BESS Working Group
Internet Draft
Category: Standard Track

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

Expires: January 06, 2020 July 05, 2019

Seamless Multicast Interoperability between EVPN and MVPN PEs
draft-sajassi-bess-evpn-mvpn-seamless-interop-04

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 COs toward 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 to IETF in full conformance with the provisions of BCP 78 and BCP 79.

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

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

Table of Contents

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 COs toward 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 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 results in non-optimum forwarding but also it can result in tromboing 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.

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).

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 documents 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

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.

L3VNI: A VNI in the tenant VRF, which is associated with the core facing interface.

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. The solution MUST prevent loop during steady and transient states just like EVPN baseline solution [RFC7432] and [RFC8365] for all multi-homing types.

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 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, then multicast traffic SHOULD only be forwarded to the remote PE which have receivers - i.e., if there are no receivers at a remote PE, the multicast traffic SHOULD NOT be forwarded to that PE and 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. IRB Unicast versus IRB Multicast

[EVPN-IRB] describes the operation for EVPN PEs in IRB mode for unicast traffic. The same IRB model used for unicast traffic in [EVPN-IRB], 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. However, there are some noticeable differences between the IRB operation for unicast traffic described in [EVPN-IRB] versus for multicast traffic described in this document. 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 (ie, 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 tenants'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. Emulated Virtual LAN Service

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 [EVPN-IGMP-PROXY] and [EVPN-PIM-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 an 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.

6. 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.

6.1. 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 [RFC 6513]. 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.

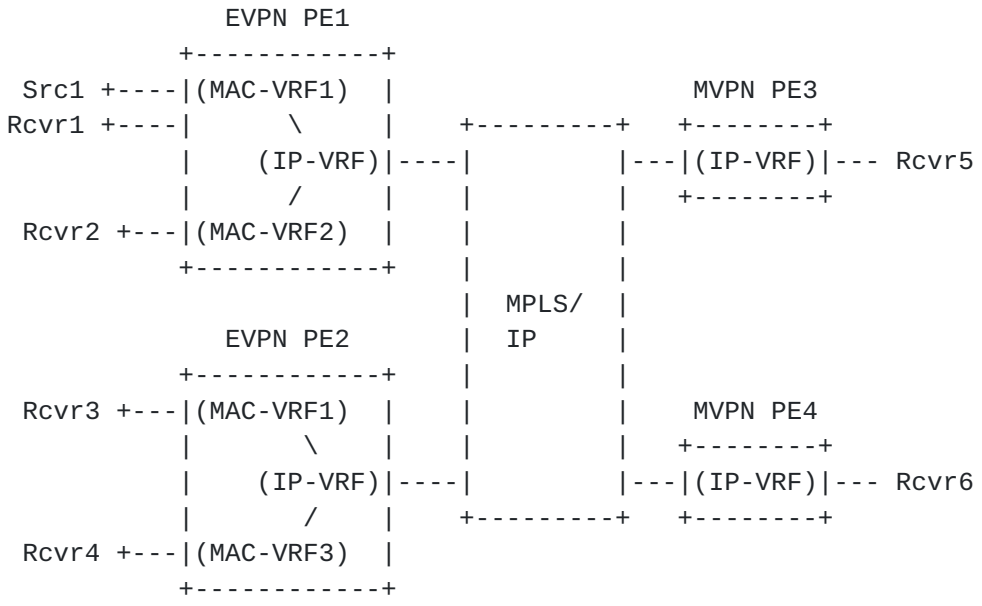

Figure-1: EVPN & MVPN PEs Seamless Interop

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.

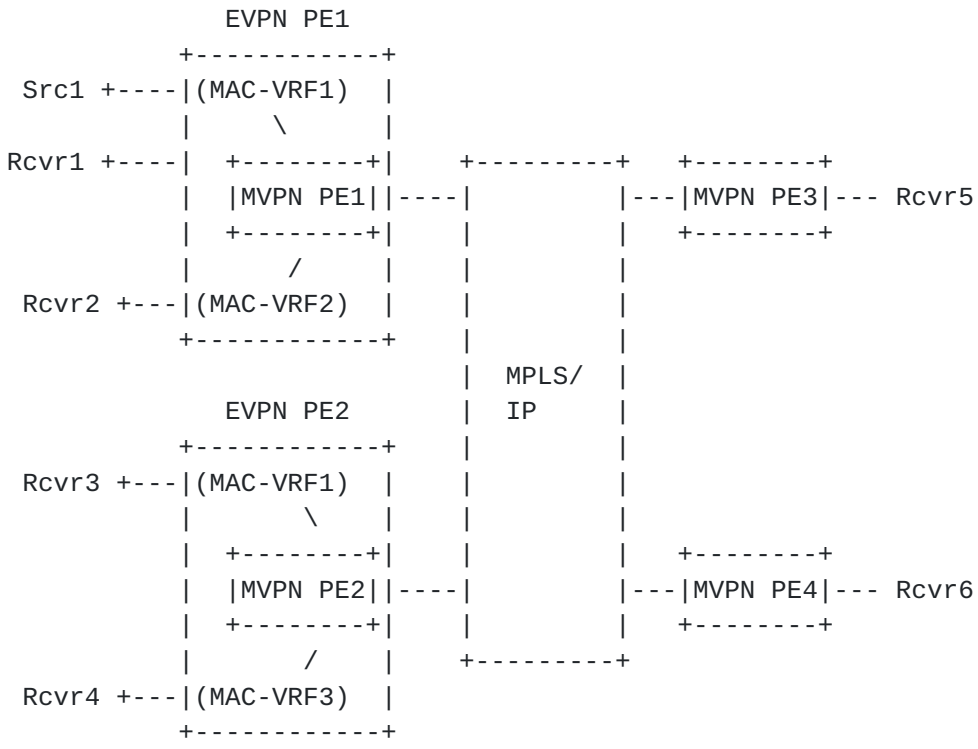

Figure-2: Modeling EVPN PEs as MVPN PEs

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.

1) Unicast route advertisements for IP multicast source
2) Multi-homing of IP multicast sources and receivers
3) Mobility for Tenant's sources and receivers
4) non-IP multicast traffic handling

6.2. 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 incase of seamless interop mode.

Section 6.6 discusses connecting EVPN and MVPN networks with gateway model. Section 9 extends seamless interop procedures to EVPN only fabrics as an IRB solution for multicast.

EVPN-PEs only need to advertise unicast routes using EVPN route-type 2 or route-type 5 and don't need to advertise IPVPN routes within EVPN only fabric. No L3VPN provisioning is needed between EVPN-PEs.

In gateway model, EVPN-PE advertises unicast routes as IPVPN routes along with VRI extended community for all multicast sources attached behind EVPN-PEs. All IPVPN routes SHOULD be summarized while adverting to MVPN-PEs.

6.3. 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.

6.3.1. 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 a Upstream Multicast Hop (UMH) which is the upstream PE that receives the IP multicast flow in case of Singe-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]. Any of the three algorithms described in that section works fine.

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 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.

6.3.2. 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 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 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 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 tunnel to other multi-homing PEs (only one other in case of dual-homing) for a given ES can be either fixed or on demand basis. If on-demand basis, then one of the other multi-homing PEs that is selected as a UMH upon receiving a join message from a downstream PE, sends a request to receive this multicast flow from the source multi-homing PE over the special intra-ES tunnel.

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 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 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. Split-horizon filtering for non-MPLS encapsulations such as VxLAN is described in section 9.2.2 that deals with a DC network that consists of only EVPN PEs.

6.4. 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 IP-VPN routes) is made by the EVPN PE where the source has moved to per section 6.3 above and unicast route withdraw (for both EVPN and IP-VPN 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.

6.5. Intra-Subnet BUM Traffic Handling

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 as well as non-IP multicast/broadcast traffic are sent per EVPN [RF7432] 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 give 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 8 for data plane operation).

6.6 EVPN and MVPN interworking with gateway model

The procedures specified in this document offers optimal multicast forwarding within a data center and also enables seamless interoperability of multicast traffic between EVPN and MVPN networks, when same tunnel types are used in the data plane.

There are few other use cases in connecting MVPN networks in the EVPN fabric other than seamless interop model, where gateway model is used to interconnect both networks.

Case1: All EVPN-PEs in the fabric can be made as MVPN exit points
Case2: MVPN network can be attached behind a EVPN PE or subset of EVPN-PEs
Case3: MVPN network (MVPN-PEs) which uses different tunnel model can be directly attached to EVPN fabric.

In gateway model, MVPN routes from one domain are terminated at the gateway PE and re-originated for another domain.

With use case 1 & 2, All PEs connected to an EVPN fabric can use one data plane to send & receive traffic within the fabric/data center. Also, IPVPN routes need not be advertised inside the fabric. Instead, PE where MVPN is terminated should advertise IPVPN as EVPN routes.

With use case 3, Fabric will get two copies per multicast flow, if receivers exist both MVPN and EVPN networks. (Two different data planes are used to send the traffic in the fabric; one for EVPN network and one for MVPN network).

7. Control Plane Operation

In seamless interop between EVPN and MVPN PEs, the control plane may need to setup the following three types of multicast tunnels. The first two are among EVPN PEs only but the third one is among EVPN and MVPN PEs.

1) Intra-ES IP multicast tunnel

2) Intra-subnet BUM tunnel

3) Inter-subnet IP multicast tunnel

7.1. Intra-ES/Intra-Subnet IP Multicast Tunnel

As described in section 6.3.2, 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/Intra-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 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], [EVPN-DF-Framework], 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 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 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 ingress replication is used for intra-ES 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/Intra-Subnet IP multicast tunnel among themselves.

7.2. Intra-Subnet BUM Tunnel

As the name implies, this tunnel is setup to carry BUM traffic for a given subnet/BD among EVNP PEs. In [RFC7432], this overlay tunnel is used for transmission of all BUM traffic including user IP multicast traffic. However, for multicast traffic handling in EVPN-IRB PEs, 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 user 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 tunnel, there SHALL be no duplication of multicast traffic over the network because they carry different types of multicast traffic - i.e., intra-ES tunnel among multi-homing PEs

carries only user IP multicast traffic; whereas, intra-subnet BUM tunnel carries link-local IP multicast traffic and BUM traffic (w/ non-IP multicast).

7.3. 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 user 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.

7.4. IGMP Hosts as TSes

If a tenant system which is an IGMP host is multi-homed to two or more EVPN PEs using All-Active multi-homing, then IGMP join and leave messages are synchronized between these EVPN PEs using EVPN IGMP Join Synch route (route type 7) and EVPN IGMP Leave Synch route (route type 8) per [IGMP-PROXY]. IGMP states are built in the corresponding

BDs of the multi-homing EVPN PEs. In [IGMP-PROXY] the DF PE for that BD originates an EVPN Selective Multicast Tag route (SMET route) route to other EVPN PEs. However, in here there is no need to use SMET because the 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]. 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.

7.5. TS PIM Routers

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].

8 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/Intra-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. 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.

8.1 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.

8.2 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 will form local bridging entry in MAC-VRF2 due to IGMP joins received from Rcvr3 and Rcvr4 respectively. PE2 now adds another OIF 'MAC-VRF2' to its existing routing entry. But there is no change in control plane states since its already sent MVPN route and no further signaling is

required. Also since Src1 is not part of MAC-VRF2 subnet, it is treated as routing OIF and hence MAC header gets modified as per normal procedures for routing. 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

9. 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). It should be noted that the scope of intra-subnet forwarding for the solution described in this document, is limited to a single EVPN PE for Single-Active multi-homing and to multi-homing PEs for All-Active multi-homing. In other words, 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, Section 11 discusses some options to support applications having TTL value 1. The procedure discussed in Section 11 may be used to support applications that require intra-subnet multicast traffic to be L2 forwarded.

9.1. 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 MVPN. 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 IP-VPN 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 as detailed in section 6.2.

The VRI received with EVPN route type 2 or 5 NLRI from source PE will be appended as an export route-target extended community. More details about handling of various types of local receivers are in section 10. 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.

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]. This is done in order to choose a single forwarder PE and to suppress receiving duplicate traffic. In such scenarios, the active multicast source is used by the receiver PEs to join the SPT if they have not received tenant (S,G) joins and by the RPT PEs to prune off the tenant (S,G) state from the RPT. The Source Active A-D route is also used for MVPN scenarios without tenant RP-shared trees. In such scenarios, the receiver PEs with tenant (*,G) states use the Source Active A-D route to know which upstream PEs with sources behind them to join per section 14 of [RFC6514] - i.e., to suppress joining Overlay shared tree.

9.2. 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.

9.2.1. 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.

9.2.2 VxLAN Encapsulation

In order to signal VXLAN, the corresponding BGP encapsulation extended community [TUNNEL-ENCAP] 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.

9.2.3. Other Encapsulation

In order to signal a different tunneling encapsulation such as NVGRE, GPE, or GENEVE the corresponding BGP encapsulation extended community [TUNNEL-ENCAP] 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 [TUNNEL-ENCAP], 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.

10. DCI with MPLS in WAN and VxLAN in DCs

This section describers 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 [INTERCON-EVPN] are not applicable here. Some aspects of the multi-homing between VxLAN DC networks and MPLS WAN is in common with [INTERCON-EVPN].

10.1. 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 [INTERCON-EVPN]. 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 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 7 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.

10.2. Data plane inter-connect

Traffic forwarding procedures on gateways are same as those described for PEs in section 5 and 6 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 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 sides. If there are more than one gateway between two domains, the multi-

homing procedures described in the following section must be considered so that incoming traffic from one side is not looped back to the other gateway.

The multicast traffic from local sources on each gateway flows to the other gateway with the preferred WAN encapsulation.

11. 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 6 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. Implementation MAY support any of the following model.

11.1. 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.

11.2. 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.

11.3. 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 brigading and routing (L2+L3). The procedure mentioned here leverages this ASIC capability.

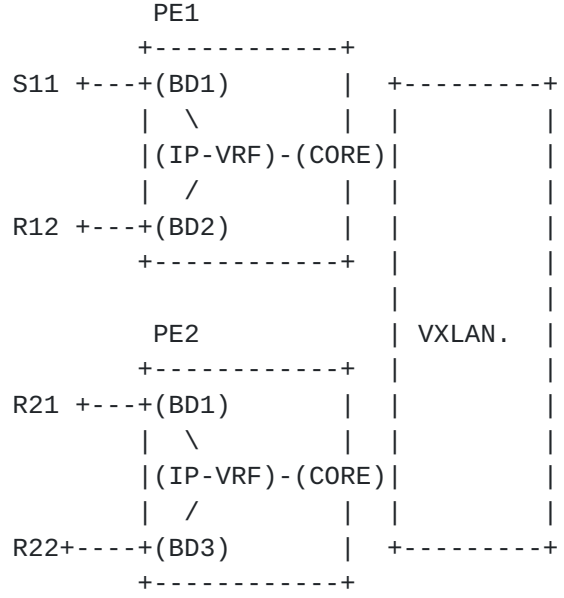

Figure 3 Intra-subnet bridging

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(L3VNI). 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.

12. 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 native EVPN-PEs(L2EVPN PE). The gateway device uses BUM tunnel when interworking with L2EVPN-PEs.

Interop procedure will be covered in the next version of the draft.

13. 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. Multicast network or PIM router can also be attached to any IRB enabled BDI interface or L3 enabled interface or set of interfaces. The fabric can be used as a Transit network. All PIM signaling is terminated at EVPN-PEs.

No additional procedures are required while connecting external multicast networks.

14. RP handling

This section describes various RP models for a tenant VRF. The RP model SHOULD be consistent across all EVPN-PEs for given group/group range in the tenant VRF.

14.1. Various RP deployment options

14.1.1. RP-less mode

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 [RFC 6514].

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.

14.1.2. Fabric anycast RP

In this model, anycast GW IP address is configured as RP in all EVPN-PE. When an EVPN-PE is operating in Fabric anycast-RP mode, an EVPN-PE MUST advertise all sources behind that PE to other EVPN PEs using procedure specified in [RFC 6514]. In this model, Sources may be directly attached to tenant BDs or sources may be attached behind a PIM router (In that case EVPN-PE learns source information due to PIM register terminating at RP interface at the tenant VRF side)

In RP-less mode and Fabric anycast RP mode, EVPN-PE operates SPT-only mode as per section 14 of RFC 6514.

14.1.3. Static RP

The procedure specified in this document supports configuring EVPN fabric with static RP. 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 RFC 6514.

14.1.4. Co-existence of Fabric anycast RP and external RP

External multicast network using its own RP may be connected to EVPN fabric operating with Fabric anycast RP mode. In this case, subset of EVPN-PEs may be designated as border leafs. Anycast RP may be configured between border leafs and external RP. Border leafs originates SA-AD routes for external sources towards fabric PEs. Border leaf acts as FHR for the sources inside the fabric. Configuration option may be provided to define the PE role as BL.

14.2. RP configuration options

PIM Bidir and PIM-SM ASM mode require Rendezvous point (RP) configuration, which acts as a shared root for a multicast shared tree. RP can be configured using static configuration or by using BSR

or Auto-RP procedures on the tenant VRF. This document only discusses static RP configuration. The use of BSR or Auto-RP procedure in the EVPN fabric is beyond the scope of this document.

15. IANA Considerations

IANA is requested to assign new flags in the "Multicast Flags Extended Community Flags" registry for the following.

- Seamless interop capable PE

16. Security Considerations

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

17. Acknowledgements

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

18. References

18.1. Normative References

[RFC7432] A. Sajassi, et al., "BGP MPLS Based Ethernet VPN", RFC 7432 , February 2015.

[RFC8365] A. Sajassi, et al., "A Network Virtualization Overlay Solution using EVPN", RFC 8365, February 2018.

[RFC6513] E. Rosen, et al., "Multicast in MPLS/BGP IP VPNs", RFC6513, February 2012.

[RFC6514] R. Aggarwal, et al., "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC6514, February 2012.

[EVPN-IRB] A. Sajassi, et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-03, February 2017.

[EVPN-IRB-MCAST] A. Rosen, et al., "EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding", draft-lin-bess-evpn-irb-

mcast-04, October 24, 2017.

18.2. Informative References

[RFC7080] A. Sajassi, et al., "Virtual Private LAN Service (VPLS) Interoperability with Provider Backbone Bridges", RFC 7080, December 2013.

[RFC7209] D. Thaler, et al., "Requirements for Ethernet VPN (EVPN)", RFC 7209, May 2014.

[RFC4389] A. Sajassi, et al., "Neighbor Discovery Proxies (ND Proxy)", RFC 4389, April 2006.

[RFC4761] K. Kompella, et al., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, Jauary 2007.

[INTERCON-EVPN] J. Rabadan, et al., "Interconnect Solution for EVPN Overlay networks", https://tools.ietf.org/html/draft-ietf-bess-dci-evpn-overlay-04, September 2016

[TUNNEL-ENCAPS] E. Rosen, et al. "The BGP Tunnel Encapsulation Attribute", https://tools.ietf.org/html/draft-ietf-idr-tunnel-encaps-06, work in progress, June 2017.

[EVPN-IGMP-PROXY] A. Sajassi, et. al., "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-01, work in progress, March 2018.

[EVPN-PIM-PROXY] J. Rabadan, et. al., "PIM Proxy in EVPN Networks", draft-skr-bess-evpn-pim-proxy-00, work in progress, July 3, 2017.

19. Authors' Addresses

Ali Sajassi
Csco
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
Avi Networks
Email: ashutosh@avinetworks.com

Luay Jalil
Verizon
Email: luay.jalil@verizon.com

Appendix A. Use Cases

A.1. DCs with only IGMP/MLD hosts w/o tenant router

In a EVPN network consisting of only IGMP/MLD hosts, PE's will receive IGMP (*, G) or (S, G) joins from their locally attached host and would originate MVPN C-Multicast Route Type 6 and 7 NLRI's respectively. As described in RFC 6514 these NLRI's are directed towards RP-PE for Type 6 or Source-PE for Type 7. In case of (*, G) join a Shared-Path Tree will be built in the core from RP-PE towards all Receiver-PE's. Once a Source starts to send Multicast data to specified multicast-group, the PE directly connected to Source will do PIM-registration with RP. Since there are existing receivers for the Group, RP will originate a PIM (S, G) join towards Source. This will

be converted to MVPN Type 7 NLRI by RP-PE. Please note that the router RP-PE would be the PE configured as RP (e.g., using static configuration or by using BSR or Auto-RP procedures). The detailed working of such protocols is beyond the scope of this document. Upon receiving Type 7 NLRI, Source-PE will include MVPN Tunnel in its Outgoing Interface List. Furthermore, Source-PE will follow the procedures in RFC-6514 to originate MVPN SA-AD route (RT 5) to avoid duplicate traffic and allow all Receiver-PE's to shift from Share-Tree to Shortest-Path-Tree rooted at Source-PE. Section 13 of [RFC6514] describes it.

However a network operator can chose to have only Shortest-Path-Tree built in MVPN core as described in section 14 of [RFC6514]. One way to achieve this, is for all PE's act as RP for its locally connected hosts and thus avoid sending any Shared-Tree Join (MVPN Type 6) into the core. In this scenario, there will be no PIM registration needed since all PE's are first-hop router as well as acting RP. Once a source starts to send multicast data, the PE directly connected to it originates Source-Active AD (RT 5) to all other PE's in network. Upon Receiving Source-Active AD route a PE must cache it in its local database and also look for any matching interest for (*, G) where G is the multicast group described in received Source-Active AD route. If it finds any such matching entry, it must originate a C-Multicast route (RT 7) in order to start receiving traffic from Source-PE. This procedure must be repeated on reception of any further Source-Active AD routes.

A.2. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-SSM

This scenario has multicast routers which can send PIM SSM (S, G) joins. Upon receiving these joins and if source described in join is learnt to be behind a MVPN peer PE, local PE will originate C-Multicast Join (RT 7) towards Source-PE. It is expected that PIM SSM group ranges are kept separate from ASM range for which IGMP hosts can send (*, G) joins. Hence both ASM and SSM groups shall operate without any overlap. There is no RP needed for SSM range groups and Shortest Path tree rooted at Source is built once a receiver interest is known.

A.3. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-ASM

This scenario includes reception of PIM (*, G) joins on PE's local AC. These joins are handled similar to IGMP (*, G) join as explained in sections above. Another interesting case can arise here is when one of the tenant routers can act as RP for some of the ASM Groups. In such scenario, a Upstream Multicast Hop (UMH) will be elected by other PE's in order to send C-Multicast Routes (RT 6). All procedures described in RFC 6513 with respect to UMH should be used to avoid traffic duplication due to incoherent selection of RP-PE by different Receiver-PE's.

A.4. DCs with mixed of IGMP/MLD hosts & multicast routers running PIM-Bidir

Creating Bidirectional (*, G) trees is useful when a customer wants least amount of control state in network. But on downside all receivers for a particular multicast group receive traffic from all sources sending to that group. However for the purpose of this document, all procedures as described in RFC 6513 and RFC 6514 apply when PIM-Bidir is used.