BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: 3 September 2025

A. Sajassi
K. Thiruvenkatasamy
S. Thoria
Cisco
A. Gupta
VMware
L. Jalil
Verizon

2 March 2025

Seamless Multicast Interoperability between EVPN and MVPN PEs
draft-ietf-bess-evpn-mvpn-seamless-interop-08

Abstract

Ethernet Virtual Private Network (EVPN) solution is becoming pervasive for Network Virtualization Overlay (NVO) services in data center (DC), Enterprise networks as well as 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/.

Copyright Notice

Table of Contents

1. Introduction

Ethernet Virtual Private Network (EVPN) solution is becoming pervasive for Network Virtualization Overlay (NVO) services in data center (DC) and Enterprise networks as well 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 specified in [RFC6513] & [RFC6514] 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 a 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 Interior Gateway Protocol (IGP) domain). In such scenarios, the service providers want to have optimum forwarding among these PE devices without the use of gateway devices. If gateway devices are used, then the IP multicast traffic between an EVPN and MVPN PEs can no longer be optimum and in some cases, 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 the 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 needs to configure per-tenant info on the gateways. In other words, for each tenant that is configured, one (or maybe two) additional touchpoints 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 broadcast domains of the same tenant system, a router attached to an EVPN PE would send multiple copies into the EVPN fabric resulting in bandwidth wastage. [RFC9135] only covers procedures for efficient inter-subnet connectivity among these Tenant Systems and End Devices while maintaining the multihoming 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. 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 Data Center Interconnect (DCI) use cases.

An EVPN network can consist of a mix of L2 and L3 PEs. The multicast operation of such a 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", "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

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

All-Active Redundancy Mode: When all PEs are 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

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

Border Leafs: A set of EVPN PEs acting as an exit point for EVPN fabric

CO: Central Office of a service provider

C-Prefix: Refers Prefix in the tenant context. C-S refers tenant/customer source; C-G refers tenent/customer group. Wildcard is used in place of S/G to indicate all sources/all groups.

DCI: Data Center Interconnect

EC: BGP Extended Community

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.

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

EVPN: Ethernet VPN

FHR: First Hop Router

GENEVE: Generic Network Virtualization Encapsulation

IMET: Inclusive Multicast Ethernet Tag

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

LATA: Local Access and Transport Area

LHR: Last Hop Router

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

NV: Network Virtualization

NVE: Network Virtualization Endpoint

NVGRE: Network Virtualization using Generic Routing Encapsulation

NVO: Network Virtualization Overlay

PE: Provider Edge device

PIM-SM: Protocol Independent Multicast - Sparse-Mode

PIM-SSM: Protocol Independent Multicast - Source Specific Multicast
Bidir PIM: Bidirectional PIM

PoD: Point of Delivery

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

SPDC: Service Provider Data Center

UMH: Upstream Multicast Hop

VNI: Virtual Network Identifier (for VXLAN)

VXLAN: Virtual Extensible LAN

TS: Tenant Systems

4. Requirements

This section describes the requirements specific to 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 (I.e, only one copy will be used to service both EVPN and MVPN PE when same tunnel encapsulation is used). 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 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 Multihoming

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

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]:

* VLAN-based service interface

* VLAN-bundle service interface

* VLAN-aware bundle service interface.

4.6. Distributed Anycast Gateway

The solution SHALL support distributed anycast gateways as specified in [RFC9135] 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 an 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 broadcast domains (BDs) are provisioned and removed on a 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 Section 6.3 of [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 an EVPN PE needs to be modeled very similar to a 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 is 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 Destination Address (MAC-DA) is performed); whereas, the inter-subnet traffic is routed in the corresponding IP-VRF (i.e. a lookup based on IP Destination Address (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 extended for intra- subnet multicast traffic described in [RFC9251]. 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 the 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 the 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. In other words, the TTL value of IP packet is 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, non-IP multicast and 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 to 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 the adjacency of that PE. Furthermore, the PE uses MVPN BGP protocol and procedures per [RFC6513] and [RFC6514]. With respect to multicast routing protocol between the 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 multicast VPN 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 an 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.

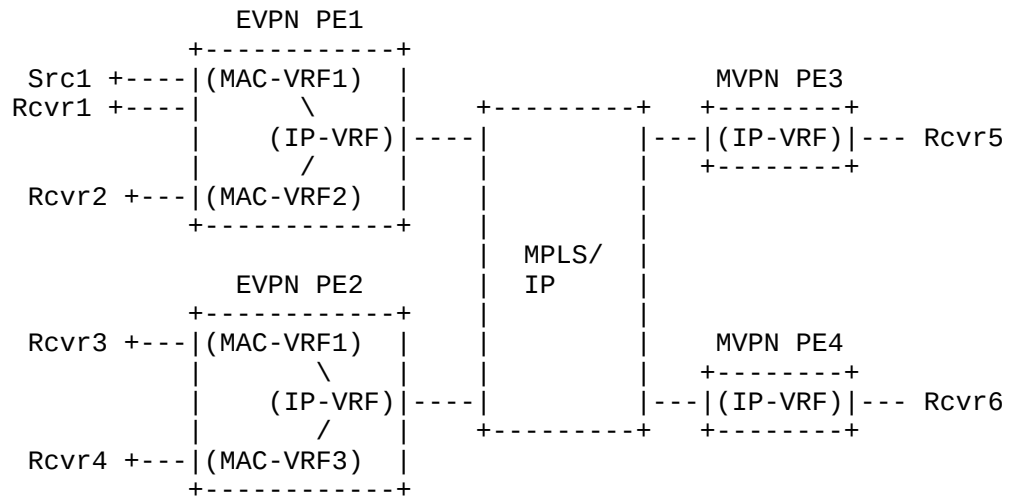

Figure 1: EVPN & MVPN PEs Seamless Interop

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 an 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 an MVPN PE where an attachment circuit serves as a routed multicast interface for an IP-VRF associated with an 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.

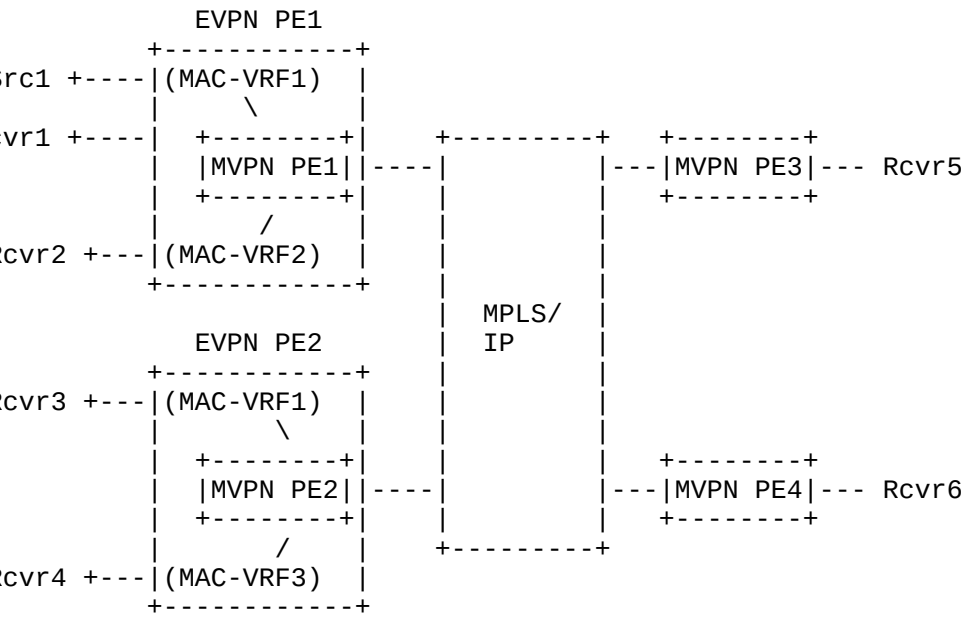

Figure 2: EVPN PEs as MVPN PEs

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.

* Unicast route advertisements for IP multicast source

* multihoming of IP multicast sources and receivers

* 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 the source is attached to a Single-Active multihomed Ethernet Segment (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 VPN-IP 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, the VPN-IP unicast route is

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

When the multicast source is attached to an All-Active multihomed ES, then the PE that learns the source advertises the unicast route for that source using EVPN route type 2 and VPN-IP 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, the VPN-IP unicast route is advertised with RT corresponding to the IP-VRF. When the other multihoming 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 VPN-IP 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 multihoming PEs associated with that All-Active ES even though one of the multihoming 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 is attached to itself or sources that are connected to a BD, which is not extended across EVPN fabric and advertises those routes with EVPN route type 5(i.e., IP Prefix route as per Section 3 of [RFC9136]). EVPN host routes are also advertised as VPN-IP 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 among 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 VPN-IP routes within EVPN only fabric.

Section 9 discusses Data Center Interconnect (DCI) use cases, where EVPN and MVPN networks are connected using a gateway model. In the gateway model, EVPN PE advertises unicast routes as VPN-IP routes along with VRI extended community for all multicast sources are attached behind EVPN PEs. All VPN-IP routes SHOULD be summarized while adverting to MVPN PEs.

5.4. Multihoming of IP Multicast Source and Receivers

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

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

5.4.1. Single-Active Multihoming

When a TS source resides on an ES that is multihomed 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 multihoming PE learns the unicast route of the TS source and advertises that using EVPN and VPN-IP to other PEs as described in Section 5.3.

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 the case of Singe-Active multihoming. 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 the case of (*,G). Since the active PE (which is also the UMH PE) has been advertised unicast route for X along with the VRF Route Import EC, the downstream PEs select the UMH without any ambiguity based on MVPN procedures described in Section 5.1 of [RFC6513].

The multihoming 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 based on incoming IGMP/MLD join for that subnet.

- L3 routes the multicast traffic to other BTs for other subnets if there are any interested receivers based on incoming IGMP/MLD or PIM join for those subnets.

* L3 routes the multicast traffic to other PEs per [RFC6513] & [RFC6513] MVPN procedures.

The multicast traffic can be sent on Inclusive, Selective, or Aggregate Selective Tree as specified in Section 2.1.1 of [RFC6513]. 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 Multihoming

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

The multihoming EVPN PE that receives the IP multicast flow on its local AC needs to do the following tasks in addition to the ones listed in the previous section for Single-Active multihoming: L2 switch the multicast traffic to other multihoming EVPN PEs for that ES via a multicast tunnel which 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 multihoming 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 multihoming. 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 the vast majority of multihoming 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 multihoming PE is sent to other multihoming PEs via this intra-ES subnet tunnel, there is no need for sending the multicast traffic via MVPN tunnel to these multihoming 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 multihoming PEs for a given ES can be either fixed or on a demand basis.

By feeding IP multicast flow received on one of the EVPN multihoming PEs to the interested EVPN PEs in the same multihoming group, we have essentially enabled all the EVPN PEs in the multihoming 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 the 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 the highest UMH PE IP address and the UMH selection algorithm based on a 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 multihoming 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 multihoming 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 multihoming 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 the 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 multihoming 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] Section 8.3 uses ESI MPLS label for split-horizon filtering of Broadcast/Unknown unicast/ multicast (BUM) traffic from an All-Active multihoming Ethernet Segment to ensure that BUM traffic doesn’t get looped back to the same Ethernet Segment that it came from. This split-horizon filtering mechanism applies as-is for multicast IRB scenarios because of using the intra- ES tunnel among multihoming PEs. Since the multicast traffic received from a TS source on an All-Active ES by a multihoming PE is bridged to all other multihoming 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 multihomed behind a group of multihoming EVPN PEs, then TS mobility is supported among EVPN PEs. Furthermore, such TS mobility will 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 PE, then the EVPN mobility procedure discovers this move and a new unicast route advertisement (using both EVPN and VPN-IP routes) is made by the EVPN PE where the source has moved to per Section 5.3 above and unicast route withdrawal (for both EVPN and VPN-IP 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 needs to setup the following three types of multicast tunnels. The first two are among EVPN PEs and are associated with the attached BD, but the third one is among EVPN and MVPN PEs and is associated with tenant’s IP-VRF

1) Intra-ES subnet tunnel

2) Intra-subnet BUM tunnel

3) Inter-subnet IP multicast tunnel

While advertising IMET routes, all seamless interop capable PEs SHOULD attach EVPN Multicast Flags Extended Community with "EVPN/MVPN Seamless Interop Supported" flag set. This is a new flag defined in this document (Refer Section 13)

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 multihoming EVPN PEs for that ES to carry multicast flow received by one of the multihoming PEs to the other PEs in that ES. We refer to this multicast tunnel as an Intra-ES subnet tunnel. It is recommended to use Ingress replication for Intra-ES subnet tunnel(especially for dual-homing usecases), In case of multihoming 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 multihoming and it is not required for Single-Active multihoming.

The EVPN PEs belonging to a given All-Active ES discover each other using EVPN Ethernet Segment route per procedures described in [RFC7432] Section 8. 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 multihoming 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 multihoming 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 multihoming 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 [RFC9251], traffic may be forwarded on demand basis. Based on IGMP/MLD synchronization procedure specified in [I-D.ietf-bess-evpn-igmp-mld-proxy], the 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 an IGMP/MLD 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.

In case of a TS, receiver sits behind an All-Active Multihoming ES and a TS source sits behind an inter-subnet tunnel (with respect to the multihomed PE), it is possible that more than one multihomed PEs

sends MVPN join toward 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 multihoming 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 for tenant IP multicast traffic. This tunnel is setup using the 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 multihoming 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 multihoming 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 the 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 optimizing 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 originated/terminated among MAC-VRFs of EVPN PEs.

6.4. IGMP/MLD Hosts as TSes

IGMP/MLD 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/MLD states are terminated at IRB interfaces, and local interest are expressed in the context of IP-VRF to remote PEs.

In the case of a network with only IGMP/MLD 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 an 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 the 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, when IGMP/MLD snooping is enabled. 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. In snooping disabled case, no L2 state will be created. Irrespective of snooping state, the IGMP/MLD join message is then 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 multihoming 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/MLD join for (C-S, C-G), IGMP/MLD snooping will record this state in the 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/MLD 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 the 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, gets added as another OIF to the 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 by 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 the 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 a 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 IP-VRF (Since EVPN PE is modeled as MVPN PE, source and receivers are announced to remote PE in the context of tenant IP-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 VPN-IP 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 may be constructed as following:

* The 4-octet Global Administrator field can be set to an IP address of the PE (PE’s loopback address or VTEP address) or BGP router-id.

* The 2-octet Local Administrator can be set to a number that uniquely identifies the IP-VRF within the PE.

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

* The Local Administrator field of the MVPN RT MAY be set to the VNI associated with the tenant IP-VRF.

Every PE that detects a local receiver via a local IGMP/MLD 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 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/MLD/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 messages 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 PEs supporting [RFC6514] based MVPN over IP/MPLS.

8.2.2. VxLAN Encapsulation

When the encapsulation mode is configured as the 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].

A gateway is needed for inter-operation between the EVPN MVPN-capable 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 that 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 the VXLAN case, a gateway is needed for inter-operation between the EVPN MVPN-capable PEs and non-EVPN MVPN PEs. Any other encapsulation models other than specificied in sections 8.2.2 & 8.2.3 are outside the scope of this document and may be covered in a separate document if required.

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 purposes. In such scenarios, a DC network can be represented as a customer network that is multihomed to two or more MVPN PEs via L3 interfaces and thus standard MVPN multihoming procedures are applicable here. It should be noted that an 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 purposes, 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:

* 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).

* 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

* Source Active A-D Route: same as joins

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

There are multiple possibilities that need to be considered here.

* L2EVPN PE may or may not have support for [RFC9251]

* Seamless interop PE may or may not support [RFC9251]

* Network may only have L2EVPN PE and Seamless interop capable PE

* 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.

* Case1: [RFC9251] is supported both at seamless interop capable PE and L2EVPN PE.

* Case2: [RFC9251] is supported only at seamless interop capable PE.

* Case3: [RFC9251] is not supported at interop capable PE.

[RFC9251] support is recommended for seamless interop capable PE. SEMG can group L2 EVPN PEs into two separate groups (one that supports the [RFC9251] 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: [RFC9251] 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.

* Process EVPN SMET routes from the remote L2 EVPN PEs that support [RFC9251] and creates L2 multicast state. SMET route in-turn triggers the creation of L3 multicast state similar to the IGMP/ MLD join received on the local AC. SEMG-DF exercises the MVPN procedures for the join.

* It should not process IGMP/MLD control packets from L2EVPN PE that supports [RFC9251].

* Originate SMET(*,*) route towards L2 EVPN PEs. This is to receive traffic from multicast sources that are connected behind L2 EVPN PEs.

* 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.

* When SEMG-DF receives traffic from the MVPN tunnel, it does the following

- Sends the traffic toward the IRB interfaces, where the 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: [RFC9251] 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. The rest of the procedures specified above is applicable for this case.

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

* Process IGMP/MLD control packets from remote L2 EVPN PEs that doesn’t support [RFC9251] and create L2 and L3 state.

* When an IGMP/MLD query is received on the intra-subnet tunnel on BD-X, SEMG-DF needs to send proxy IGMP/MLD reports for all groups that it has learned from remote L2-EVPN PEs on that BD.

* Connecting multicast router behind L2 EVPN PE is not recommended. If a multicast router is connected behind L2 EVPN PE, the BD corresponds to the VRF tunnel needs to be configured in the L2 EVPN PE so that the PIM router may get all joins that are received in the BD corresponds to the MVPN tunnel interface at SEMG-DF.

* SEMG-DF should get all multicast traffic from L2EVPN PEs. This may be achieved by sending an IGMP/MLD query or PIM hello on the intra- subnet tunnel

Case 3: [RFC9251] is not supported at seamless interop capable PE

The procedure for handling this use case is exactly the same as the 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/MLD control packets, if any received on the intra-subnet tunnel. SEMG should discard incoming SMET routes and IGMP/MLD joins from L2EVPN PEs, if it is not the DF for the incoming BD.

When [RFC9251] 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 the VPN-IP unicast route along with the 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

External multicast networks (multicast sources, multicast receivers and PIM routers) can be attached to any EVPN MVPN-capable 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 the 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 following flags for this document from the "Multicast Flags Extended Community " registry under the "Border Gateway Protocol (BGP) Extended Communities" registry group available at IANA-BGP-EC (https://www.iana.org/assignments/bgp-extended-communities/bgp-extended-communities.xhtml)

Bit	Name	Reference
5	SEMG	This document
6	EVPN/MVPN Seamless Interop Supported	This document

14. Security Considerations

All the security considerations in [RFC7432], [RFC6513], [RFC6514] and [RFC9251] apply directly to this document because this document leverages these RFCs control planes 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

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

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[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>.

[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>.

[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>.

[RFC9251] Sajassi, A., Thoria, S., Mishra, M., Patel, K., Drake, J., and W. Lin, "Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Proxies for Ethernet VPN (EVPN)", RFC 9251, DOI 10.17487/RFC9251, June 2022, <https://www.rfc-editor.org/info/rfc9251>.

16.2. Informative References

Appendix A. Supporting application with TTL value 1

It is possible that some deployments may have a host on the tenant domain that sends intra subnet 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 a few possible ways to support traffic having TTL value 1 or traffic that requires L2 bridging behavior. An implementation MAY support any of the following models.

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.

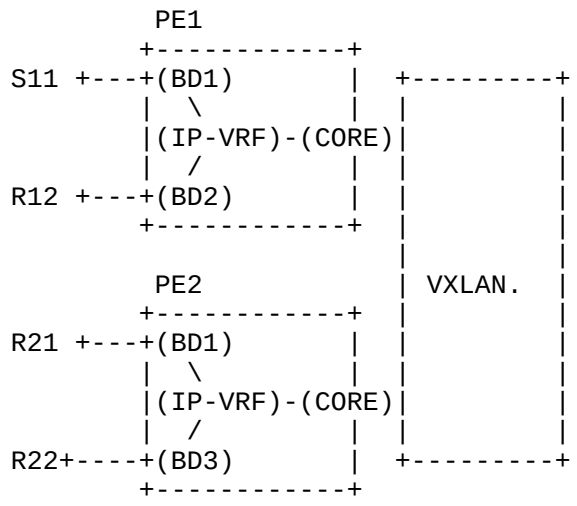

Figure 3: Intra-subnet bridging

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/MLD 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/MLD 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 the CORE interface.

PE2 inherits (*, G11) OIFs to (S11, G11) entry. While inheriting OIF, PE2 checks whether the source is attached to OIF’s subnet. OIF matching source subnet is added with a flag indicating bridge only interface. In the 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 the CORE interface. (PE1 retains the TTL and source-MAC). The traffic is encapsulated with VNI associated with the 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 an 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 a 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 Thiruvenkatasamy
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