	9
6. IANA Considerations	10
7. Security Considerations	10
8. Acknowledgements	11
9. References	11
9.1. Normative References	11
9.2. References	11
Authors' Addresses	11

1. Introduction

[RFC7880] defines the Seamless Bidirectional Forwarding Detection (S-BFD) mechanism. S-BFD is a simplified mechanism for using BFD with a large proportion of negotiation aspects eliminated, thus providing benefits such as quick provisioning, as well as improved control and flexibility for network nodes initiating path monitoring. Currently, S-BFD can be used to simplify the service deployment.

During network construction, carriers usually deploy active and standby nodes to improve network reliability. In this way, when a single node is faulty, a protection switchover can be performed quickly. To accelerate fault detection, BFD is generally used. BFD sessions must be deployed on both ends of the BFD session, which occupies resources on both ends of the PE.

[RFC7880] defines Seamless Bidirectional Forwarding Detection (S-BFD), a simplified mechanism for using BFD with a large proportion of negotiation aspects eliminated, thus providing benefits such as quick provisioning, as well as improved control and flexibility for network nodes initiating path monitoring. This mechanism is useful for asymmetric scenarios, such as 3PE scenarios. In dual-homing scenarios, BFD does not need to be deployed to detect single-homing nodes. In this scenario, S-BFD greatly saves resources on the dual-homing side.

To deploy S-BFD, the initiator needs to know the reflector's endpoint and identifier. When a large number of PEs need to be deployed, the deployment is complicated. [RFC7883] and [RFC7884] introduced an IGP-based S-BFD discriminator advertisement mechanism to simplify S-BFD deployment. VPN service may be deployed across inter-area or inter-AS. In this case, the IGP flooding mechanism does not work.

It is recommended to use BGP to distribute BFD discriminator information. BGP can transmit routes across domains, and service routes can drive to generate the end-to-end S-BFD sessions on demand.

2. Terminology

BFD : Bidirectional Forwarding Detection

S-BFD : Seamless Bidirectional Forwarding Detection

APE : Access PE, used to access users

SPE: Service PE, used to support service for users

UCE: User CE

SCE: Service CE

3. Scenarios

In some EVPN deployments, for example, when it spans over multiple domains, only one of a pair of interconnected PEs benefits from monitoring the status of the connection. In such a case, using S-BFD [RFC7880] is advantageous as it reduces the load on one of the PEs while providing the benefit of faster convergence. The following sections provide examples of EVPNs that would benefit from using S-BFD.

For SRv6 services, there are two different service types. One is service over SRv6 BE, the other is service over SRv6 Policy. For the service over SRv6 BE, it will use the VPNSID to resolve the forwarding information. Thus we must generate an S-BFD session to detect the VPNSID's reachability. This is an IP-routed S-BFD. We may use the remote VPNSID's locator as the destination of the S-BFD session. For the service over SRv6 Policy, it will use <nexthop, color> of the service route to resolve an SRv6 Policy. Then we must generate an S-BFD session to detect the reachability of the SR Policy.

3.1. EVPN Layer 3 Service Over SRv6 BE Use Case

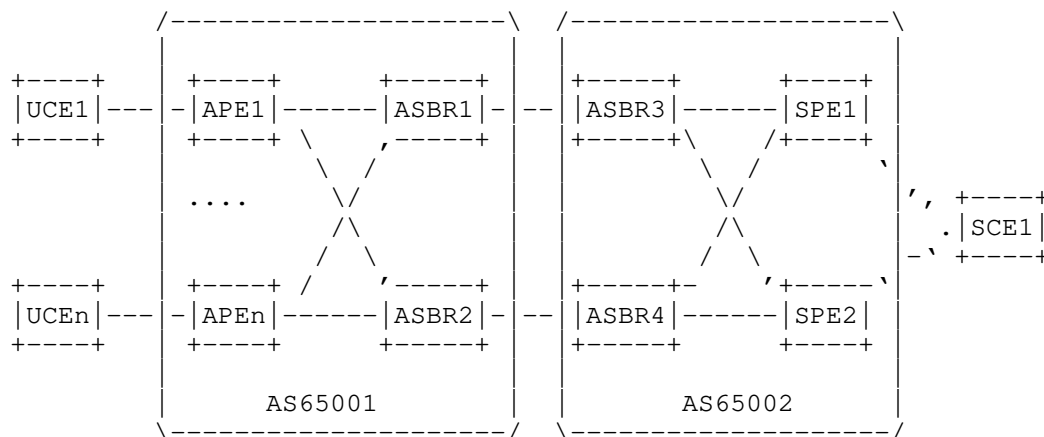


Figure 1: EVPN Layer 3 Service Over SRv6 BE

Figure 1 shows a SRv6 BE based seamless scenario. UCE is single-homed to APE, and SCE is dual-homed to SPE1 and SPE2. The service is across multiple ASes.

SCE1 accesses SPE1 and SE2 through Layer 3 and advertises its private network routes to them. SPE1 and SPE2 encapsulate the routes into Type 5 routes in the EVPN format and sends them to APE1. After receiving Type 5 routes advertised by SPE1 and SPE2, APE1 generates

primary and backup entries for the routes to speed up service switchover. In this scenario, the SRv6 BE service mode is used. APE1 will resolve SPE1's VPN routes reachability through the VPNSID. To ensure that APE1 can properly route to PE1, PE1 needs to advertise its own locator route. The advertisement of the locator route is not in the scope of this document.

To speed up fault detection, we may configure an S-BFD session on APE1 to detect SPE1 or SPE2's reachability. In traditional mode, a discriminator needs to be assigned by SPE1 and SPE2, and two S-BFD sessions need to be configured on APE1 to detect the VPN SID's reachability of SPE1 and SPE2. It needs to generate an S-BFD session with the destination set to the VPN SID. To reduce the number of S-BFD sessions, locator-based S-BFD sessions can be used instead of S-BFD sessions for VPNSIDs.

There are a large number of such APEs that exist on the network. Each APE is configured with several S-BFD sessions to detect PE1 and PE2, which increases the deployment complexity.

3.2. EVPN Layer 3 Service Over SRv6 Policy Use Case

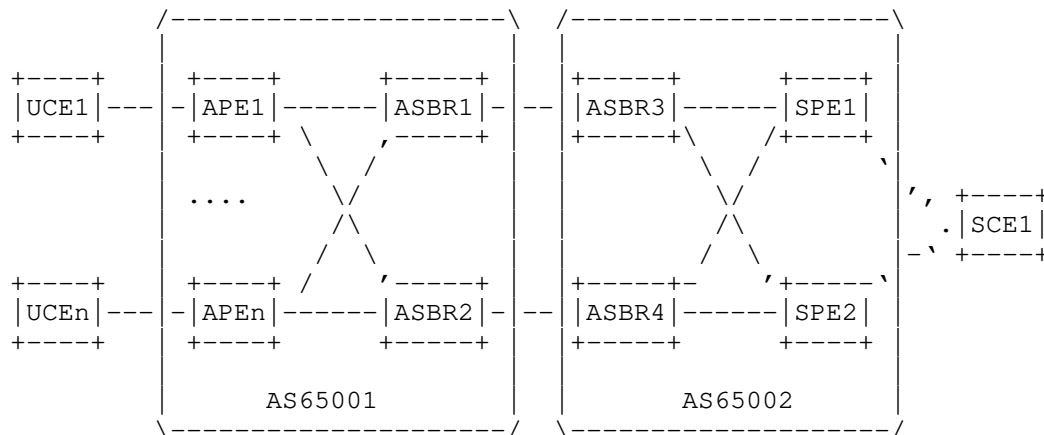


Figure 2: EVPN Layer 3 Service Over SRv6 Policy

Figure 2 shows a SRv6 Policy scenario. SCE1 is dual-homed to SPE1 and SPE2, and UCE1 is accessed to APE1. SPE1, SPE2, and APE1 are cross BGP ASes.

SCE1 accesses SPE1 and SPE2 through Layer 3 and advertises its private network routes to APE1. SPE1 and SPE2 encapsulate the routes into Type 5 routes in the EVPN format and sends them to APE1.

After receiving Type 5 routes advertised by SPE1 and SPE2, APE1 generates primary and backup entries for the routes, speeding up service switchover. APE1 parses the tunnel based on the <nexthop, color> of the service routes advertised by SPE1 and SPE2, and matches an SRv6 Policy. After receiving the traffic from UCE1 to SCE1, APE1 encapsulates and forwards the traffic based on the SRv6 Policy.

An S-BFD session needs to be established for these SRv6 Policy-based forwarding paths to swiftly detect the availability of the paths. When detecting a fault on the SRv6 Policy path of the primary service route, services can be swiftly switched to the backup path, providing more reliable protection for services.

There are a large number of such PEs that exist on the network. Each PE is configured with several S-BFD sessions to detect PE1 and PE2, which increases the deployment complexity.

Certainly, this scenario may also be implemented in other methods. For example, when delivering an SRv6 policy, specify a tunnel to generate an S-BFD session.

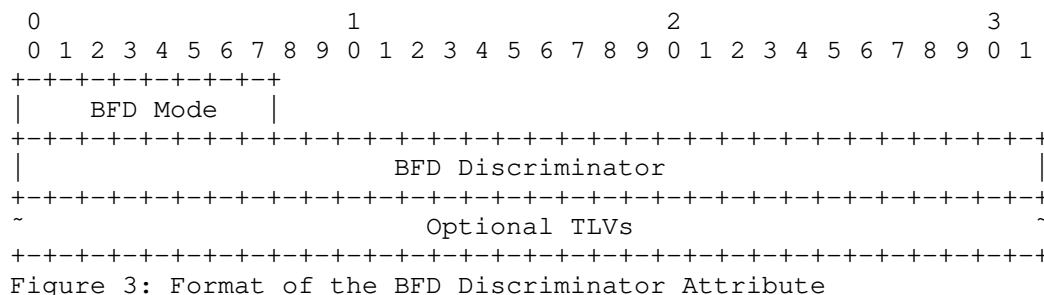
4. Procedure

4.1. BGP Encoding

[RFC9026] specifies the "BFD Discriminators" (38) attribute, which is an optional transitive BGP attribute that conveys the Discriminators and other optional attributes used to establish BFD sessions.

The attribute defined in [RFC9026] is used to transmit P2MP BFD session creation information through the BFD Discriminator attribute in MVPN scenarios. For non-multicast services, such as L3VPN services, L2VPN services, and native IP services, BFD discriminators are also required to create an S-BFD session.

The S-BFD Discriminator attribute introduced in this document is defined as follows:



o BFD Mode:

The BFD Mode field is 1 octet. [RFC9026] defines only the P2MP BFD session for MVPN. This document defines two new types of S-BFD session types based on the preceding scenarios.

As described in the preceding scenario. There are two types of S-BFD sessions for SRv6 services. For service over SRv6 BE, an IP-routed S-BFD session needs to be created to detect the locator route. For service over SRv6 Policy, an S-BFD session for SRv6 Policy path needs to be created to detect the SRv6 Policy path. So two new BFD modes should be introduced here.

S-BFD for SRv6 Locator Session Mode, which is dedicated to detecting the locator. The temporary type is 176, and is to be allocated by IANA.

S-BFD for Common Session Mode, which is for general S-BFD session. The temporary type is 177, and is to be allocated by IANA. This mode is not only for SRv6, but also can be used for other scenarios.

o BFD Discriminators:

The field length is 4 octets. Used to specify the discriminator for S-BFD session.

o Optional TLVs:

Variable-length fields are optional. Indicates the additional information required for creating a S-BFD session. The format is as follows:

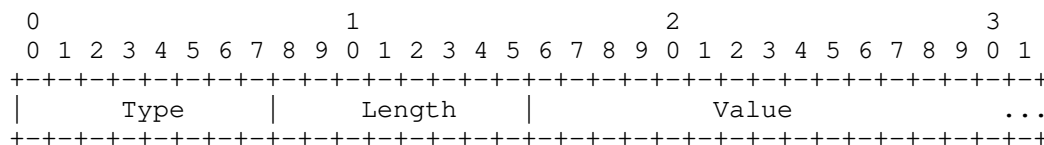


Figure 4: Format of the Optional TLV

If a transit node changes the next hop or reassigns a VPN SID when forwarding a route, the transit node needs to use the locally allocated S-BFD discriminator to advertise the S-BFD discriminator attribute. If the transit node does not recognize the S-BFD Discriminator attribute in the learned route and continues to advertise the route to the remote PE, the receiver may use incorrect information when creating an S-BFD session. Therefore, the advertised S-BFD Discriminator attribute needs to carry the IP address for receiver verification.

In this document, S-BFD for SRv6 Locator Session and S-BFD for Common Session must carry IP addresses except discriminators, which reuse the Source IP Address TLV defined in [RFC9026].

If the mode is set to S-BFD for SRv6 Locator Session, the SRv6 Locator address used for the service is carried.

If the mode is set to S-BFD for Common Session, the next-hop address used for the service is carried.

For details about the error handling, see section "Error Handling".

4.2. Router Procedure

In BGP address families, such as L3VPN or EVPN, routes can carry the S-BFD Discriminator attribute as required so that S-BFD sessions can be established based on the attribute. The following uses S-BFD for SRv6 Locator as an example. If mode is set to S-BFD for Common Session, the processing method is similar.

4.2.1. Egress Node Process

As shown in figure 1, the S-BFD discriminator is configured on PE1. After obtaining the information, BGP encapsulates the attribute into the EVPN route and sets the BFD Mode to S-BFD for Locator Session, when advertising the EVPN route. The Discriminator value is local discriminator value. The optional TLV carries the local PE's locator address used by the VPN.

4.2.2. Transit Node Process

Here is the end-to-end SRv6 BE scenario. The ASBR does not re-allocate the VPN SID. Thus, the ASBR does not require to modify the VPN SID, and not to alter the BFD discriminator attribute.

4.2.3. Ingress Node Process

After receiving the EVPN Type 5 routes from PE1 and PE2, PE3 imports the routes to the VRF of PE3 based on the route targets. Routes triggers establish the S-BFD sessions based on <S-BFD discriminator, locator ip>.

Then, routes with the same prefix from PE1 and PE2 form primary and backup paths. When the primary path or the egress node is in fault, S-BFD detects that fault and forms switch to backup path quickly.

To avoid the waste of redundant resources, assume that the ASBR re-assigns the SID in Option B and the ASBR does not recognize the attribute. In this case, the SID and locator carried in the route received by PE3 do not match the Source IP carried in the Optional TLV in the BFD attribute. Therefore, PE3 does not need to establish an S-BFD session to remote PE, which can avoid resource waste.

5. Error handling

Error handling complies with [RFC7606]. In this document, the BFD discriminator information is used only to establish an S-BFD session. Therefore, if the BFD discriminator information is invalid, the BFD attribute will be discard and not transmit to other devices.

For BFD discriminator attribute, the following case will be processed:

- o The BFD Discriminator value in receiving BFD Discriminator attribute is 0, the attribute is invalid.

For BFD mode type is S-BFD for SRv6 Locator Session, the following case will be processed:

- o The BFD discriminator attribute doesn't contain optional TLV with type set to 1, the attribute is invalid.
- o The optional TLV type is 1 but the length is not 16, the attribute is invalid.
- o The optional TLV type is 1 but the value is all 0, the attribute is invalid.

- o If multiple Source IP Optional TLVs are carried, the first source IP address should be used as the destination to establish an S-BFD session. For EVPN type 2 MAC-IP routes may use the first and the second IP address because it may carry two SRv6 SIDs with different locators. Other source IP addresses should be ignored.

- o If a non-Source IP Optional TLV is carried, the Optional TLV will be ignored.

For BFD mode type is S-BFD for Common Session, the following case will be processed:

- o The BFD discriminator attribute doesn't contain optional TLV with type set to 1, the attribute is invalid.

- o The optional TLV type is 1 but the length is not 4 or 16, the attribute is invalid.

- o The optional TLV type is 1 but the value is all 0, the attribute is invalid.

- o If multiple Source IP Optional TLVs are carried, only the first source IP address should be used as the destination to establish an S-BFD session. Other source IP addresses should be ignored.

- o If a non-Source IP Optional TLV is carried, the Optional TLV will be ignored.

6. IANA Considerations

This document defines two new BFD modes in the BFD Discriminator attribute. The following values are recommended to be assigned by IANA:

Value	Description
-----	-----
176	S-BFD for SRv6 Locator Session
177	S-BFD for Common Session

7. Security Considerations

The new S-BFD Discriminators sub-TLV does not introduce any new security risks for BGP.

When creating an S-BFD session, the initiator verifies the S-BFD session based on routing information. This reduces the number of invalid S-BFD sessions and avoid attribute attack.

8. Acknowledgements

The authors would like to thank Greg Mirsky for their review and comments.

9. References

9.1. Normative References

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

9.2. References

- [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>>.
- [RFC7880] Pignataro, C., Ward, D., Akiya, N., Bhatia, M., and S. Pallagatti, "Seamless Bidirectional Forwarding Detection (S-BFD)", RFC 7880, DOI 10.17487/RFC7880, July 2016, <<https://www.rfc-editor.org/info/rfc7880>>.
- [RFC7883] Ginsberg, L., Akiya, N., and M. Chen, "Advertising Seamless Bidirectional Forwarding Detection (S-BFD) Discriminators in IS-IS", RFC 7883, DOI 10.17487/RFC7883, July 2016, <<https://www.rfc-editor.org/info/rfc7883>>.
- [RFC7884] Pignataro, C., Bhatia, M., Aldrin, S., and T. Ranganath, "OSPF Extensions to Advertise Seamless Bidirectional Forwarding Detection (S-BFD) Target Discriminators", RFC 7884, DOI 10.17487/RFC7884, July 2016, <<https://www.rfc-editor.org/info/rfc7884>>.
- [RFC9026] Morin, T., Ed., Kebler, R., Ed., and G. Mirsky, Ed., "Multicast VPN Fast Upstream Failover", RFC 9026, DOI 10.17487/RFC9026, April 2021, <<https://www.rfc-editor.org/info/rfc9026>>.

Authors' Addresses

Haibo Wang
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: rainsword.wang@huawei.com

Jie Dong
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: jie.dong@huawei.com

Greg Mirsky
Ericsson
Email: gregimirsky@gmail.com

Yang Huang
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: yang.huang@huawei.com

routing
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2022

Z. Zhang
Juniper Networks
October 25, 2021

MVPN Inter/Intra-region Tunnel Segmentation
draft-zzhang-bess-mvpn-regional-segmentation-01

Abstract

RFC7524 specifies procedures for Inter-Area Point-to-Multipoint Segmented Label Switched Paths (aka MVPN tunnel segmentation). This document updates RFC7524 by extending the inter-area segmentation concept to inter-region and intra-region segmentation where a region is no longer tied to an IGP area.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 28, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
1.1. Tunnel Segmentation	2
1.2. Intra-region Segmentation	3
1.3. Bud Node Support	3
2. Specifications	3
2.1. Inter-region Segmentation	3
2.2. Intra-region Segmentation	4
2.3. Bud Node Support	5
3. Security Considerations	6
4. IANA Considerations	6
5. Acknowledgements	6
6. References	6
6.1. Normative References	6
6.2. Informative References	7
Author's Address	7

1. Introduction

1.1. Tunnel Segmentation

[RFC6514] specifies (among other things) inter-AS MVPN tunnel segmentation procedures and [RFC7524] specifies inter-area MVPN tunnel segmentation procedures. The procedures for inter-AS and inter-area are similar in that the segmentation points - ASBRs or ABRs - change the PMSI Tunnel Attribute (PTA) attached to I/S-PMSI routes to specify the type and identification of tunnel to be used in the next AS/area, when they re-advertise the PMSI routes into the next AS/area.

This change of tunnel at the segmentation points and stitching of upstream and downstream tunnel segments not only allows different tunnel technology/instance to be used in different AS/area, but also limits the replication of traffic to only PEs and segmentation points in the local AS/area, instead of to all PEs.

The inter-area segmentation points are route reflectors and when they re-advertise the x-PMSI routes to different downstream areas they may use different BGP neighbor groups so that different tunnel type/identification can be encoded in PTA for different downstream areas. If the ABR is also responsible for reflecting the routes to PEs in the same area, the ABR does not modify the PTA (because of that those local PEs are also put into a different neighbor group).

As a result, a segmentation point will likely have different neighbor groups (one group for each area) so that the PTA and Inter-Area P2MP

Segmented Next-Hop Extended Community (referred to as Segmentation EC) can be set accordingly when it re-advertise the x-PMSI routes.

The provisioning of a RR with these different neighbor groups for segmentation purpose can actually be done on any router (as a segmentation point) - not necessarily on an ABR. As a result, the procedures in RFC7524, while specified for inter-area, can be extended inter-regiona as well - the segmentation points can be any border routers between arbitrarily defined "regions".

This concept is already described in Section 6 of [I-D.ietf-bess-evpn-bum-procedure-updates], but specified formally in this document for MVPN.

1.2. Intra-region Segmentation

Even with the inter-area segmentation extended to inter-region, when a regional border router (RBR) reflects routes to PEs in the same region, it does not modify the PTA or Segmentation EC. But if the RBR also modifies the two attributes when reflecting routes to the local PEs, tunnel segmentation is achieved even intra-region - both the upstream and downstream tunnel segments are in the same region.

This Intra-region Segmentation is one way to achieve Assisted Replication in MVPN: a PE sends traffic to assisting replicators who will then relay traffic to other PEs (even in the same region).

1.3. Bud Node Support

A segmentation point may have both local receivers off a VRF and downstream receivers off a remote PE for traffic arriving on an upstream segment. This segmentation point is referred to a bud node, just like that a node can be both a transit and leaf node for a P2MP tree.

Depending on implementation, a bud node may need to receive two copies of a packet, one for local delivery and one for remote delivery. If so, the bud node may request the upstream PE or segmentation point to send two copies.

2. Specifications

2.1. Inter-region Segmentation

The procedures in RFC7524 are extended to beyond IGP area-based. A provider network can be arranged into "regions" connected by "Regional Border Routers" (RBRs). On a segmentation point a region MAY be defined as a BGP neighbor group - all peers in the group are

subject to the same export policy, which can be used to control the modification of attributes for the purpose of segmentation.

RFC7524 procedures apply as is, though "area" is replaced with "region" and "Area Border Router" (ABR) is replaced with "Regional Border Router" (RBR).

The concept of Per-region Aggregation, as explained in Section 6.1 of [I-D.ietf-bess-evpn-bum-procedure-updates], is also applicable to MVPN. A future revision of this document will specify details of Per-region Aggregation for MVPN.

2.2. Intra-region Segmentation

The following procedures are applicable for intra-region segmentation. One use of intra-region segmentation is for Assisted Replication where PE-PE traffic goes through a relay point (assisting replicator).

If it is known that the local PEs are only peered with the RBRs (as RRs and segmentation points), the PEs and RBRs follow the procedures in RFC7524. In addition, the local RBRs modify the PTA and Segmentation EC even when they re-advertise x-PMSI routes to PEs in the ingress region, thus achieving Intra-region Segmentation.

Otherwise (i.e., if a local PE may import BGP-MVPN routes directly unless with the modified procedures specified below), the following modified procedures apply:

- o When an ingress PE advertises an x-PMSI route, it attaches an Extended Community (EC) derived from the Route Target for the VPN (RT-VPN) [I-D.zzhang-idr-rt-derived-community] but not the RT-VPN itself. Call this EC as EC-VPN. The route also carry a Segmentation EC as specified in RFC7524.
- o When the local RBRs (as RRs and segmentation points) receive this route, it replaces the EC-VPN with the corresponding RT-VPN (the EC-VPN and RT-VPN can be derived from each other), and then re-advertise the route to its peers, with the Segmentation EC and PTA modified as specified in RFC7524. The modification applies even when re-advertising to peers in the same ingress region.

This is to ensure that local egress PEs will only import the routes re-advertised by the RBRs after the modification of PTA and Segmentation EC.

Additionally, if there are intermediate RRs between the ingress PE and local RBRs, and Route Target Constrain [RFC4684] is in use, the

ingress PE MUST also attach a Route Target (referred to as RT-RBR) and the local RBRs MUST be provisioned to import routes with RT-RBR (otherwise the intermediate RRs will not re-advertise the routes towards the RBRs because the routes carry only EC-VPN but not RT-VPN). The local RBRs MUST remove the RT-RBR when they re-advertise the routes.

2.3. Bud Node Support

This section applies only if the segmentation point can not both route traffic arriving on the upstream segment to local receivers and label switch the traffic to downstream segments due to implementation limitation.

If a segmentation point is a bud node for a segmented x-PMSI tunnel with the above mentioned limitation, it SHOULD request an additional copy to be sent by the upstream RSVP neighbor if the upstream segment is a RSVP-TE P2MP tunnel, or by the upstream PE/RBR when the upstream segment is an IR or mLDP tunnel.

The RSVP-TE P2MP case is outside the scope of this document (though there are known implementations). For the IR/mLDP case, it is done by including a Tunnel Encapsulation Attribute (TEA) [RFC9012] in the Leaf A-D route in response to the x-PMSI route for the upstream segment. Note that the leaf A-D route is sent for this purpose even if the Leaf Information Required (LIR) flag is not set in the x-PMSI route (e.g. for mLDP tunnel).

The TEA includes one tunnel of a desired type (e.g. MPLS or Any Encapsulation [I-D.ietf-bess-bgp-multicast-controller]) that is used for the upstream PE/RBR to send the additional copy to this bud node. The tunnel MUST include a Tunnel Egress Endpoint sub-TLV set to a local address on the bud node, and MUST include a Tree Label Stack sub-TLV that includes a single label. The node MUST program a label forwarding entry to pop the label and forward packet based on IP lookup in a VRF identified by the label (while the tunnel label for the upstream segment or the label in the PTA of the x-PMSI/Leaf route for the upstream segment is used to stitch the upstream and downstream segments together).

When the upstream PE/RBR decodes the TEA in the Leaf A-D route in response to an x-PMSI A-D route that it (re-)advertises (even if it did set the LIR flag in the x-PMSI A-D route), it SHOULD send an extra copy via unicast tunneling with the label encoded in the Tree Label Stack sub-TLV. If the extra copy is not sent the downstream bud node segmentation point will not be able to send traffic to its local receivers.

3. Security Considerations

No additional security considerations are needed beyond what are discussed in RFC7524.

4. IANA Considerations

This document requests the IANA to create a "PMSI Tunnel Attribute Extension sub-TLV Type Registry". Allocation from the registry is First Come First Serve, with an initial allocation for "Additional Label".

5. Acknowledgements

The author thanks Sanoj Vivekanandan for his review and comments.

6. References

6.1. Normative References

- [I-D.ietf-bess-bgp-multicast-controller]
Zhang, Z., Raszuk, R., Pacella, D., and A. Gulko,
"Controller Based BGP Multicast Signaling", draft-ietf-bess-bgp-multicast-controller-07 (work in progress), July 2021.
- [I-D.zzhang-idr-rt-derived-community]
Zhang, Z., "Extended Communities Derived from Route Targets", draft-zzhang-idr-rt-derived-community-01 (work in progress), March 2021.
- [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>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.

6.2. Informative 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-11 (work in progress), October 2021.

[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>>.

Author's Address

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

bess
Internet-Draft
Intended status: Standards Track
Expires: 28 April 2022

Z. Zhang
Juniper Networks
R. Parekh
Cisco Systems
Z. Zhang
ZTE
H. Bidgoli
Nokia
25 October 2021

MVPN and EVPN BUM Signaling with Controllers
draft-zzhang-mvpn-evpn-controller-01

Abstract

This document specifies optional procedures for BGP-MVPN and EVPN BUM signaling with controllers. When P2MP tunnels used for BGP-MVPN and EVPN BUM are to be signaled from controllers, the controllers can learn tunnel information (identifier, root, leaf) by participating BGP-MVPN and EVPN BUM signaling, instead of relying on ingress PEs to collect the information and then pass to the controllers. Additionally, Inclusive/Selective PMSI Auto Discovery Routes can be originated from controllers based on central provisioning, instead of from PEs based on local provisioning.

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 28 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Terminologies	2
2. Introduction	3
3. Specification	4
3.1. Controller Address Extended Community	4
3.2. Targeting Leaf A-D Routes to Controllers	4
3.3. Controller Originated I/S-PMSI Routes	5
3.3.1. Inter-AS/Region Segmentation	5
3.4. Automatic DCB Label Allocation by Controllers	6
4. Security Considerations	6
5. IANA Considerations	6
6. Acknowledgements	7
7. References	7
7.1. Normative References	7
7.2. Informative References	7
Authors' Addresses	8

1. Terminologies

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

- * 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/BD.
- * I-PMSI: Inclusive PMSI - to all PEs in the same VPN/BD.
- * S-PMSI: Selective PMSI - to some of the PEs in the same VPN/BD.

- * Leaf A-D routes: For explicit leaf tracking purpose. Triggered by S-PMSI A-D routes and targeted at triggering route's originator.
- * IMET A-D route: Inclusive Multicast Ethernet Tag A-D route. The EVPN equivalent of MVPN Intra-AS I-PMSI A-D route.

As pointed out above, 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. Introduction

Consider a provider network with BGP-MVPN/EVPN where controllers are used to set up P2MP tunnels per [I-D.ietf-bess-bgp-multicast-controller] or [I-D.ietf-pim-sr-p2mp-policy]. For a controller to calculate the corresponding trees and set up the tunnels, it needs to learn the (ID, root, leaf) information for those trees. Currently, [I-D.ietf-bess-mvpn-evpn-sr-p2mp] specifies that an ingress PE assigns the SR P2MP ID and collects leaf information via Leaf A-D routes, and then pass onto the controller. Observing that BGP-MVPN/EVPN signaling typically involves Router Reflectors, which may typically be hosted on or co-located with controllers, it makes sense to have the controllers participating BGP-MVPN/EVPN signaling to learn (ID, root, leaf) information. This will relieve the PEs from maintaining Leaf A-D routes, and remove the extra hop of leaf information propagation.

Also Consider that in the same network many selective tunnels are used, and their usages are dynamically provisioned based on specific needs at different time. For example, the provider provides video transmission services for events at various time, location and to various receivers. With traditional methods the provider would provision the PEs at the transmission sources with various selective tunnels, which triggers corresponding S-PMSI A-D routes. The provisioning is put in place shortly before an event takes place and removed shortly after the event ends. Alternatively and preferably, a controller can originate S-PMSI A-D routes based on centralized provisioning on behalf of the source PEs. The controller also collects the leaf information (either based on centralized provisioning or based on Leaf A-D routes), calculates the tree and signal tree nodes. Additionally, when tunnel aggregation labels are allocated from Domain-wide Common Block (DCB), originating I/S-PMSI A-D routes from controllers makes the DCB label allocation a lot easier.

It is possible that an operator prefers automatic DCB aggregation label allocation by the controller but prefers I/S-PMSI A-D routes origination from individual PEs. In that case, a PE can target an I/S-PMSI A-D route at the controller and the controller will allocate a DCB label and return it in a corresponding Leaf A-D route.

3. Specification

The procedures specified in this section applies if one or more controllers participate MVPN/EVPN signaling for the purpose of leaf discovery for P2MP tree calculation, and/or if controllers are to originate I/S-PMSI A-D routes or BGP-MVPN and/or BGP-EVPN BUM.

3.1. Controller Address Extended Community

This document defines a new Transitive IPv4-Address-Specific Extended Community Sub-Type: "Controller Address". This document also defines a new BGP Transitive IPv6-Address-Specific Extended Community Sub-Type: "Controller Address".

A Controller Address Extended Community (referred to as Controller EC) is constructed by setting the Global Administrator field to the IP address of the controller and the Local Administrator field to 0.

3.2. Targeting Leaf A-D Routes to Controllers

When a PE originates an I/S-PMSI A-D route with PTA's tunnel type set to PIM-SSM/ASM, mLDP or SR P2MP that are to be set up by controllers, the PE MUST attach a Controller EC constructed as above. If there are multiple controllers, then one Controller EC is attached for each of the controllers.

In case of tunnel segmentation and a new controller is used for the next segmentation region, when an ABR/ASBR/RBR re-advertises the I/S-PMSI A-D route to the next segmentation region it MUST modify the Controller EC to specify the new controller address.

When a PE/ABR/ASBR/RBR receives an I/S-PMSI A-D route with the Controller EC, it MUST originate a corresponding Leaf A-D route. The PTA from the I/S-PMSI A-D route is copied to the Leaf A-D route, and an IP Address Specific Route Target is attached to the Leaf A-D route. The Global Administrator field of the RT is set to the address of the controller (as encoded in the received Controller EC), and the Local Administrator field is set to 0.

Note that, the above is done even if the Leaf Information Required (LIR) bit in the Flags field of the I/S-PMSI A-D route's PMSI Tunnel Attribute (PTA) is not set. If the LIR bit in the Flags field of the

I/S-PMSI A-D route's PTA is set, then the above mentioned RTs are in addition to the RT that the PE attaches according to the procedures in [RFC6514], [RFC7524], or [I-D.ietf-bess-evpn-bum-procedure-updates]. In other words, the Leaf A-D route will have RTs for both the controllers and the upstream PE or segmentation points in this case.

When a controller receives the advertisement and/or withdrawl of Leaf A-D routes, it derives the set of leaves for the tunnel identified in the PTA, calculate and set up the tree according to procedurs in [I-D.ietf-bess-bgp-multicast-controller] or [I-D.ietf-pim-sr-p2mp-policy]. The controller does not further propagate the received advertisement and/or withdrawl, unless there are other RTs attached.

3.3. Controller Originated I/S-PMSI Routes

When I/S-PMSI A-D routes are to be originated from the controllers, it is expected that the controller, based on central planning, has the knowledge of each VPN/BD's Route Target, each PE's RD for the VPN/BD, and the Tunnel Type and Identifier for each I/S-PMSI. If the tunnel aggregation is used, the controllers also allocate labels from the DCB for the I/S-PMSIs.

The controller constructs the I/S-PMSI A-D route the same way as if an ingress PE would be originating the routes. There are some exceptions in case inter-AS/region segmentation is used, as specified in Section 3.3.1.

Specifically, the controller uses the ingress PE's RD and RTs for the VPN/BD, and use the ingress PE's address as "Originating Router's IP Address" when constructing the I/S-PMSI A-D routes. The routes are sent with the controller's address as next-hop initially, though the next-hop may change as the routes propagates.

When the Ingress PE router receives the I/S-PMSI A-D routes, it sets up corresponding forwarding state as if it originated the routes per its local provisioning. Note that the next-hop address of the routes will be different from the case where the ingress PE originates the routes, but that does not matter.

3.3.1. Inter-AS/Region Segmentation

In case of segmentation, instead of using the Route Target for the VPN/BD, the controller constructs an IP Address specific Route Target with the Global Administrator Field set to the corresponding ingress PE's address and the Local Administrator Field set to 0. This targets the I/S-PMSI A-D routes to the Ingress PEs only.

The controller also sets the Originating Router's IP Address field of the I/S-PMSI A-D route to its own address.

The receiving Ingress PE associate the I/S-PMSI A-D route to the corresponding VRF/BD based on the RD of the received route. It then re-originate a corresponding I/S-PMSI A-D route based on the received I/S-PMSI A-D route from the controller by doing the following:

- * Changing the Originating Router's IP address to its own
- * Replacing the Route Target with the Route Target for the VPN/BD

3.4. Automatic DCB Label Allocation by Controllers

If it is desired for a PE to originate I/S-PMSI A-D routes on its own but with DCB labels dynamically allocated by a controller, the PE originates the I/S-PMSI A-D route with the Tunnel Type in the PTA set to "no tunnel information present", the LIR bit in the PTA'S Flags field set to 1, and attaches an IP Address Specific RT. The RT's Global Administrator Field is set to the Controller's address and Local Administrator field is set to 0.

When the controller receives the I/S-PMSI A-D route, it allocates a DCB label and responds with a Leaf A-D route. The Label field of the Leaf A-D route's PTA is set to the allocated DCB label.

When the PE receives the Leaf A-D route, it re-advertises the I/S-PMSI A-D route, with an additional RT for the corresponding VPN/BD. The PTA's tunnel information is set as needed and the Label field is set to the DCB label received in the Leaf A-D route. The LIR bit in the Flags field of the PTA is set to 1 or 0 as needed. If it is set to 0, the controller withdraws the Leaf A-D route but does not release the allocated label.

When the PE withdraws the I/S-PMSI A-D route, the controller release the DCB label and withdraws the corresponding Leaf A-D route if it had not been withdrawn before.

4. Security Considerations

This document does not change security aspects as discussed in [RFC4360], [6514], [7432], and [I-D.ietf-bess-evpn-bum-procedure-updates].

5. IANA Considerations

To be added.

6. Acknowledgements

7. References

7.1. 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", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-bum-procedure-updates-11, 7 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-evpn-bum-procedure-updates-11.txt>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

7.2. Informative References

- [I-D.ietf-bess-bgp-multicast-controller]
Zhang, Z., Raszuk, R., Pacella, D., and A. Gulko, "Controller Based BGP Multicast Signaling", Work in Progress, Internet-Draft, draft-ietf-bess-bgp-multicast-controller-07, 12 July 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-bgp-multicast-controller-07.txt>>.
- [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", Work in Progress, Internet-Draft, draft-ietf-bess-mvpn-evpn-aggregation-label-06, 19 April 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-mvpn-evpn-aggregation-label-06.txt>>.

[I-D.ietf-bess-mvpn-evpn-sr-p2mp]

Parekh, R., Filsfils, C., Venkateswaran, A., Bidgoli, H., Voyer, D., and Z. Zhang, "Multicast and Ethernet VPN with Segment Routing P2MP", Work in Progress, Internet-Draft, draft-ietf-bess-mvpn-evpn-sr-p2mp-04, 19 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-mvpn-evpn-sr-p2mp-04.txt>>.

[I-D.ietf-pim-sr-p2mp-policy]

(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", Work in Progress, Internet-Draft, draft-ietf-pim-sr-p2mp-policy-03, 23 August 2021, <<https://www.ietf.org/archive/id/draft-ietf-pim-sr-p2mp-policy-03.txt>>.

[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>>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

Rishabh Parekh
Cisco Systems

Email: riparekh@cisco.com

Zheng Zhang
ZTE

Email: zhang.zheng@zte.com.cn

Hooman Bidgoli
Nokia

Email: hooman.bidgoli@nokia.com