BESS
Internet-Draft
Intended status: Standards Track
Expires: August 17, 2018

W. Lin
Z. Zhang
J. Drake
E. Rosen, Ed.
Juniper Networks, Inc.
J. Rabadan
Nokia
A. Sajassi
Cisco Systems

February 13, 2018

EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding
draft-ietf-bess-evpn-irb-mcast-00

Abstract

Ethernet VPN (EVPN) provides a service that allows a single Local Area Network (LAN), i.e., a single IP subnet, to be distributed over multiple sites. The sites are interconnected by an IP or MPLS backbone. Intra-subnet traffic (either unicast or multicast) always appears to the endusers to be bridged, even when it is actually carried over the IP backbone. When a single "tenant" owns multiple such LANs, EVPN also allows IP unicast traffic to be routed between those LANs. This document specifies new procedures that allow inter-subnet IP multicast traffic to be routed among the LANs of a given tenant, while still making intra-subnet IP multicast traffic appear to be bridged. These procedures can provide optimal routing of the inter-subnet multicast traffic, and do not require any such traffic to leave a given router and then reenter that same router. These procedures also accommodate IP multicast traffic that needs to travel to or from systems that are outside the EVPN domain.

Status of This Memo

Table of Contents

1. Introduction

1.1. Background

Ethernet VPN (EVPN) [RFC7432] provides a Layer 2 VPN (L2VPN) solution, which allows IP backbone provider to offer ethernet service to a set of customers, known as "tenants".

In this section (as well as in [EVPN-IRB]), we provide some essential background information on EVPN.

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

1.1.1. Segments, Broadcast Domains, and Tenants

One of the key concepts of EVPN is the Broadcast Domain (BD). A BD is essentially an emulated ethernet. Each BD belongs to a single tenant. A BD typically consists of multiple ethernet "segments", and each segment may be attached to a different EVPN Provider Edge (EVPN-PE) router. EVPN-PE routers are often referred to as "Network Virtualization Endpoints" or NVEs. However, this document will use the term "EVPN-PE", or, when the context is clear, just "PE".

In this document, we use the term "segment" to mean the same as "Ethernet Segment" or "ES" in [RFC7432].

Attached to each segment are "Tenant Systems" (TSes). A TS may be any type of system, physical or virtual, host or router, etc., that can attach to an ethernet.

When two TSes are on the same segment, traffic between them does not pass through an EVPN-PE. When two TSes are on different segments of the same BD, traffic between them does pass through an EVPN-PE.

When two TSes, say TS1 and TS2 are on the same BD, then:

o If TS1 knows the MAC address of TS2, TS1 can send unicast ethernet frames to TS2. TS2 will receive the frames unaltered. That is, TS1's MAC address will be in the MAC Source Address field. If the

frame contains an IP datagram, the IP header is not modified in any way during the transmission.

- o If TS1 broadcasts an ethernet frame, TS2 will receive the unaltered frame.

- o If TS1 multicasts an ethernet frame, TS2 will receive the unaltered frame, as long as TS2 has been provisioned to receive ethernet multicasts.

When we say that TS2 receives an unaltered frame from TS1, we mean that the frame still contains TS1's MAC address, and that no alteration of the frame's payload has been done.

EVPN allows a single segment to be attached to multiple PE routers. This is known as "EVPN multi-homing". EVPN has procedures to ensure that a frame from a given segment, arriving at a particular PE router, cannot be returned to that segment via a different PE router. This is particularly important for multicast, because a frame arriving at a PE from a given segment will already have been seen by all systems on the segment that need to see it. If the frame were sent back to the originating segment, receivers on that segment would receive the packet twice. Even worse, the frame might be sent back to a PE, which could cause an infinite loop.

1.1.2. Inter-BD (Inter-Subnet) IP Traffic

If a given tenant has multiple BDs, the tenant may wish to allow IP communication among these BDs. Such a set of BDs is known as an "EVPN Tenant Domain" or just a "Tenant Domain".

If tenant systems TS1 and TS2 are not in the same BD, then they do not receive unaltered ethernet frames from each other. In order for TS1 to send traffic to TS2, TS1 encapsulates an IP datagram inside an ethernet frame, and uses ethernet to send these frames to an IP router. The router decapsulates the IP datagram, does the IP processing, and re-encapsulates the datagram for ethernet. The MAC source address field now has the MAC address of the router, not of TS1. The TTL field of the IP datagram should be decremented by exactly 1; this hides the structure of the provider's IP backbone from the tenants.

EVPN accommodates the need for inter-BD communication within a Tenant Domain by providing an integrated L2/L3 service for unicast IP traffic. EVPN's Integrated Routing and Bridging (IRB) functionality is specified in [EVPN-IRB]. Each BD in a Tenant Domain is assumed to be a single IP subnet, and each IP subnet within a a given Tenant Domain is assumed to be a single BD. EVPN's IRB functionality allows

IP traffic to travel from one BD to another, and ensures that proper IP processing (e.g., TTL decrement) is done.

A brief overview of IRB, including the notion of an "IRB interface", can be found in Appendix A. As explained there, an IRB interface is a sort of virtual interface connecting an L3 routing instance to a BD. A BD may have multiple attachment circuits (ACs) to a given PE, where each AC connects to a different ethernet segment of the BD. However, these ACs are not visible to the L3 routing function; from the perspective of an L3 routing instance, a PE has just one interface to each BD, viz., the IRB interface for that BD.

The "L3 routing instance" depicted in Appendix A is associated with a single Tenant Domain, and may be thought of as an IP-VRF for that Tenant Domain.

1.1.3. EVPN and IP Multicast

[EVPN-IRB] and [EVPN_IP_Prefix] cover inter-subnet (inter-BD) IP unicast forwarding, but they do not cover inter-subnet IP multicast forwarding.

[RFC7432] covers intra-subnet (intra-BD) ethernet multicast. The intra-subnet ethernet multicast procedures of [RFC7432] are used for ethernet Broadcast traffic, for ethernet unicast traffic whose MAC Destination Address field contains an Unknown address, and for ethernet traffic whose MAC Destination Address field contains an ethernet Multicast MAC address. These three classes of traffic are known collectively as "BUM traffic" (Broadcast/UnknownUnicast/ Multicast), and the procedures for handling BUM traffic are known as "BUM procedures".

[IGMP-Proxy] extends the intra-subnet ethernet multicast procedures by adding procedures that are specific to, and optimized for, the use of IP multicast within a subnet. However,that document does not cover inter-subnet IP multicast.

The purpose of this document is to specify procedures for EVPN that provide optimized IP multicast functionality within an EVPN tenant domain. This document also specifies procedures that allow IP multicast packets to be sourced from or destined to systems outside the Tenant Domain. We refer to the entire set of these procedures as "OISM" (Optimized Inter-Subnet Multicast) procedures.

In order to support the OISM procedures specified in this document, an EVPN-PE MUST also support [EVPN-IRB] and [IGMP-Proxy].

1.1.4. BDs, MAC-VRFS, and EVPN Service Models

[RFC7432] defines the notion of "MAC-VRF". A MAC-VRF contains one or more "Bridge Tables" (see section 3 of [RFC7432] for a discussion of this terminology), each of which represents a single Broadcast Domain.

In the IRB model (outlined in Appendix A) a L3 routing instance has one IRB interface per BD, NOT one per MAC-VRF. The procedures of this document are intended to work with all the EVPN service models. This document does not distinguish between a "Broadcast Domain" and a "Bridge Table", and will use the terms interchangeably (or will use the acronym "BD" to refer to either). The way the BDs are grouped into MAC-VRFs is not relevant to the procedures specified in this document.

Section 6 of [RFC7432] also defines several different EVPN service models:

o In the "vlan-based service", each MAC-VRF contains one "bridge table", where the bridge table corresponds to a particular Virtual LAN (VLAN). (See section 3 of [RFC7432] for a discussion of this terminology.) Thus each VLAN is treated as a BD.

o In the "vlan bundle service", each MAC-VRF contains one bridge table, where the bridge table corresponds to a set of VLANs. Thus a set of VLANs are treated as constituting a single BD.

o In the "vlan-aware bundle service", each MAC-VRF may contain multiple bridge tables, where each bridge table corresponds to one BD. If a MAC-VRF contains several bridge tables, then it corresponds to several BDs.

The procedures of this document are intended to work for all these service models.

1.2. Need for EVPN-aware Multicast Procedures

Inter-subnet IP multicast among a set of BDs can be achieved, in a non-optimal manner, without any specific EVPN procedures. For instance, if a particular tenant has n BDs among which he wants to send IP multicast traffic, he can simply attach a conventional multicast router to all n BDs. Or more generally, as long as each BD has at least one IP multicast router, and the IP multicast routers communicate multicast control information with each other, conventional IP multicast procedures will work normally, and no special EVPN functionality is needed.

However, that technique does not provide optimal routing for multicast. In conventional multicast routing, for a given multicast flow, there is only one multicast router on each BD that is permitted to send traffic of that flow to the BD. If that BD has receivers for a given flow, but the source of the flow is not on that BD, then the flow must pass through that multicast router. This leads to the "hair-pinning" problem described (for unicast) in Appendix A.

For example, consider an (S,G) flow that is sourced by a TS S and needs to be received by TSes R1 and R2. Suppose S is on a segment of BD1, R1 is on a segment of BD2, but both are attached to PE1. Suppose also that the tenant has a multicast router, attached to a segment of BD1 and to a segment of BD2. However, the segments to which that router is attached are both attached to PE2. Then the flow from S to R would have to follow the path: S-->PE1-->PE2-->Tenant Multicast Router-->PE2-->PE1-->R1. Obviously, the path S-->PE1-->R would be preferred.

Now suppose that there is a second receiver, R2. R2 is attached to a third BD, BD3. However, it is attached to a segment of BD3 that is attached to PE1. And suppose also that the Tenant Multicast Router is attached to a segment of BD3 that attaches to PE2. In this case, the Tenant Multicast Router will make two copies of the packet, one for BD2 and one for BD3. PE2 will send both copies back to PE1. Not only is the routing sub-optimal, but PE2 sends multiple copies of the same packet to PE1. This is a further sub-optimality.

This is only an example; many more examples of sub-optimal multicast routing can easily be given. To eliminate sub-optimal routing and extra copies, it is necessary to have a multicast solution that is EVPN-aware, and that can use its knowledge of the internal structure of a Tenant Domain to ensure that multicast traffic gets routed optimally. The procedures of this document allow us to avoid all such sub-optimalities when routing inter-subnet multicasts within a Tenant Domain.

1.3. Additional Requirements That Must be Met by the Solution

In addition to providing optimal routing of multicast flows within a Tenant Domain, the EVPN-aware multicast solution is intended to satisfy the following requirements:

o The solution must integrate well with the procedures specified in [IGMP-Proxy]. That is, an integrated set of procedures must handle both intra-subnet multicast and inter-subnet multicast.

o With regard to intra-subnet multicast, the solution MUST maintain the integrity of multicast ethernet service. This means:

  * If a source and a receiver are on the same subnet, the MAC source address (SA) of the multicast frame sent by the source will not get rewritten.

  * If a source and a receiver are on the same subnet, no IP processing of the ethernet payload is done. The IP TTL is not decremented, the header checksum is not changed, no fragmentation is done, etc.

- On the other hand, if a source and a receiver are on different subnets, the frame received by the receiver will not have the MAC Source address of the source, as the frame will appear to have come from a multicast router. Also, proper processing of the IP header is done, e.g., TTL decrement by 1, header checksum modification, possibly fragmentation, etc.

- If a Tenant Domain contains several BDs, it MUST be possible for a multicast flow (even when the multicast group address is an "any source multicast" (ASM) address), to have sources in one of those BDs and receivers in one or more of the other BDs, without requiring the presence of any system performing PIM Rendezvous Point (RP) functions ([RFC7761]). Multicast throughout a Tenant Domain must not require the tenant systems to be aware of any underlying multicast infrastructure.

- Sometimes a MAC address used by one TS on a particular BD is also used by another TS on a different BD. Inter-subnet routing of multicast traffic MUST NOT make any assumptions about the uniqueness of a MAC address across several BDs.

- If two EVPN-PEs attached to the same Tenant Domain both support the OISM procedures, each may receive inter-subnet multicasts from the other, even if the egress PE is not attached to any segment of the BD from which the multicast packets are being sourced. It MUST NOT be necessary to provision the egress PE with knowledge of the ingress BD.

- There must be a procedure that that allows EVPN-PE routers supporting OISM procedures to send/receive multicast traffic to/from EVPN-PE routers that support only [RFC7432], but that do not support the OISM procedures or even the procedures of [EVPN-IRB]. However, when interworking with such routers (which we call "non-OISM PE routers"), optimal routing may not be achievable.

- It MUST be possible to support scenarios in which multicast flows with sources inside a Tenant Domain have "external" receivers, i.e., receivers that are outside the domain. It must also be possible to support scenarios where multicast flows with external

sources (sources outside the Tenant Domain) have receivers inside the domain.

This presupposes that unicast routes to multicast sources outside the domain can be distributed to EVPN-PEs attached to the domain, and that unicast routes to multicast sources within the domain can be distributed outside the domain.

Of particular importance are the scenario in which the external sources and/or receivers are reachable via L3VPN/MVPN, and the scenario in which external sources and/or receivers are reachable via IP/PIM.

The solution for external interworking MUST allow for deployment scenarios in which EVPN does not need to export a host route for every multicast source.

- The solution for external interworking must not presuppose that the same tunneling technology is used within both the EVPN domain and the external domain. For example, MVPN interworking must be possible when MVPN is using MPLS P2MP tunneling, and EVPN is using Ingress Replication or VXLAN tunneling.

- The solution must not be overly dependent on the details of a small set of use cases, but must be adaptable to new use cases as they arise. (That is, the solution must be robust.)

1.4. Terminology

In this document we make frequent use of the following terminology:

- OISM: Optimized Inter-Subnet Multicast. EVPN-PEs that follow the procedures of this document will be known as "OISM" PEs. EVPN-PEs that do not follow the procedures of this document will be known as "non-OISM" PEs.

- IP Multicast Packet: An IP packet whose IP Destination Address field is a multicast address that is not a link-local address. (Link-local addresses are IPv4 addresses in the 224/8 range and IPv6 address in the FF02/16 range.)

- IP Multicast Frame: An ethernet frame whose payload is an IP multicast packet (as defined above).

- (S,G) Multicast Packet: An IP multicast packet whose IP Source Address field contains S and whose IP Destination Address field contains G.

o (S,G) Multicast Frame: An IP multicast frame whose payload contains S in its IP Source Address field and G in its IP Destination Address field.

o Broadcast Domain (BD): an emulated ethernet, such that two systems on the same BD will receive each other's link-local broadcasts.

Note that EVPN supports models in which a single EVPN Instance (EVI) contains only one BD, and models in which a single EVI contains multiple BDs. Both models are supported by this draft. However, a given BD belongs to only one EVI.

o Designated Forwarder (DF). As defined in [RFC7432], an ethernet segment may be multi-homed (attached to more than one PE). An ethernet segment may also contain multiple BDs, of one or more EVIs. For each such EVI, one of the PEs attached to the segment becomes that EVI's DF for that segment. Since a BD may belong to only one EVI, we can speak unambiguously of the BD's DF for a given segment.

When the text makes it clear that we are speaking in the context of a given BD, we will frequently use the term "a segment's DF" to mean the given BD's DF for that segment.

o AC: Attachment Circuit. An AC connects the bridging function of an EVPN-PE to an ethernet segment of a particular BD. ACs are not visible at the router (L3) layer.

o L3 Gateway: An L3 Gateway is a PE that connects an EVPN tenant domain to an external multicast domain by performing both the OISM procedures and the Layer 3 multicast procedures of the external domain.

o PEG (PIM/EVPN Gateway): A L3 Gateway that connects an EVPN tenant domain to an external multicast domain whose Layer 3 multicast procedures are those of PIM ([RFC7761]).

o MEG (MVPN/EVPN Gateway): A L3 Gateway that connects an EVPN tenant domain to an external multicast domain whose Layer 3 multicast procedures are those of MVPN ([RFC6513], [RFC6514]).

o IPMG (IP Multicast Gateway): A PE that is used for interworking OISM EVPN-PEs with non-OISM EVPN-PEs.

o DR (Designated Router): A PE that has special responsibilities for handling multicast on a given BD.

o Use of the "C-" prefix. In many documents on VPN multicast, the prefix "C-" appears before any address or wildcard that refers to an address or addresses in a tenant's address space, rather than to an address of addresses in the address space of the backbone network. This document omits the "C-" prefix in many cases where it is clear from the context that the reference is to the tenant's address space.

This document also assumes familiarity with the terminology of [RFC4364], [RFC6514], [RFC7432], [RFC7761], [IGMP-Proxy], [EVPN_IP_Prefix] and [EVPN-BUM].

1.5. Model of Operation: Overview

1.5.1. Control Plane

In this section, and in the remainder of this document, we assume the reader is familiar with the procedures of IGMP/MLD (see [RFC2236] and [RFC2710]), by which hosts announce their interest in receiving particular multicast flows.

Consider a Tenant Domain consisting of a set of k BDs: BD1, ..., BDk. To support the OISM procedures, each Tenant Domain must also be associated with a "Supplementary Broadcast Domain" (SBD). An SBD is treated in the control plane as a real BD, but it does not have any ACs. The SBD has several uses, that will be described later in this document. (See Section 2.1.)

Each PE that attaches to one or more of the BDs in a given tenant domain will be provisioned to recognize that those BDs are part of the same Tenant Domain. Note that a given PE does not need to be configured with all the BDs of a given Tenant Domain. In general, a PE will only be attached to a subset of the BDs in a given Tenant Domain, and will be configured only with that subset of BDs. However, each PE attached to a given Tenant Domain must be configured with the SBD for that Tenant Domain.

Suppose a particular segment of a particular BD is attached to PE1. [RFC7432] specifies that PE1 must originate an Inclusive Multicast Ethernet Tag (IMET) route for that BD, and that the IMET must be propagated to all other PEs attached to the same BD. If the given segment contains a host that has interest in receiving a particular multicast flow, either an (S,G) flow or a (*,G) flow, PE1 will learn of that interest by participating in the IGMP/MLD procedures, as specified in [IGMP-Proxy]. In this case, we will say that:

o PE1 is interested in receiving the flow;

o The AC attaching the interested host to PE1 is also said to be interested in the flow;

o The BD containing an AC that is interested in a particular flow is also said to be interested in that flow.

Once PE1 determines that it has interest in receiving a particular flow or set of flows, it uses the procedures of [IGMP-Proxy] to advertise its interest in those flows. It advertises its interest in a given flow by originating a Selective Multicast Ethernet Tag (SMET) route. An SMET route is propagated to the other PEs that attach to the same BD.

OISM PEs MUST follow the procedures of [IGMP-Proxy]. In this document, we extend the procedures of [IGMP-Proxy] so that IMET and SMET routes for a particular BD are distributed not just to PEs that attach to that BD, but to PEs that attach to any BD in the Tenant Domain.

In this way, each PE attached to a given Tenant Domain learns, from each other PE attached to the same Tenant Domain, the set of flows that are of interest to each of those other PEs.

An OISM PE that is provisioned with several BDs in the same Tenant Domain may originate an IMET route for each such BD. To indicate its support of [IGMP-Proxy], it MUST attach the EVPN Multicast Flags Extended Community to each such IMET route.

Suppose PE1 is provisioned with both BD1 and BD2, and is provisioned to consider them to be part of the same Tenant Domain. It is possible that PE1 will receive from PE2 both an IMET route for BD1 and an IMET route for BD2. If either of these IMET routes has the EVPN Multicast Flags Extended Community, PE1 MUST assume that PE2 is supporting the procedures of [IGMP-Proxy] for ALL BDs in the Tenant Domain.

If a PE supports OISM functionality, it MUST indicate that by attaching an "OISM-supported" flag or Extended Community (EC) to all its IMET routes. (Details to be specified in next revision.) An OISM PE SHOULD attach this flag or EC to all the IMET routes it originates. However, if PE1 imports IMET routes from PE2, and at least one of PE2's IMET routes indicates that PE2 is an OISM PE, PE1 will assume that PE2 is following OISM procedures.

1.5.2. Data Plane

Suppose PE1 has an AC to a segment in BD1, and PE1 receives from that AC an (S,G) multicast frame (as defined in Section 1.4).

There may be other ACs of PE1 on which TSes have indicated an interest (via IGMP/MLD) in receiving (S,G) multicast packets. PE1 is responsible for sending the received multicast packet out those ACs. There are two cases to consider:

- Intra-Subnet Forwarding: In this case, an attachment AC with interest in (S,G) is connected to a segment that is part of the source BD, BD1. If the segment is not multi-homed, or if PE1 is the Designated Forwarder (DF) (see [RFC7432]) for that segment, PE1 sends the multicast frame on that AC without changing the MAC SA. The IP header is not modified at all; in particular, the TTL is not decremented.

- Inter-Subnet Forwarding: An AC with interest in (S,G) is connected to a segment of BD2, where BD2 is different than BD1. If PE1 is the DF for that segment (or if the segment is not multi-homed), PE1 decapsulates the IP multicast packet, performs any necessary IP processing (including TTL decrement), then re-encapsulates the packet appropriately for BD2. PE1 then sends the packet on the AC. Note that after re-encapsulation, the MAC SA will be PE1's MAC address on BD2. The IP TTL will have been decremented by 1.

In addition, there may be other PEs that are interested in (S,G) traffic. Suppose PE2 is such a PE. Then PE1 tunnels a copy of the IP multicast frame (with its original MAC SA, and with no alteration of the payload's IP header). The tunnel encapsulation contains information that PE2 can use to associate the frame with a source BD. If the source BD is BD1:

- If PE2 is attached to BD1, the tunnel encapsulation used to send the frame to PE2 will cause PE2 to identify BD1 as the source BD.

- If PE2 is not attached to BD1, the tunnel encapsulation used to send the frame to PE2 will cause PE2 to identify the SBD as the source BD.

The way in which the tunnel encapsulation identifies the source BD is of course dependent on the type of tunnel that is used. This will be specified later in this document.

When PE2 receives the tunneled frame, it will forward it on any of its ACs that have interest in (S,G).

If PE2 determines from the tunnel encapsulation that the source BD is BD1, then

o For those ACs that connect PE2 to BD1, the intra-subnet forwarding procedure described above is used, except that it is now PE2, not PE1, carrying out that procedure. Unmodified EVPN procedures from [RFC7432] are used to ensure that a packet originating from a multi-homed segment is never sent back to that segment.

o For those ACs that do not connect to BD1, the inter-subnet forwarding procedure described above is used, except that it is now PE2, not PE1, carrying out that procedure.

If the tunnel encapsulation identifies the source BD as the SBD, PE2 applies the inter-subnet forwarding procedures described above to all of its ACs that have interest in the flow.

These procedures ensure that an IP multicast frame travels from its ingress PE to all egress PEs that are interested in receiving it. While in transit, the frame retains its original MAC SA, and the payload of the frame retains its original IP header. Note that in all cases, when an IP multicast packet is sent from one BD to another, these procedures cause its TTL to be decremented by 1.

So far we have assumed that an IP multicast packet arrives at its ingress PE over an AC that belongs to one of the BDs in a given Tenant Domain. However, it is possible for a packet to arrive at its ingress PE in other ways. Since an EVPN-PE supporting IRB has an IP-VRF, it is possible that the IP-VRF will have a "VRF interface" that is not an IRB interface. For example, there might be a VRF interface that is actually a physical link to an external ethernet switch, or to a directly attached host, or to a router. When an EVPN-PE, say PE1, receives a packet through such means, we will say that the packet has an "external" source (i.e., a source "outside the tenant domain"). There are also other scenarios in which a multicast packet might have an external source, e.g., it might arrive over an MVPN tunnel from an L3VPN PE. In such cases, we will still refer to PE1 as the "ingress EVPN-PE".

When an EVPN-PE, say PE1, receives an externally sourced multicast packet, and there are receivers for that packet inside the Tenant Domain, it does the following:

o Suppose PE1 has an AC in BD1 that has interest in (S,G). Then PE1 encapsulates the packet for BD1, filling in the MAC SA field with the MAC address of PE1 itself on BD1. It sends the resulting frame on the AC.

o Suppose some other EVPN-PE, say PE2, has interest in (S,G). PE1 encapsulates the packet for ethernet, filling in the MAC SA field with PE1's own MAC address on the SBD. PE1 then tunnels the packet to PE2. The tunnel encapsulation will identify the source BD as the SBD. Since the source BD is the SBD, PE2 will know to treat the frame as an inter-subnet multicast.

When ingress replication is used to transmit IP multicast frames from an ingress EVPN-PE to a set of egress PEs, then of course the ingress PE has to send multiple copies of the frame. Each copy is the original ethernet frame; decapsulation and IP processing take place only at the egress PE.

If a Point-to-Multipoint (P2MP) tree or BIER ([EVPN-BIER]) is used to transmit an IP multicast frame from an ingress PE to a set of egress PEs, then the ingress PE only has to send one copy of the frame to each of its next hops. Again, each egress PE receives the original frame and does any necessary IP processing.

2. Detailed Model of Operation

The model described in Section 1.5.2 can be expressed more precisely using the notion of "IRB interface" (see Appendix A). However, this requires that the semantics of the IRB interface be modified for multicast packets. It is also necessary to have an IRB interface that connects the L3 routing instance of a particular Tenant Domain (in a particular PE) to the SBD of that Tenant Domain.

In this section we assume that PIM is not enabled on the IRB interfaces. In general, it is not necessary to enable PIM on the IRB interfaces unless there are PIM routers on one of the Tenant Domain's BDs, or unless there is some other scenario requiring a Tenant Domain's L3 routing instance to become a PIM adjacency of some other system. These cases will be discussed in Section 7.

2.1. Supplementary Broadcast Domain

Suppose a given Tenant Domain contains three BDs (BD1, BD2, BD3) and two PEs (PE1, PE2). PE1 attaches to BD1 and BD2, while PE2 attaches to BD2 and BD3.

To carry out the procedures described above, all the PEs attached to the Tenant Domain must be provisioned to have the SBD for that tenant domain. An RT must be associated with the SBD, and provisioned on each of those PEs. We will refer to that RT as the "SBD-RT".

A Tenant Domain is also configured with an IP-VRF ([EVPN-IRB]), and the IP-VRF is associated with an RT. This RT MAY be the same as the SBD-RT.

Suppose an (S,G) multicast frame originating on BD1 has a receiver on BD3. PE1 will transmit the packet to PE2 as a frame, and the encapsulation will identify the frame's source BD as BD1. Since PE2 is not provisioned with BD1, it will treat the packet as if its source BD were the SBD. That is, a packet can be transmitted from BD1 to BD3 even though its ingress PE is not configured for BD3, and/or its egress PE is not configured for BD1.

EVPN supports service models in which a given EVPN Instance (EVI) can contain only one BD. It also supports service models in which a given EVI can contain multiple BDs. The SBD can be treated either as its own EVI, or it can be treated as one BD within an EVI that contains multiple BDs. The procedures specified in this document accommodate both cases.

2.2. When is a Route About/For/From a Particular BD

In this document, we will frequently say that a particular route is "about" a particular BD, or is "from" a particular BD, or is "for" a particular BD or is "related to" a particular BD. These terms are used interchangeably. In this section, we explain exactly what that means.

In EVPN, each BD is assigned an RT. In some service models, each BD is assigned a unique RT. In other service models, a set of BDs (all in the same Tenant Domain) may be assigned the same RT. (An RT is actually assigned to a MAC-VRF, and hence is shared by all the BDs that share the MAC-VRF.) The RT is a BGP extended community that may be attached to the BGP routes used by the EVPN control plane.

In those service models that allow a set of BDs to share a single RT, each BD is assigned a non-zero Tag ID. The Tag ID appears in the Network Layer Reachability Information (NLRI) of many of the BGP routes that are used by the EVPN control plane.

A route is about a particular BD if it carries the RT that has been assigned to that BD, and its NLRI contains the Tag ID that has been assigned to that BD.

Note that a route that is about a particular BD may also carry additional RTs.

2.3. Use of IRB Interfaces at Ingress PE

When an (S,G) multicast frame is received from an AC belonging to a particular BD, say BD1:

1. The frame is sent unchanged to other EVPN-PEs that are interested in (S,G) traffic. The encapsulation used to send the frame to the other EVPN-PEs depends on the tunnel type being used for multicast transmission. (For our purposes, we consider Ingress Replication (IR), Assisted Replication (AR) and BIER to be "tunnel types", even though IR, AR and BIER do not actually use P2MP tunnels.) At the egress PE, the source BD of the frame can be inferred from the tunnel encapsulation. If the egress PE is not attached to the real source BD, it will infer that the source BD is the SBD.

   Note that the the inter-PE transmission of a multicast frame among EVPN-PEs of the same Tenant Domain does NOT involve the IRB interfaces, as long as the multicast frame was received over an AC attached to one of the Tenant Domain's BDs.

2. The frame is also sent up the IRB interface that attaches BD1 to the Tenant Domain's L3 routing instance in this PE. That is, the L3 routing instance, behaving as if it were a multicast router, receives the IP multicast frames that arrive at the PE from its local ACs. The L3 routing instance decapsulates the frame's payload to extract the IP multicast packet, decrements the IP TTL, adjusts the header checksum, and does any other necessary IP processing (e.g., fragmentation).

3. The L3 routing instance keeps track of which BDs have local receivers for (S,G) traffic. (A "local receiver" is a tenant system, reachable via a local attachment circuit that has expressed interest in (S,G) traffic.) If the L3 routing instance has an IRB interface to BD2, and it knows that BD2 has a LOCAL receiver interested in (S,G) traffic, it encapsulates the packet in an ethernet header for BD2, putting its own MAC address in the MAC SA field. Then it sends the packet down the IRB interface to BD2.

If a packet is sent from the L3 routing instance to a particular BD via the IRB interface (step 3 in the above list), and if the BD in question is NOT the SBD, the packet is sent ONLY to LOCAL ACs of that BD. If the packet needs to go to other PEs, it has already been sent to them in step 1. Note that this is a change in the IRB interface semantics from what is described in [EVPN-IRB] and Figure 2.

Existing EVPN procedures ensure that a packet is not sent by a given PE to a given locally attached segment unless the PE is the DF for that segment. Those procedures also ensure that a packet is never sent by a PE to its segment of origin. Thus EVPN segment multi-homing is fully supported; duplicate delivery to a segment or looping on a segment are thereby prevented, without the need for any new procedures to be defined in this document.

What if an IP multicast packet is received from outside the tenant domain? For instance, perhaps PE1's IP-VRF for a particular tenant domain also has a physical interface leading to an external switch, host, or router, and PE1 receives an IP multicast packet or frame on that interface. Or perhaps the packet is from an L3VPN, or a different EVPN Tenant Domain.

Such a packet is first processed by the L3 routing instance, which decrements TTL and does any other necessary IP processing. Then the packet is sent into the Tenant Domain by sending it down the IRB interface to the SBD of that Tenant Domain. This requires encapsulating the packet in an ethernet header, with the PE's own MAC address, on the SBD, in the MAC SA field.

An IP multicast packet sent by the L3 routing instance down the IRB interface to the SBD is treated as if it had arrived from a local AC, and steps 1-3 are applied. Note that the semantics of sending a packet down the IRB interface to the SBD are thus slightly different than the semantics of sending a packet down other IRB interfaces. IP multicast packets sent down the SBD's IRB interface may be distributed to other PEs, but IP multicast packets sent down other IRB interfaces are distributed only to local ACs.

If a PE sends a link-local multicast packet down the SBD IRB interface, that packet will be distributed (as an ethernet frame) to other PEs of the Tenant Domain, but will not appear on any of the actual BDs.

2.4. Use of IRB Interfaces at an Egress PE

Suppose an egress EVPN-PE receives an (S,G) multicast frame from the frame's ingress EVPN-PE. As described above, the packet will arrive as an ethernet frame over a tunnel from the ingress PE, and the tunnel encapsulation will identify the source BD of the ethernet frame.

We define the notion of the frame's "inferred source BD" as follows. If the egress PE is attached to the actual source BD, the actual source BD is the inferred source BD. If the egress PE is not attached to the actual source BD, the inferred source BD is the SBD.

The egress PE now takes the following steps:

1. If the egress PE has ACs belonging to the inferred source BD of the frame, it sends the frame unchanged to any ACs of that BD that have interest in (S,G) packets. The MAC SA of the frame is not modified, and the IP header of the frame's payload is not modified in any way.

2. The frame is also sent to the L3 routing instance by being sent up the IRB interface that attaches the L3 routing instance to the inferred source BD. Steps 2 and 3 of Section 2.3 are then applied.

2.5. Announcing Interest in (S,G)

[IGMP-Proxy] defines the procedures used by an egress PE to announce its interest in a multicast flow or set of flows. This is done by originating an SMET route. If an egress PE determines it has LOCAL receivers in a particular BD that are interested in a particular set of flows, it originates one or more SMET routes for that BD. The SMET route specifies a flow or set of flows, and identifies the egress PE. The SMET route is specific to a particular BD. A PE that originates an SMET route is announcing "I have receivers for (S,G) or (*,G) in BD-x".

In [IGMP-Proxy], an SMET route for a particular BD carries a Route Target (RT) that ensures it will be distributed to all PEs that are attached to that BD. In this document, it is REQUIRED that an SMET route also carry the RT that is assigned to the SBD. This ensures that every ingress PE attached to a particular Tenant Domain will learn of all other PEs (attached to the same Tenant Domain) that have interest in a particular set of flows. Note that it is not necessary for the ingress PE to have any BDs other than the SBD in common with the egress PEs.

Since the SMET routes from any BD in a given Tenant Domain are propagated to all PEs of that Tenant Domain, an (S,G) receiver on one BD can receive (S,G) packets that originate in a different BD. Within an EVPN domain, a given IP source address can only be on one BD. Therefore inter-subnet multicasting can be done, within the Tenant Domain, without requiring any Rendezvous Points, shared trees, or other complex aspects of multicast routing infrastructure. (Note that while the MAC addresses do not have to be unique across all the BDs in a Tenant Domain, the IP addresses to have to be unique across all those BDs.)

If some PE attached to the Tenant Domain does not support [IGMP-Proxy], it will be assumed to be interested in all flows. Whether a

particular remote PE supports [IGMP-Proxy] is determined by the presence of the Multicast Flags Extended Community in its IMET route; this is specified in [IGMP-Proxy].)

2.6. Tunneling Frames from Ingress PE to Egress PEs

[RFC7432] specifies the procedures for setting up and using "BUM tunnels". A BUM tunnel is a tunnel used to carry traffic on a particular BD if that traffic is (a) broadcast traffic, or (b) unicast traffic with an unknown MAC DA, or (c) ethernet multicast traffic.

This document allows the BUM tunnels to be used as the default tunnels for transmitting intra-subnet IP multicast frames. It also allows a separate set of tunnels to be used, instead of the BUM tunnels, as the default tunnels for carrying intra-subnet IP multicast frames. Let's call these "IP Multicast Tunnels".

When the tunneling is done via Ingress Replication or via BIER, this difference is of no significance. However, when P2MP tunnels are used, there is a significant advantages to having separate IP multicast tunnels.

It is desirable for an ingress PE to transmit a copy of a given (S,G) multicast frame on only one tunnel. All egress PEs interested in (S,G) packets must then join that tunnel. If the source BD/PE for an (S,G) packet is BD1/PE1, and PE2 has receivers for (S,G) on BD2, PE2 must join the P2MP LSP on which PE1 transmits the frame. PE2 must join this P2MP LSP even if PE2 is not attached to the source BD (BD1). If PE1 were transmitting the multicast frame on its BD1 BUM tunnel, then PE2 would have to join the BD1 BUM tunnel, even though PE2 has no BD1 attachment circuits. This would cause PE2 to pull all the BUM traffic from BD1, most of which it would just have to discard. Thus we RECOMMEND that the default IP multicast tunnels be distinct from the BUM tunnels.

Whether or not the default IP multicast tunnels are distinct from the BUM tunnels, selective tunnels for particular multicast flows can still be used. Traffic sent on a selective tunnel would not be sent on the default tunnel.

Notwithstanding the above, link local IP multicast traffic MUST always be carried on the BUM tunnels, and ONLY on the BUM tunnels. Link local IP multicast traffic consists of IPv4 traffic with a destination address prefix of 224/8 and IPv6 traffic with a destination address prefix of FF02/16. In this document, the terms "IP multicast packet" and "IP multicast frame" are defined in Section 1.4 so as to exclude the link-local traffic.

2.7. Advanced Scenarios

There are some deployment scenarios that require special procedures:

1. Some multicast sources or receivers are attached to PEs that support [RFC7432], but do not support this document or [EVPN-IRB]. To interoperate with these "non-OISM PEs", it is necessary to have one or more gateway PEs that interface the tunnels discussed in this document with the BUM tunnels of the legacy PEs. This is discussed in Section 5.

2. Sometimes multicast traffic originates from outside the EVPN domain, or needs to be sent outside the EVPN domain. This is discussed in Section 6. An important special case of this, integration with MVPN, is discussed in Section 6.1.2.

3. In some scenarios, one or more of the tenant systems is a PIM router, and the Tenant Domain is used for as a transit network that is part of a larger multicast domain. This is discussed in Section 7.

3. EVPN-aware Multicast Solution Control Plane

3.1. Supplementary Broadcast Domain (SBD) and Route Targets

Every Tenant Domain is associated with a single Supplementary Broadcast Domain (SBD), as discussed in Section 2.1. Recall that a Tenant Domain is defined to be a set of BDs that can freely send and receive IP multicast traffic to/from each other. If an EVPN-PE has one or more ACs in a BD of a particular Tenant Domain, and if the EVPN-PE supports the procedures of this document, that EVPN-PE must be provisioned with the SBD of that Tenant Domain.

At each EVPN-PE attached to a given Tenant Domain, there is an IRB interface leading from the L3 routing instance of that Tenant Domain and the SBD. However, the SBD has no ACs.

The SBD may be in an EVPN Instance (EVI) of its own, or it may be one of several BDs (of the same Tenant Domain) in an EVI.

Each SBD is provisioned with a Route Target (RT). All the EVPN-PEs supporting a given SBD are provisioned with that RT as an import RT.

Each SBD is also provisioned with a "Tag ID" (see Section 6 of [RFC7432]).

- If the SBD is the only BD in its EVI, the mapping from RT to SBD is one-to-one. The Tag ID is zero.

o If the SBD is one of several BDs in its EVI, it may have its own RT, or it may share an RT with one or more of those other BDs. In either case, it must be assigned a non-zero Tag ID. The mapping from <RT, Tag ID> is always one-to-one.

We will use the term "SBD-RT" to denote the RT has has been assigned to an SBD. Routes carrying this RT will be propagated to all EVPN-PEs in the same Tenant Domain as the originator.

An EVPN-PE that receives a route can always determine whether a received route "belongs to" a particular SBD, by seeing if that route carries the SBD-RT and has the Tag ID of the SBD in its NLRI.

If the VLAN-based service model is being used for a particular Tenant Domain, and thus each BD is in a distinct EVI, it is natural to have the SBD be in a distinct EVI as well. If the VLAN-aware bundle service is being used, it is natural to include the SBD in the same EVI that contains the other BDs. However, it is not required to do so; the SBD can still be placed in an EVI of its own, if that is desired.

Note that an SBD, just like any other BD, is associated on each EVPN-PE with a MAC-VRF. Per [RFC7432], each MAC-VRF is associated with a Route Distinguisher (RD). When constructing a route that is "about" an SBD, an EVPN-PE will place the RD of the associated MAC-VRF in the "Route Distinguisher" field of the NLRI. (If the Tenant Domain has several MAC-VRFs on a given PE, the EVPN-PE has a choice of which RD to use.)

If Assisted Replication (AR, see [EVPN-AR]) is used, each AR-REPLICATOR for a given Tenant Domain must be provisioned with the SBD of that Tenant Domain, even if the AR-REPLICATOR does not have any L3 routing instance.

3.2. Advertising the Tunnels Used for IP Multicast

The procedures used for advertising the tunnels that carry IP multicast traffic depend upon the type of tunnel being used. If the tunnel type is neither Ingress Replication, Assisted Replication, nor BIER, there are procedures for advertising both "inclusive tunnels" and "selective tunnels".

When IR, AR or BIER are used to transmit IP multicast packets across the core, there are no P2MP tunnels. Once an ingress EVPN-PE determines the set of egress EVPN-PEs for a given flow, the IMET routes contain all the information needed to transport packets of that flow to the egress PEs.

If AR is used, the ingress EVPN-PE is also an AR-LEAF and the IMET route coming from the selected AR-REPLICATOR contains the information needed. The AR-REPLICATOR will behave as an ingress EVPN-PE when sending a flow to the egress EVPN-PEs.

If the tunneling technique requires P2MP tunnels to be set up (e.g., RSVP-TE P2MP, mLDP, PIM), some of the tunnels may be selective tunnels and some may be inclusive tunnels.

Selective tunnels are always advertised by the ingress PE using S-PMSI A-D routes ([EVPN-BUM]).

For inclusive tunnels, there is a choice between using a BD's ordinary "BUM tunnel" [RFC7432] as the default inclusive tunnel for carrying IP multicast traffic, or using a separate IP multicast tunnel as the default inclusive tunnel for carrying IP multicast. In the former case, the inclusive tunnel is advertised in an IMET route. In the latter case, the inclusive tunnel is advertised in a (C-*,C-*) S-PMSI A-D route ([EVPN-BUM]). Details may be found in subsequent sections.

3.2.1. Constructing SBD Routes

3.2.1.1. Constructing an SBD-IMET Route

In general, an EVPN-PE originates an IMET route for each real BD. Whether an EVPN-PE has to originate an IMET route for the SBD (of a particular Tenant Domain) depends upon the type of tunnels being used to carry EVPN multicast traffic across the backbone. In some cases, an IMET route does not need to be originated for the SBD, but the other IMET routes have to carry the SBD-RT as well as any other RTs they would ordinarily carry (per [RFC7432].

Subsequent sections will specify when it is necessary for an EVPN-PE to originate an IMET route for the SBD. We will refer to such a route as an "SBD-IMET route".

When an EVPN-PE needs to originate an SBD-IMET route that is "for" the SBD, it constructs the route as follows:

o the RD field of the route's NLRI is set to the RD of the MAC-VRF that is associated with the SBD;

o a Route Target Extended Community containing the value of the SBD-RT is attached to that route;

o the "Tag ID" field of the NLRI is set to the Tag ID that has been assigned to the SBD. This is most likely 0 if a VLAN-based or

VLAN-bundle service is being used and non-zero if a VLAN-aware bundle service is being used.

3.2.1.2. Constructing an SBD-SMET Route

An EVPN-PE can originate an SMET route to indicate that it has receivers, on a specified BD, for a specified multicast flow. In some scenarios, an EVPN-PE must originate an SMET route that is for the SBD, which we will call an "SBD-SMET route". Whether an EVPN-PE has to originate an SMET route for the SBD (of a particular tenant domain) depends upon various factors, detailed in subsequent sections.

When an EVPN-PE needs to originate an SBD-SMET route that is "for" the SBD, it constructs the route as follows:

o the RD field of the route's NLRI is set to the RD of the MAC-VRF that is associated with the SBD;

o a Route Target Extended Community containing the value of the SBD-RT is attached to that route;

o the "Tag ID" field of the NLRI is set to the Tag ID that has been assigned to the SBD. This is most likely 0 if a VLAN-based or VLAN-bundle service is being used and non-zero if a VLAN-aware bundle service is being used.

3.2.1.3. Constructing an SBD-SPMSI Route

An EVPN-PE can originate an S-PMSI A-D route (see [EVPN-BUM]) to indicate that it is going to use a particular P2MP tunnel to carry the traffic of particular IP multicast flows. In general, an S-PMSI A-D route is specific to a particular BD. In some scenarios, an EVPN-PE must originate an S-PMSI A-D route that is for the SBD, which we will call an "SBD-SPMSI route". Whether an EVPN-PE has to originate an SBD-SPMSI route for (of a particular Tenant Domain) depends upon various factors, detailed in subsequent sections.

When an EVPN-PE needs to originate an SBD-SPMSI route that is "for" the SBD, it constructs the route as follows:

o the RD field of the route's NLRI is set to the RD of the MAC-VRF that is associated with the SBD;

o a Route Target Extended Community containing the value of the SBD-RT is attached to that route;

o the "Tag ID" field of the NLRI is set to the Tag ID that has been assigned to the SBD. This is most likely 0 if a VLAN-based or VLAN-bundle service is being used and non-zero if a VLAN-aware bundle service is being used.

3.2.2. Ingress Replication

When Ingress Replication (IR) is used to transport IP multicast frames of a given Tenant Domain, each EVPN-PE attached to that Tenant Domain MUST originate an SBD-IMET route, as described in Section 3.2.1.1.

The SBD-IMET route MUST carry a PMSI Tunnel attribute (PTA), and the MPLS label field of the PTA MUST specify a downstream-assigned MPLS label that maps uniquely (in the context of the originating EVPN-PE) to the SBD.

An EVPN-PE MUST also originate an IMET route for each BD to which it is attached, following the procedures of [RFC7432]. Each of these IMET routes carries a PTA that specifying a downstream-assigned label that maps uniquely (in the context of the originating EVPN-PE) to the BD in question. These IMET routes need not carry the SBD-RT.

When an ingress EVPN-PE needs to use IR to send an IP multicast frame from a particular source BD to an egress EVPN-PE, the ingress PE determines whether the egress PE has originated an IMET route for that BD. If so, that IMET route contains the MPLS label that the egress PE has assigned to the source BD. The ingress PE uses that label when transmitting the packet to the egress PE. Otherwise, the ingress PE uses the label that the egress PE has assigned to the SBD (in the SBD-IMET route originated by the egress).

Note that the set of IMET routes originated by a given egress PE, and installed by a given ingress PE, will change over time. If the egress PE withdraws its IMET route for the source BD, the ingress PE must stop using the label carried in that IMET route, and start using the label carried in the SBD-IMET route from that egress PE.

3.2.3. Assisted Replication

When Assisted Replication is used to transport IP multicast frames of a given Tenant Domain, each EVPN-PE (including the AR-REPLICATOR) attached to the Tenant Domain MUST originate an SBD-IMET route, as described in Section 3.2.1.1.

An AR-REPLICATOR attached to a given Tenant Domain is considered to be an EVPN-PE of that Tenant Domain. It is attached to all the BDs in the Tenant Domain, but it has no IRB interfaces.

As with Ingress Replication, the SBD-IMET route carries a PTA where the MPLS label field specifies the downstream-assigned MPLS label that identifies the SBD. However, the AR-REPLICATOR and AR-LEAF EVPN-PEs will set the PTA's flags differently, as per [EVPN-AR].

In addition, each EVPN-PE originates an IMET route for each BD to which it is attached. As in the case of Ingress Replication, these routes carry the downstream-assigned MPLS labels that identify the BDs and do not carry the SBD-RT.

When an ingress EVPN-PE, acting as AR-LEAF, needs to send an IP multicast frame from a particular source BD to an egress EVPN-PE, the ingress PE determines whether there is any AR-REPLICATOR that originated an IMET route for that BD. After the AR-REPLICATOR selection (if there are more than one), the AR-LEAF uses the label contained in the IMET route of the AR-REPLICATOR when transmitting packets to it. The AR-REPLICATOR receives the packet and, based on the procedures specified in [EVPN-AR], transmits the packets to the egress EVPN-PEs using the labels contained in the IMET routes received from the egress PEs.

If an ingress AR-LEAF for a given BD has not received any IMET route for that BD from an AR-REPLICATOR, the ingress AR-LEAF follows the procedures in Section 3.2.2.

3.2.4. BIER

When BIER is used to transport multicast packets of a given Tenant Domain, each EVPN-PE attached to that Tenant Domain MUST originate an SBD-IMET route, as described in Section 3.2.1.1.

In addition, IMET routes that are originated for other BDs in the Tenant Domain MUST carry the SBD-RT.

Each IMET route (including but not limited to the SBD-IMET route) MUST carry a PMSI Tunnel attribute (PTA). The MPLS label field of the PTA MUST specify an upstream-assigned MPLS label that maps uniquely (in the context of the originating EVPN-PE) to the BD for which the route is originated.

When an ingress EVPN-PE uses BIER to send an IP multicast packet (inside an ethernet frame) from a particular source BD to a set of egress EVPN-PEs, the ingress PE follows the BIER encapsulation with the upstream-assigned label it has assigned to the source BD. (This label will come from the originated SBD-IMET route ONLY if the traffic originated from outside the Tenant Domain.) An egress PE can determine from that label whether the packet's source BD is one of the BDs to which the egress PE is attached.

Further details on the use of BIER to support EVPN can be found in [EVPN-BIER].

3.2.5. Inclusive P2MP Tunnels

3.2.5.1. Using the BUM Tunnels as IP Multicast Inclusive Tunnels

The procedures in this section apply only when it is desired to use the BUM tunnels to carry IP multicast traffic across the backbone. In this cases, an IP multicast frame (whether inter-subnet or intra-subnet) will be carried across the backbone in the BUM tunnel belonging to its source BD. An EVPN-PE attached to a given Tenant Domain will then need to join the BUM tunnels for each BD in the Tenant Domain, even if the EVPN-PE is not attached to all of those BDs. The reason is that an IP multicast packet from any source BD might be needed by an EVPN-PE that is not attached to that source domain.

Note that this will cause BUM traffic from a given BD in a Tenant Domain to be sent to all PEs that attach to that tenant domain, even the PEs that don't attach to the given BD. To avoid this, it is RECOMMENDED that the BUM tunnels not be used as IP Multicast inclusive tunnels, and that the procedures of Section 3.2.5.2 be used instead.

3.2.5.1.1. RSVP-TE P2MP

When BUM tunnels created by RSVP-TE P2MP are used to transport IP multicast frames of a given Tenant Domain, each EVPN-PE attached to that Tenant Domain MUST originate an SBD-IMET route, as described in Section 3.2.1.1.

In addition, IMET routes that are originated for other BDs in the Tenant Domain MUST carry the SBD-RT.

Each IMET route (including but not limited to the SBD-IMET route) MUST carry a PMSI Tunnel attribute (PTA).

If received IMET route is not the SBD-IMET route, it will also be carrying the RT for its source BD. The route's NLRI will carry the Tag ID for the source BD. From the RT and the Tag ID, any PE receiving the route can determine the route's source BD.

If the MPLS label field of the PTA contains zero, the specified RSVP-TE P2MP tunnel is used only to carry frames of a single source BD.

If the MPLS label field of the PTA does not contain zero, it MUST contain an upstream-assigned MPLS label that maps uniquely (in the context of the originating EVPN-PE) to the source BD (or, in the case of an SBD-IMET route, the SBD). The tunnel may be used to carry frames of multiple source BDs, and the source BD for a particular packet is inferred from the label carried by the packet.

IP multicast traffic originating outside the Tenant Domain is transmitted with the label corresponding to the SBD, as specified in the ingress EVPN-PE's SBD-IMET route.

3.2.5.1.2. mLDP or PIM

When either mLDP or PIM is used to transport multicast packets of a given Tenant Domain, an EVPN-PE attached to that tenant domain originates an SBD-IMET route only if it is the ingress PE for IP multicast traffic originating outside the tenant domain. Such traffic is treated as having the SBD as its source BD.

An EVPN-PE MUST originate an IMET routes for each BD to which it is attached. These IMET routes MUST carry the SBD-RT of the Tenant Domain to which the BD belongs. Each such IMET route must also carry the RT of the BD to which it belongs.

When an IMET route (other than the SBD-IMET route) is received by an egress PE, the route will be carrying the RT for its source BD and the route's NLRI will contain the Tag ID for that source BD. This allows any PE receiving the route to determine the source BD associated with the route.

If the MPLS label field of the PTA contains zero, the specified mLDP or PIM tunnel is used only to carry frames of a single source BD.

If the MPLS label field of the PTA does not contain zero, it MUST contain an upstream-assigned MPLS label that maps uniquely (in the context of the originating EVPN-PE) to the source BD. The tunnel may be used to carry frames of multiple source BDs, and the source BD for a particular packet is inferred from the label carried by the packet.

The EVPN-PE advertising these IMET routes is specifying the default tunnel that it will use (as ingress PE) for transmitting IP multicast packets. The upstream-assigned label allows an egress PE to determine the source BD of a given packet.

The procedures of this section apply whenever the tunnel technology is based on the construction of the multicast trees in a "receiver-driven" manner; mLDP and PIM are two ways of constructing trees in a receiver-driven manner.

3.2.5.2. Using Wildcard S-PMSI A-D Routes to Advertise Inclusive Tunnels Specific to IP Multicast

The procedures of this section apply when (and only when) it is desired to transmit IP multicast traffic on an inclusive tunnel, but not on the same tunnel used to transmit BUM traffic.

However, these procedures do NOT apply when the tunnel type is Ingress Replication or BIER, EXCEPT in the case where it is necessary to interwork between non-OISM PEs and OISM PEs, as specified in Section 5.

Each EVPN-PE attached to the given Tenant Domain MUST originate an SBD-SPMSI A-D route. The NLRI of that route MUST contain (C-*,C-*) (see [RFC6625]). Additional rules for constructing that route are given in Section 3.2.1.3.

In addition, an EVPN-PE MUST originate an S-PMSI A-D route containing (C-*,C-*) in its NLRI for each of the other BDs in the Tenant Domain to which it is attached. All such routes MUST carry the SBD-RT. This ensures that those routes are imported by all EVPN-PEs attached to the Tenant Domain.

The route carrying the PTA will also be carrying the RT for that source BD, and the route's NLRI will contain the Tag ID for that source BD. This allows any PE receiving the route to determine the source BD associated with the route.

If the MPLS label field of the PTA contains zero, the specified tunnel is used only to carry frames of a single source BD.

If the MPLS label field of the PTA does not contain zero, it MUST specify an upstream-assigned MPLS label that maps uniquely (in the context of the originating EVPN-PE) to the source BD. The tunnel may be used to carry frames of multiple source BDs, and the source BD for a particular packet is inferred from the label carried by the packet.

The EVPN-PE advertising these S-PMSI A-D route routes is specifying the default tunnel that it will use (as ingress PE) for transmitting IP multicast packets. The upstream-assigned label allows an egress PE to determine the source BD of a given packet.

3.2.6. Selective Tunnels

An ingress EVPN-PE for a given multicast flow or set of flows can always assign the flow to a particular P2MP tunnel by originating an S-PMSI A-D route whose NLRI identifies the flow or set of flows. The NLRI of the route could be (C-*,C-G), or (C-S,C-G). The S-PMSI A-D

route MUST carry the SBD-RT, so that it is imported by all EVPN-PEs attached to the Tenant Domain.

An S-PMSI A-D route is "for" a particular source BD. It MUST carry the RT associated with that BD, and it MUST have the Tag ID for that BD in its NLRI.

Each such route MUST contain a PTA, as specified in Section 3.2.5.2.

An egress EVPN-PE interested in the specified flow or flows MUST join the specified tunnel. Procedures for joining the specified tunnel are specific to the tunnel type. (Note that if the tunnel type is RSVP-TE P2MP LSP, the Leaf Information Required (LIR) flag of the PTA SHOULD NOT be set. An ingress OISM PE knows which OISM EVPN PEs are interested in any given flow, and hence can add them to the RSVP-TE P2MP tunnel that carries such flows.)

When an EVPN-PE imports an S-PMSI A-D route, it infers the source BD from the RTs and the Tag ID. If the EVPN-PE is not attached to the source BD, the tunnel it specifies is treated as belonging to the SBD. That is, packets arriving on that tunnel are treated as having been sourced in the SBD. Note that a packet is only considered to have arrived on the specified tunnel if the packet carries the upstream-assigned label specified in in the PTA, or if there is no upstream-assigned label specified in the PTA.

It should be noted that when either IR or BIER is used, there is no need for an ingress PE to use S-PMSI A-D routes to assign specific flows to selective tunnels. The procedures of Section 3.3, along with the procedures of Section 3.2.2, Section 3.2.3, or Section 3.2.4, provide the functionality of selective tunnels without the need to use S-PMSI A-D routes.

3.3. Advertising SMET Routes

[IGMP-Proxy] allows an egress EVPN-PE to express its interest in a particular multicast flow or set of flows by originating an SMET route. The NLRI of the SMET route identifies the flow or set of flows as (C-*,C-*) or (C-*,C-G) or (C-S,C-G).

Each SMET route belongs to a particular BD. The Tag ID for the BD appears in the NLRI of the route, and the route carries the RT associated that that BD. From this <RT, tag> pair, other EVPN-PEs can identify the BD to which a received SMET route belongs. (Remember though that the route may be carrying multiple RTs.)

There are two cases to consider:

1. Case 1: When it is known that no BD of a Tenant Domain contains a multicast router.

   In this case, an egress PE can advertise its interest in a flow or set of flows by originating a single SMET route. The SMET route will belong to the SBD. We refer to this as an SBD-SMET route. The SBD-SMET route carries the SBD-RT, and has the Tag ID for the SBD in its NLRI. SMET routes for the individual BDs are not needed.

2. Case 2: When it is possible that a BD of a Tenant Domain contains a multicast router.

   Suppose that an egress PE is attached to a BD on which there might be a tenant multicast router. (The tenant router is not necessarily on a segment that is attached to that PE.) And suppose that the PE has one or more ACs attached to that BD which are interested in a given multicast flow. In this case, IN ADDITION to the SMET route for the SBD, the egress PE MUST originate an SMET route for that BD. This will enable the ingress PE(s) to send IGMP/MLD messages on ACs for the BD, as specified in [IGMP-Proxy].

   If an SMET route is not an SBD-SMET route, and if the SMET route is for (C-S,C-G) (i.e., no wildcard source), and if the EVPN-PE originating it knows the source BD of C-S, it MAY put only the RT for that BD on the route. Otherwise, the route MUST carry the SBD-RT, so that it gets distributed to all the EVPN-PEs attached to the tenant domain.

As detailed in [IGMP-Proxy], an SMET route carries flags saying whether it is to result in the propagation of IGMP v1, v2, or v3 messages on the ACs of the BD to which the SMET route belongs. These flags SHOULD be set to zero in an SBD-SMET route.

Note that a PE only needs to originate the set SBD-SMET routes that are needed to pull in all the traffic in which it is interested. Suppose PE1 has ACs attached to BD1 that are interested in (C-*,C-G) traffic, and ACs attached to BD2 that are interested in (C-S,C-G) traffic. A single SBD-SMET route specifying (C-*,C-G) will pull in all the necessary flows.

As another example, suppose the ACs attached to BD1 are interested in (C-*,C-G) but not in (C-S,C-G), while the ACs attached to BD2 are interested in (C-S,C-G). A single SBD-SMET route specifying (C-*,C-G) will pull in all the necessary flows.

In other words, to determine the set of SBD-SMET routes that have to be sent for a given C-G, the PE has to merge the IGMP/MLD state for all the BDs (of the given Tenant Domain) to which it is attached.

Per [IGMP-Proxy], importing an SMET route for a particular BD will cause IGMP/MLD state to be instantiated for the IRB interface to that BD. This applies as well when the BD is the SBD.

However, traffic originating in a BD of a particular Tenant Domain MUST NOT be sent down the IRB interface that connects the L3 routing instance of that Tenant Domain to the SBD of that Tenant Domain. That would cause duplicate delivery of traffic, since traffic arriving at L3 over the IRB interface from the SBD has already been distributed throughout the Tenant Domain. When setting up the IGMP/MLD state based on SBD-SMET routes, care must be taken to ensure that the IRB interface to the SBD is not added to the Outgoing Interface (OIF) list if the traffic originates within the Tenant Domain.

4. Constructing Multicast Forwarding State

4.1. Layer 2 Multicast State

An EVPN-PE maintains "layer 2 multicast state" for each BD to which it is attached.

Let PE1 be an EVPN-PE, and BD1 be a BD to which it is attached. At PE1, BD1's layer 2 multicast state for a given (C-S,C-G) or (C-*,C-G) governs the disposition of an IP multicast packet that is received by BD1's layer 2 multicast function on an EVPN-PE.

An IP multicast (S,G) packet is considered to have been received by BD1's layer 2 multicast function in PE1 in the following cases:

o The packet is the payload of an ethernet frame received by PE1 from an AC that attaches to BD1.

o The packet is the payload of an ethernet frame whose source BD is BD1, and which is received by the PE1 over a tunnel from another EVPN-PE.

o The packet is received from BD1's IRB interface (i.e., has been transmitted by PE1's L3 routing instance down BD1's IRB interface).

According to the procedures of this document, all transmission of IP multicast packets from one EVPN-PE to another is done at layer 2. That is, the packets are transmitted as ethernet frames, according to the layer 2 multicast state.

Each layer 2 multicast state (S,G) or (*,G) contains a set "output interfaces" (OIF list). The disposition of an (S,G) multicast frame received by BD1's layer 2 multicast function is determined as follows:

- o The OIF list is taken from BD1's layer 2 (S,G) state, or if there is no such (S,G) state, then from BD1's (*,G) state. (If neither state exists, the OIF list is considered to be null.)

- o The rules of Section 4.1.2 are applied to the OIF list. This will generally result in the frame being transmitted to some, but not all, elements of the OIF list.

Note that there is no RPF check at layer 2.

4.1.1. Constructing the OIF List

In this document, we have extended the procedures of [IGMP-Proxy] so that IMET and SMET routes for a particular BD are distributed not just to PEs that attach to that BD, but to PEs that attach to any BD in the Tenant Domain. In this way, each PE attached to a given Tenant Domain learns, from each other PE attached to the same Tenant Domain, the set of flows that are of interest to each of those other PEs. (If some PE attached to the Tenant Domain does not support [IGMP-Proxy], it will be assumed to be interested in all flows. Whether a particular remote PE supports [IGMP-Proxy] is determined by the presence of an Extended Community in its IMET route; this is specified in [IGMP-Proxy].) If a set of remote PEs are interested in a particular flow, the tunnels used to reach those PEs are added to the OIF list of the multicast states corresponding to that flow.

An EVPN-PE may run IGMP/MLD procedures on each of its ACs, in order to determine the set of flows of interest to each AC. (An AC is said to be interested in a given flow if it connects to a segment that has tenant systems interested in that flow.) If IGMP/MLD procedures are not being run on a given AC, that AC is considered to be interested in all flows. For each BD, the set of ACs interested in a given flow is determined, and the ACs of that set are added to the OIF list of that BD's multicast state for that flow.

The OIF list for each multicast state must also contain the IRB interface for the BD to which the state belongs.

Implementors should note that the OIF list of a multicast state will change from time to time as ACs and/or remote PEs either become interested in, or lose interest in, particular multicast flows.

4.1.2. Data Plane: Applying the OIF List to an (S,G) Frame

When an (S,G) multicast frame is received by the layer 2 multicast function of a given EVPN-PE, say PE1, its disposition depends (a) the way it was received, (b) upon the OIF list of the corresponding multicast state (see Section 4.1.1), (c) upon the "eligibility" of an AC to receive a given frame (see Section 4.1.2.1 and (d) upon its source BD (see Section 3.2 for information about determining the source BD of a frame received over a tunnel from another PE).

4.1.2.1. Eligibility of an AC to Receive a Frame

A given (S,G) multicast frame is eligible to be transmitted by a given PE, say PE1, on a given AC, say AC1, only if one of the following conditions holds:

1. ESI labels are being used, PE1 is the DF for the segment to which AC1 is connected, and the frame did not originate from that same segment (as determined by the ESI label), or

2. The ingress PE for the frame is a remote PE, say PE2, local bias is being used, and PE2 is not connected to the same segment as AC1.

4.1.2.2. Applying the OIF List

Assume a given (S,G) multicast frame has been received by a given PE, say PE1. PE1 determines the source BD of the frame, finds the layer 2 (S,G) state for the source BD (or the (*,G) state if there is no (S,G) state), and takes the OIF list from that state. Note that if PE1 is not attached to the actual source BD, it will treat the frame as if its source BD is the SBD.

Suppose PE1 has determined the frame's source BD to be BD1 (which may or may not be the SBD.) There are the following cases to consider:

1. The frame was received by PE1 from a local AC, say AC1, that attaches to BD1.

    a. The frame MUST be sent out all local ACs of BD1 that appear in the OIF list, except for AC1 itself.

    b. The frame MUST also be delivered to any other EVPN-PEs that have interest in it. This is achieved as follows:

        i. If (a) AR is being used, and (b) PE1 is an AR-LEAF, and (c) the OIF list is non-null, PE1 MUST send the frame to the AR-REPLICATOR.

ii. Otherwise the frame MUST be sent on all tunnels in the OIF list.

c. The frame MUST be sent to the local L3 routing instance by being sent up the IRB interface of BD1. It MUST NOT be sent up any other IRB interfaces.

2. The frame was received by PE1 over a tunnel from another PE. (See Section 3.2 for the rules to determine the source BD of a packet received from another PE. Note that if PE1 is not attached to the source BD, it will regard the SBD as the source BD.)

   a. The frame MUST be sent out all local ACs in the OIF list that connect to BD1 and that are eligible (per Section 4.1.2.1) to receive the frame.

   b. The frame MUST be sent up the IRB interface of the source BD. (Note that this may be the SBD.) The frame MUST NOT be sent up any other IRB interfaces.

   c. If PE1 is not an AR-REPLICATOR, it MUST NOT send the frame to any other EVPN-PEs. However, if PE1 is an AR-REPLICATOR, it MUST send the frame to all tunnels in the OIF list, except for the tunnel over which the frame was received.

3. The frame was received by PE1 from the BD1 IRB interface (i.e., the frame has been transmitted by PE1's L3 routing instance down the BD1 IRB interface), and BD1 is NOT the SBD.

   a. The frame MUST be sent out all local ACs in the OIF list that are eligible (per Section 4.1.2.1 to receive the frame.

   b. The frame MUST NOT be sent to any other EVPN-PEs.

   c. The frame MUST NOT be sent up any IRB interfaces.

4. The frame was received from the SBD IRB interface (i.e., has been transmitted by PE1's L3 routing instance down the SBD IRB interface).

   a. The frame MUST be sent on all tunnels in the OIF list. This causes the frame to be delivered to any other EVPN-PEs that have interest in it.

   b. The frame MUST NOT be sent on any local ACs.

   c. The frame MUST NOT be sent up any IRB interfaces.

4.2. Layer 3 Forwarding State

If an EVPN-PE is performing IGMP/MLD procedures on the ACs of a given BD, it processes those messages at layer 2 to help form the layer 2 multicast state. If also sends those messages up that BD's IRB interface to the L3 routing instance of a particular tenant domain. This causes layer 2 (C-S,C-G) or (C-*,C-G) L3 state to be created/ updated.

A layer 3 multicast state has both an Input Interface (IIF) and an OIF list.

To set the IIF of an (C-S,C-G) state, the EVPN-PE must determine the source BD of C-S. This is done by looking up S in the local MAC-VRF(s) of the given Tenant Domain.

If the source BD is present on the PE, the IIF is set to the IRB interface that attaches to that BD. Otherwise the IIF is set to the SBD IRB interface.

For (C-*,C-G) states, traffic can arrive from any BD, so the IIF needs to be set to a wildcard value meaning "any IRB interface".

The OIF list of these states includes one or more of the IRB interfaces of the Tenant Domain. In general, maintenance of the OIF list does not require any EVPN-specific procedures. However, there is one EVPN-specific rule:

> If the IIF is one of the IRB interfaces (or the wild card meaning "any IRB interface"), then the SBD IRB interface MUST NOT be added to the OIF list. Traffic originating from within a particular EVPN Tenant Domain must not be sent down the SBD IRB interface, as such traffic has already been distributed to all EVPN-PEs attached to that Tenant Domain.

Please also see Section 6.1.1, which states a modification of this rule for the case where OISM is interworking with external Layer 3 multicast routing.

5. Interworking with non-OISM EVPN-PEs

It is possible that a given Tenant Domain will be attached to both OISM PEs and non-OISM PEs. Inter-subnet IP multicast should be possible and fully functional even if not all PEs attaching to a Tenant Domain can be upgraded to support OISM functionality.

Note that the non-OISM PEs are not required to have IRB support, or support for [IGMP-Proxy]. It is however advantageous for the non-OISM PEs to support [IGMP-Proxy].

In this section, we will use the following terminology:

o PE-S: the ingress PE for an (S,G) flow.

o PE-R: an egress PE for an (S,G) flow.

o BD-S: the source BD for an (S,G) flow. PE-S must have one or more ACs attached BD-S, at least one of which attaches to host S.

o BD-R: a BD that contains a host interested in the flow. The host is attached to PE-R via an AC that belongs to BD-R.

To allow OISM PEs to interwork with non-OISM PEs, a given Tenant Domain needs to contain one or more "IP Multicast Gateways" (IPMGs). An IPMG is an OISM PE with special responsibilities regarding the interworking between OISM and non-OISM PEs.

If a PE is functioning as an IPMG, it MUST signal this fact by attaching a particular flag or EC (details to be determined) to its IMET routes. An IPMG SHOULD attach this flag or EC to all IMET routes it originates. However, if PE1 imports any IMET route from PE2 that has the "IPMG" flag or EC present, then the PE1 will assume that PE2 is an IPMG.

An IPMG Designated Forwarder (IPMG-DF) selection procedure is used to ensure that, at any given time, there is exactly one active IPMG-DF for any given BD. Details of the IPMG-DF selection procedure are in Section 5.1. The IPMG-DF for a given BD, say BD-S, has special functions to perform when it receives (S,G) frames on that BD:

o If the frames are from a non-OISM PE-S:

   * The IPMG-DF forwards them to OISM PEs that do not attach to BD-S but have interest in (S,G).

     Note that OISM PEs that do attach to BD-S will have received the frames on the BUM tunnel from the non-OISM PE-S.

   * The IPMG-DF forwards them to non-OISM PEs that have interest in (S,G) on ACs that do not belong to BD-S.

     Note that if a non-OISM PE has multiple BDs other than BD-S with interest in (S,G), it will receive one copy of the frame

for each such BD. This is necessary because the non-OISM PEs cannot move IP multicast traffic from one BD to another.

o If the frames are from an OISM PE, the IPMG-DF forwards them to non-OISM PEs that have interest in (S,G) on ACs that do not belong to BD-S.

If a non-OISM PE has interest in (S,G) on an AC belonging to BD-S, it will have received a copy of the (S,G) frame, encapsulated for BD-S, from the OISM PE-S. (See Section 3.2.2.) If the non-OISM PE has interest in (S,G) on one or more ACs belonging to BD-R1,...,BD-Rk where the BD-Ri are distinct from BD-S, the IPMG-DF needs to send it a copy of the frame for BD-Ri.

If an IPMG receives a frame on a BD for which it is not the IPMG-DF, it just follows normal OISM procedures.

This section specifies several sets of procedures:

o the procedures that the IPMG-DF for a given BD needs to follow when receiving, on that BD, an IP multicast frame from a non-OISM PE;

o the procedures that the IPMG-DF for a given BD needs to follow when receiving, on that BD, an IP multicast frame from an OISM PE;

o the procedures that an OISM PE needs to follow when receiving, on a given BD, an IP multicast frame from a non-OISM PE, when the OISM PE is not the IPMG-DF for that BD.

To enable OISM/non-OISM interworking in a given Tenant Domain, the Tenant Domain MUST have some EVPN-PEs that can function as IPMGs. An IPMG must be configured with the SBD. It must also be configured with every BD of the Tenant Domain that exists on any of the non-OISM PEs of that domain. (Operationally, it may be simpler to configure the IPMG with all the BDs of the Tenant Domain.)

A non-OISM PE of course only needs to be configured with BDs for which it has ACs. An OISM PE that is not an IPMG only needs to be configured with the SBD and with the BDs for which it has ACs.

An IPMG MUST originate a wildcard SMET route (with (C-*,C-*) in the NLRI) for each BD in the Tenant Domain. This will cause it to receive all the IP multicast traffic that is sourced in the Tenant Domain. Note that non-OISM nodes that do not support [IGMP-Proxy] will send all the multicast traffic from a given BD to all PEs attached to that BD, even if those PEs do not originate an SMET route.

The interworking procedures vary somewhat depending upon whether packets are transmitted from PE to PE via Ingress Replication (IR) or via Point-to-Multipoint (P2MP) tunnels. We do not consider the use of BIER in this section, due to the low likelihood of there being a non-OISM PE that supports BIER.

5.1. IPMG Designated Forwarder

Each IPMG MUST be configured with an "IPMG dummy ethernet segment" that has no ACs.

EVPN supports a number of procedures that can be used to select the Designated Forwarder (DF) for a particular BD on a particular ethernet segment. Some of the possible procedures can be found, e.g., in [RFC7432], [EVPN-DF-NEW], and [EVPN-DF-WEIGHTED]. Whatever procedure is in use in a given deployment can be adapted to select an IPMG-DF for a given BD, as follows.

Each IPMG will originate an Ethernet Segment route for the IPMG dummy ethernet segment. It MUST carry a Route Target derived from the corresponding Ethernet Segment Identifier. Thus only IPMGs will import the route.

Once the set of IPMGs is known, it is also possible to determine the set of BDs supported by each IPMG. The DF selection procedure can then be used to choose a DF for each BD. (The conditions under which the IPMG-DF for a given BD changes depends upon the DF selection algorithm that is in use.)

5.2. Ingress Replication

The procedures of this section are used when Ingress Replication is used to transmit packets from one PE to another.

When a non-OISM PE-S transmits a multicast frame from BD-S to another PE, PE-R, PE-S will use the encapsulation specified in the BD-S IMET route that was originated by PE-R. This encapsulation will include the label that appears in the "MPLS label" field of the PMSI Tunnel attribute (PTA) of the IMET route. If the tunnel type is VXLAN, the "label" is actually a Virtual Network Identifier (VNI); for other tunnel types, the label is an MPLS label. In either case, we will speak of the transmitted frames as carrying a label that was assigned to a particular BD by the PE-R to which the frame is being transmitted.

To support OISM/non-OISM interworking, an OISM PE-R MUST originate, for each of its BDs, both an IMET route and an S-PMSI (C-*,C-*) A-D route. Note that even when IR is being used, interworking between

OISM and non-OISM PEs requires the OISM PEs to follow the rules of Section 3.2.5.2, as modified below.

Non-OISM PEs will not understand S-PMSI A-D routes. So when a non-OISM PE-S transmits an IP multicast frame with a particular source BD to an IPMG, it encapsulates the frame using the label specified in that IPMG's BD-S IMET route. (This is just the procedure of [RFC7432].)

The (C-*,C-*) S-PMSI A-D route originated by a given OISM PE will have a PTA that specifies IR.

- o If MPLS tunneling is being used, the MPLS label field SHOULD contain a non-zero value, and the LIR flag SHOULD be zero. (The case where the MPLS label field is zero or the LIR flag is set is outside the scope of this document.)

- o If the tunnel encapsulation is VXLAN, the MPLS label field MUST contain a non-zero value, and the LIR flag MUST be zero.

When an OISM PE-S transmits an IP multicast frame to an IPMG, it will use the label specified in that IPMG's (C-*,C-*) S-PMSI A-D route.

When a PE originates both an IMET route and a (C-*,C-*) S-PMSI A-D route, the values of the MPLS label field in the respective PTAs must be distinct. Further, each MUST map uniquely (in the context of the originating PE) to the route's BD.

As a result, an IPMG receiving an MPLS-encapsulated IP multicast frame can always tell by the label whether the frame's ingress PE is an OISM PE or a non-OISM PE. When an IPMG receives a VXLAN-encapsulated IP multicast frame it may need to determine the identity of the ingress PE from the outer IP encapsulation; it can then determine whether the ingress PE is an OISM PE or a non-OISM PE by looking the IMET route from that PE.

Suppose an IPMG receives an IP multicast frame from another EVPN-PE in the Tenant Domain, and the IPMG is not the IPMG-DF for the frame's source BD. Then the IPMG performs only the ordinary OISM functions; it does not perform the IPMG-specific functions for that frame. In the remainder of this section, when we discuss the procedures applied by an IPMG when it receives an IP multicast frame, we are presuming that the source BD of the frame is a BD for which the IPMG is the IPMG-DF.

We have two basic cases to consider: (1) a frame's ingress PE is a non-OISM node, and (2) a frame's ingress PE is an OISM node.

5.2.1. Ingress PE is non-OISM

In this case, a non-OISM PE, PE-S, has received an (S,G) multicast frame over an AC that is attached to a particular BD, BD-S. By virtue of normal EVPN procedures, PE-S has sent a copy of the frame to every PE-R (both OISM and non-OISM) in the Tenant Domain that is attached to BD-S. If the non-OISM node supports [IGMP-Proxy], only PEs that have expressed interest in (S,G) receive the frame. The IPMG will have expressed interest via a (C-*,C-*) SMET route and thus receives the frame.

Any OISM PE (including an IPMG) receiving the frame will apply normal OISM procedures. As a result it will deliver the frame to any of its local ACs (in BD-S or in any other BD) that have interest in (S,G).

An OISM PE that is also the IPMG-DF for a particular BD, say BD-S, has additional procedures that it applies to frames received on BD-S from non-OISM PEs:

1. When the IPMG-DF for BD-S receives an (S,G) frame from a non-OISM node, it MUST forward a copy of the frame to every OISM PE that is NOT attached to BD-S but has interest in (S,G). The copy sent to a given OISM PE-R must carry the label that PE-R has assigned to the SBD in an S-PMSI A-D route. The IPMG MUST NOT do any IP processing of the frame's IP payload. TTL decrement and other IP processing will be done by PE-R, per the normal OISM procedures. There is no need for the IPMG to include an ESI label in the frame's tunnel encapsulation, because it is already known that the frame's source BD has no presence on PE-R. There is also no need for the IPMG to modify the frame's MAC SA.

2. In addition, when the IPMG-DF for BD-S receives an (S,G) frame from a non-OISM node, it may need to forward copies of the frame to other non-OISM nodes. Before it does so, it MUST decapsulate the (S,G) packet, and do the IP processing (e.g., TTL decrement). Suppose PE-R is a non-OISM node that has an AC to BD-R, where BD-R is not the same as BD-S, and that AC has interest in (S,G). The IPMG must then encapsulate the (S,G) packet (after the IP processing has been done) in an ethernet header. The MAC SA field will have the MAC address of the IPMG's IRB interface to BD-R. The IPMG then sends the frame to PE-R. The tunnel encapsulation will carry the label that PE-R advertised in its IMET route for BD-R. There is no need to include an ESI label, as the source and destination BDs are known to be different.

Note that if a non-OISM PE-R has several BDs (other than BD-S) with local ACs that have interest in (S,G), the IPMG will send it one copy for each such BD. This is necessary because the non-OISM PE cannot move packets from one BD to another.

There may be deployment scenarios in which every OISM PE is configured with every BD that is present on any non-OISM PE. In such scenarios, the procedures of item 1 above will not actually result in the transmission of any packets. Hence if it is known a priori that this deployment scenario exists for a given tenant domain, the procedures of item 1 above can be disabled.

5.2.2. Ingress PE is OISM

In this case, an OISM PE, PE-S, has received an (S,G) multicast frame over an AC that attaches to a particular BD, BD-S.

By virtue of receiving all the IMET routes about BD-S, PE-S will know all the PEs attached to BD-S. By virtue of normal OISM procedures:

o PE-S will send a copy of the frame to every OISM PE-R (including the IPMG) in the Tenant Domain that is attached to BD-S and has interest in (S,G). The copy sent to a given PE-R carries the label that that the PE-R has assigned to BD-S in its (C-*,C-*) S-PMSI A-D route.

o PE-S will also transmit a copy of the (S,G) frame to every OISM PE-R that has interest in (S,G) but is not attached to BD-S. The copy will contain the label that the PE-R has assigned to the SBD. (As in Section 5.2.1, an IPMG is assumed to have indicated interest in all multicast flows.)

o PE-S will also transmit a copy of the (S,G) frame to every non-OISM PE-R that is attached to BD-S. It does this using the label advertised by that PE-R in its IMET route for BD-S.

The PE-Rs follow their normal procedures. An OISM PE that receives the (S,G) frame on BD-S applies the OISM procedures to deliver the frame to its local ACs, as necessary. A non-OISM PE that receives the (S,G) frame on BD-S delivers the frame only to its local BD-S ACs, as necessary.

Suppose that a non-OISM PE-R has interest in (S,G) on a BD, BD-R, that is different than BD-S. If the non-OISM PE-R is attached to BD-S, the OISM PE-S will send forward it the original (S,G) multicast frame, but the non-OISM PE-R will not be able to send the frame to ACs that are not in BD-S. If PE-R is not even attached to BD-S, the OISM PE-S will not send it a copy of the frame at all, because PE-R

is not attached to the SBD. In these cases, the IPMG needs to relay the (S,G) multicast traffic from OISM PE-S to non-OISM PE-R.

When the IPMG-DF for BD-S receives an (S,G) frame from an OISM PE-S, it has to forward it to every non-OISM PE-R that that has interest in (S,G) on a BD-R that is different than BD-S. The IPMG MUST decapsulate the IP multicast packet, do the IP processing, re-encapsulate it for BD-R (changing the MAC SA to the IPMG's own MAC address on BD-R), and send a copy of the frame to PE-R. Note that a given non-OISM PE-R will receive multiple copies of the frame, if it has multiple BDs on which there is interest in the frame.

5.3. P2MP Tunnels

When IR is used to distribute the multicast traffic among the EVPN-PEs, the procedures of Section 5.2 ensure that there will be no duplicate delivery of multicast traffic. That is, no egress PE will ever send a frame twice on any given AC. If P2MP tunnels are being used to distribute the multicast traffic, it is necessary have additional procedures to prevent duplicate delivery.

At the present time, it is not clear that there will be a use case in which OISM nodes need to interwork with non-OISM nodes that use P2MP tunnels. If it is determined that there is such a use case, procedures for it will be included in a future revision of this document.

6. Traffic to/from Outside the EVPN Tenant Domain

In this section, we discuss scenarios where a multicast source outside a given EVPN Tenant Domain sends traffic to receivers inside the domain (as well as, possibly, to receivers outside the domain). This requires the OISM procedures to interwork with various layer 3 multicast routing procedures.

We assume in this section that the Tenant Domain is not being used as an intermediate transit network for multicast traffic; that is, we do not consider the case where the Tenant Domain contains multicast routers that will receive traffic from sources outside the domain and forward the traffic to receivers outside the domain. The transit scenario is considered in Section 7.

We can divide the non-transit scenarios into two classes:

1. One or more of the EVPN PE routers provide the functionality needed to interwork with layer 3 multicast routing procedures.

2. One BD in the Tenant Domain contains external multicast routers ("tenant multicast routers") that are used to interwork the entire Tenant Domain with layer 3 multicast routing procedures.

6.1. Layer 3 Interworking via EVPN OISM PEs

6.1.1. General Principles

Sometimes it is necessary to interwork an EVPN Tenant Domain with an external layer 3 multicast domain (the "external domain"). This is needed to allow EVPN tenant systems to receive multicast traffic from sources ("external sources") outside the EVPN Tenant Domain. It is also needed to allow receivers ("external receivers") outside the EVPN Tenant Domain to receive traffic from sources inside the Tenant Domain.

In order to allow interworking between an EVPN Tenant Domain and an external domain, one or more OISM PEs must be "L3 Gateways". An L3 Gateway participates both in the OISM procedures and in the L3 multicast routing procedures of the external domain.

An L3 Gateway that has interest in receiving (S,G) traffic must be able to determine the best route to S. If an L3 Gateway has interest in (*,G), it must be able to determine the best route to G's RP. In these interworking scenarios, the L3 Gateway must be running a layer 3 unicast routing protocol. Via this protocol, it imports unicast routes (either IP routes or VPN-IP routes) from routers other than EVPN PEs. And since there may be multicast sources inside the EVPN Tenant Domain, the EVPN PEs also need to export, either as IP routes or as VPN-IP routes (depending upon the external domain), unicast routes to those sources.

When selecting the best route to a multicast source or RP, an L3 Gateway might have a choice between an EVPN route and an IP/VPN-IP route. When such a choice exists, the L3 Gateway SHOULD always prefer the EVPN route. This will ensure that when traffic originates in the Tenant Domain and has a receiver in the tenant domain, the path to that receiver will remain within the EVPN tenant domain, even if the source is also reachable via a routed path. This also provides protection against sub-optimal routing that might occur if two EVPN PEs export IP/VPN-IP routes and each imports the other's IP/VPN-IP routes.

Section 4.2 discusses the way layer 3 multicast states are constructed by OISM PEs. These layer 3 multicast states have IRB interfaces as their IIF and OIF list entries, and are the basis for interworking OISM with other layer 3 multicast procedures such as MVPN or PIM. From the perspective of the layer 3 multicast

procedures running in a given L3 Gateway, an EVPN Tenant Domain is a set of IRB interfaces.

When interworking an EVPN Tenant Domain with an external domain, the L3 Gateway's layer 3 multicast states will not only have IRB interfaces as IIF and OIF list entries, but also other "interfaces" that lead outside the Tenant Domain. For example, when interworking with MVPN, the multicast states may have MVPN tunnels as well as IRB interfaces as IIF or OIF list members. When interworking with PIM, the multicast states may have PIM-enabled non-IRB interfaces as IIF or OIF list members.

As long as a Tenant Domain is not being used as an intermediate transit network for IP multicast traffic, it is not necessary to enable PIM on its IRB interfaces.

In general, an L3 Gateway has the following responsibilities:

o It exports, to the external domain, unicast routes to those multicast sources in the EVPN Tenant Domain that are locally attached to the L3 Gateway.

o It imports, from the external domain, unicast routes to multicast sources that are in the external domain.

o It executes the procedures necessary to draw externally sourced multicast traffic that is of interest to locally attached receivers in the EVPN Tenant Domain. When such traffic is received, the traffic is sent down the IRB interfaces of the BDs on which the locally attached receivers reside.

One of the L3 Gateways in a given Tenant Domain becomes the "DR" for the SBD.(See Section 6.1.2.4.) This L3 gateway has the following additional responsibilities:

o It exports, to the external domain, unicast routes to multicast sources that in the EVPN Tenant Domain that are not locally attached to any L3 gateway.

o It imports, from the external domain, unicast routes to multicast sources that are in the external domain.

o It executes the procedures necessary to draw externally sourced multicast traffic that is of interest to receivers in the EVPN Tenant Domain that are not locally attached to an L3 gateway. When such traffic is received, the traffic is sent down the SBD IRB interface. OISM procedures already described in this document will then ensure that the IP multicast traffic gets distributed

throughout the Tenant Domain to any EVPN PEs that have interest in it. Thus to an OISM PE that is not an L3 gateway the externally sourced traffic will appear to have been sourced on the SBD.

In order for this to work, some special care is needed when an L3 gateway creates or modifies a layer 3 (*,G) multicast state. Suppose group G has both external sources (sources outside the EVPN Tenant Domain) and internal sources (sources inside the EVPN tenant domain). Section 4.2 states that when there are internal sources, the SBD IRB interface must not be added to the OIF list of the (*,G) state. Traffic from internal sources will already have been delivered to all the EVPN PEs that have interest in it. However, if the OIF list of the (*,G) state does not contain its SBD IRB interface, then traffic from external sources will not get delivered to other EVPN PEs.

One way of handling this is the following. When a L3 gateway receives (S,G) traffic from other than an IRB interface, and the traffic corresponds to a layer 3 (*,G) state, the L3 gateway can create (S,G) state. The IIF will be set to the external interface over which the traffic is expected. The OIF list will contain the SBD IRB interface, as well as the IRB interfaces of any other BDs attached to the PEG DR that have locally attached receivers with interest in the (S,G) traffic. The (S,G) state will ensure that the external traffic is sent down the SBD IRB interface. The following text will assume this procedure; however other implementation techniques may also be possible.

If a particular BD is attached to several L3 Gateways, one of the L3 Gateways becomes the DR for that BD. (See Section 6.1.2.4.) If the interworking scenario requires FHR functionality, it is generally the DR for a particular BD that is responsible for performing that functionality on behalf of the source hosts on that BD. (E.g., if the interworking scenario requires that PIM Register messages be sent by a FHR, the DR for a given BD would send the PIM Register messages for sources on that BD.) Note though that the DR for the SBD does not perform FHR functionality on behalf of external sources.

An optional alternative is to have each L3 gateway perform FHR functionality for locally attached sources. Then the DR would only have to perform FHR functionality on behalf of sources that are locally attached to itself AND sources that are not attached to any L3 gateway.

6.1.2. Interworking with MVPN

In this section, we specify the procedures necessary to allow EVPN PEs running OISM procedures to interwork with L3VPN PEs that run BGP-based MVPN ([RFC6514]) procedures. More specifically, the procedures

herein allow a given EVPN Tenant Domain to become part of an L3VPN/MVPN, and support multicast flows where either:

- The source of a given multicast flow is attached to an ethernet segment whose BD is part of an EVPN Tenant Domain, and one or more receivers of the flow are attached to the network via L3VPN/MVPN. (Other receivers may be attached to the network via EVPN.)

- The source of a given multicast flow is attached to the network via L3VPN/MVPN, and one or more receivers of the flow are attached to an ethernet segment that is part of an EVPN tenant domain. (Other receivers may be attached via L3VPN/MVPN.)

In this interworking model, existing L3VPN/MVPN PEs are unaware that certain sources or receivers are part of an EVPN Tenant Domain. The existing L3VPN/MVPN nodes run only their standard procedures and are entirely unaware of EVPN. Interworking is achieved by having some or all of the EVPN PEs function as L3 Gateways running L3VPN/MVPN procedures, as detailed in the following sub-sections.

In this section, we assume that there are no tenant multicast routers on any of the EVPN-attached ethernet segments. (There may of course be multicast routers in the L3VPN.) Consideration of the case where there are tenant multicast routers is deferred till Section 7.)

To support MVPN/EVPN interworking, we introduce the notion of an MVPN/EVPN Gateway, or MEG.

A MEG is an L3 Gateway (see Section 6.1.1), hence is both an OISM PE and an L3VPN/MVPN PE. For a given EVPN Tenant Domain it will have an IP-VRF. If the Tenant Domain is part of an L3VPN/MVPN, the IP-VRF also serves as an L3VPN VRF ([RFC4364]). The IRB interfaces of the IP-VRF are considered to be "VRF interfaces" of the L3VPN VRF. The L3VPN VRF may also have other local VRF interfaces that are not EVPN IRB interfaces.

The VRF on the MEG will import VPN-IP routes ([RFC4364]) from other L3VPN Provider Edge (PE) routers. It will also export VPN-IP routes to other L3VPN PE routers. In order to do so, it must be appropriately configured with the Route Targets used in the L3VPN to control the distribution of the VPN-IP routes. These Route Targets will in general be different than the Route Targets used for controlling the distribution of EVPN routes, as there is no need to distribute EVPN routes to L3VPN-only PEs and no reason to distribute L3VPN/MVPN routes to EVPN-only PEs.

Note that the RDs in the imported VPN-IP routes will not necessarily conform to the EVPN rules (as specified in [RFC7432]) for creating

RDs. Therefore a MEG MUST NOT expect the RDs of the VPN-IP routes to be of any particular format other than what is required by the L3VPN/ MVPN specifications.

The VPN-IP routes that a MEG exports to L3VPN are subnet routes and/ or host routes for the multicast sources that are part of the EVPN tenant domain. The exact set of routes that need to be exported is discussed in Section 6.1.2.2.

Each IMET route originated by a MEG SHOULD carry a flag or Extended Community (to be determined) indicating that the originator of the IMET route is a MEG. However, PE1 will consider PE2 to be a MEG if PE1 imports at least one IMET route from PE2 that carries the flag or EC.

All the MEGs of a given Tenant Domain attach to the SBD of that domain, and one of them is selected to be the SBD's Designated Router (DR) for the domain. The selection procedure is discussed in Section 6.1.2.4.

In this model of operation, MVPN procedures and EVPN procedures are largely independent. In particular, there is no assumption that MVPN and EVPN use the same kind of tunnels. Thus no special procedures are needed to handle the common scenarios where, e.g., EVPN uses VXLAN tunnels but MVPN uses MPLS P2MP tunnels, or where EVPN uses Ingress Replication but MVPN uses MPLS P2MP tunnels.

Similarly, no special procedures are needed to prevent duplicate data delivery on ethernet segments that are multi-homed.

The MEG does have some special procedures (described below) for interworking between EVPN and MVPN; these have to do with selection of the Upstream PE for a given multicast source, with the exporting of VPN-IP routes, and with the generation of MVPN C-multicast routes triggered by the installation of SMET routes.

6.1.2.1. MVPN Sources with EVPN Receivers

6.1.2.1.1. Identifying MVPN Sources

Consider a multicast source S. It is possible that a MEG will import both an EVPN unicast route to S and a VPN-IP route (or an ordinary IP route), where the prefix length of each route is the same. In order to draw (S,G) multicast traffic for any group G, the MEG SHOULD use the EVPN route rather than the VPN-IP or IP route to determine the "Upstream PE" (see section 5 of [RFC6513]).

Doing so ensures that when an EVPN tenant system desires to receive a multicast flow from another EVPN tenant system, the traffic from the source to that receiver stays within the EVPN domain. This prevents problems that might arise if there is a unicast route via L3VPN to S, but no multicast routers along the routed path. This also prevents problem that might arise as a result of the fact that the MEGs will import each others' VPN-IP routes.

In the Section 6.1.2.1.2, we describe the procedures to be used when the selected route to S is a VPN-IP route.

6.1.2.1.2. Joining a Flow from an MVPN Source

Suppose a tenant system R wants to receive (S,G) multicast traffic, where source S is not attached to any PE in the EVPN Tenant Domain, but is attached to an MVPN PE.

- Suppose R is on a singly homed ethernet segment of BD-R, and that segment is attached to PE1, where PE1 is a MEG. PE1 learns via IGMP/MLD listening that R is interested in (S,G). PE1 determines from its VRF that there is no route to S within the Tenant Domain (i.e., no EVPN RT-2 route with S's IP address), but that there is a route to S via L3VPN (i.e., the VRF contains a subnet or host route to S that was received as a VPN-IP route). PE1 thus originates (if it hasn't already) an MVPN C-multicast Source Tree Join(S,G) route. The route is constructed according to normal MVPN procedures.

  The layer 2 multicast state is constructed as specified in Section 4.1.

  In the layer 3 multicast state, the IIF is the appropriate MVPN tunnel, and the IRB interface to BD-R is added to the OIF list.

  When PE1 receives (S,G) traffic from the appropriate MVPN tunnel, it performs IP processing of the traffic, and then sends the traffic down its IRB interface to BD-R. Following normal OISM procedures, the (S,G) traffic will be encapsulated for ethernet and sent out the AC to which R is attached.

- Suppose R is on a singly homed ethernet segment of BD-R, and that segment is attached to PE1, where PE1 is an OISM PE but is NOT a MEG. PE1 learns via IGMP/MLD listening that R is interested in (S,G). PE1 follows normal OISM procedures, originating an SMET route in BD-R for (S,G). Since this route will carry the SBD-RT, it will be received by the MEG that is the DR for the Tenant Domain. The MEG DR can determine from PE1's IMET route whether PE1 is itself a MEG. If PE1 is not a MEG, the MEG DR will

originate (if it hasn't already) an MVPN C-multicast Source Tree Join(S,G) route. This will cause the DR MEG to receive (S,G) traffic on an MVPN tunnel.

The layer 2 multicast state is constructed as specified in Section 4.1.

In the layer 3 multicast state, the IIF is the appropriate MVPN tunnel, and the IRB interface to the SBD is added to the OIF list.

When the DR MEG receives (S,G) traffic on an MVPN tunnel, it performs IP processing of the traffic, and the sends the traffic down its IRB interface to the SBD. Following normal OISM procedures, the traffic will be encapsulated for ethernet and delivered to all PEs in the Tenant Domain that have interest in (S,G), including PE1.

o If R is on a multi-homed ethernet segment of BD-R, one of the PEs attached to the segment will be its DF (following normal EVPN procedures), and the DF will know (via the procedures of [IGMP-Proxy] that a tenant system reachable via one of its local ACs to BD-R is interested in (S,G) traffic. The DF is responsible for originating an SMET route for (S,G), following normal OISM procedures. If the DF is a MEG, it will originate the corresponding MVPN C-multicast Source Tree Join(S,G) route; if the DF is not a MEG, the MEG that is the DR will originate the C-multicast route when it receives the SMET route.

o If R is attached to a non-OISM PE, it will receive the traffic via an IPMG, as specified in Section 5.

If an EVPN-attached receiver is interested in (*,G) traffic, and if it is possible for there to be sources of (*,G) traffic that are attached only to L3VPN nodes, the MEGs will have to know the group-to-RP mappings. That will enable them to originate MVPN C-multicast Shared Tree Join(*,G) routes and to send them towards the RP. (Since we are assuming in this section that there are no tenant multicast routers attached to the EVPN Tenant Domain, the RP must be attached via L3VPN. Alternatively, the MEG itself could be configured to function as an RP for group G.)

The layer 2 multicast states are constructed as specified in Section 4.1.

In the layer 3 (*,G) multicast state, the IIF is the appropriate MVPN tunnel. A MEG will add to the (*,G) OIF list its IRB interfaces for any BDs containing locally attached receivers. If there are receivers attached to other EVPN PEs, then whenever (S,G) traffic

from an external source matches a (*,G) state, the MEG will create (S,G) state, with the MVPN tunnel as the IIF, the OIF list copied from the (*,G) state, and the SBD IRB interface added to the OIF list. (Please see the discussion in Section 6.1.1 regarding the inclusion of the SBD IRB interface in a (*,G) state; the SBD IRB interface is used in the OIF list only for traffic from external sources.)

Normal MVPN procedures will then result in the MEG getting the (*,G) traffic from all the multicast sources for G that are attached via L3VPN. This traffic arrives on MVPN tunnels. When the MEG removes the traffic from these tunnels, it does the IP processing. If there are any receivers on a given BD, BD-R, that are attached via local EVPN ACs, the MEG sends the traffic down its BD-R IRB interface. If there are any other EVPN PEs that are interested in the (*,G) traffic, the MEG sends the traffic down the SBD IRB interface. Normal OISM procedures then distribute the traffic as needed to other EVPN-PEs.

6.1.2.2. EVPN Sources with MVPN Receivers

6.1.2.2.1. General procedures

Consider the case where an EVPN tenant system S is sending IP multicast traffic to group G, and there is a receiver R for the (S,G) traffic that is attached to the L3VPN, but not attached to the EVPN Tenant Domain. (We assume in this document that the L3VPN/MVPN-only nodes will not have any special procedures to deal with the case where a source is inside an EVPN domain.)

In this case, an L3VPN PE through which R can be reached has to send an MVPN C-multicast Join(S,G) route to one of the MEGs that is attached to the EVPN Tenant Domain. For this to happen, the L3VPN PE must have imported a VPN-IP route for S (either a host route or a subnet route) from a MEG.

If a MEG determines that there is multicast source transmitting on one of its ACs, the MEG SHOULD originate a VPN-IP host route for that source. This determination SHOULD be made by examining the IP multicast traffic that arrives on the ACs. (It MAY be made by provisioning.) A MEG SHOULD NOT export a VPN-IP host route for any IP address that is not known to be a multicast source (unless it has some other reason for exporting such a route). The VPN-IP host route for a given multicast source MUST be withdrawn if the source goes silent for a configurable period of time, or if it can be determined that the source is no longer reachable via a local AC.

A MEG SHOULD also originate a VPN-IP subnet route for each of the BDs in the Tenant Domain.

VPN-IP routes exported by a MEG must carry any attributes or extended communities that are required by L3VPN and MVPN. In particular, a VPN-IP route exported by a MEG must carry a VRF Route Import Extended Community corresponding to the IP-VRF from which it is imported, and a Source AS Extended Community.

As a result, if S is attached to a MEG, the L3VPN nodes will direct their MVPN C-multicast Join routes to that MEG. Normal MVPN procedures will cause the traffic to be delivered to the L3VPN nodes. The layer 3 multicast state for (S,G) will have the MVPN tunnel on its OIF list. The IIF will be the IRB interface leading to the BD containing S.

If S is not attached to a MEG, the L3VPN nodes will direct their C-multicast Join routes to whichever MEG appears to be on the best route to S's subnet. Upon receiving the C-multicast Join, that MEG will originate an EVPN SMET route for (S,G). As a result, the MEG will receive the (S,G) traffic at layer 2 via the OISM procedures. The (S,G) traffic will be sent up the appropriate IRB interface, and the layer 3 MVPN procedures will ensure that the traffic is delivered to the L3VPN nodes that have requested it. The layer 3 multicast state for (S,G) will have the MVPN tunnel in the OIF list, and the IIF will be one of the following:

o If S belongs to a BD that is attached to the MEG, the IIF will be the IRB interface to that BD;

o Otherwise the IIF will be the SBD IRB interface.

Note that this works even if S is attached to a non-OISM PE, per the procedures of Section 5.

6.1.2.2.2. Any-Source Multicast (ASM) Groups

Suppose the MEG DR learns that one of the PEs in its Tenant Domain is interested in (*,G), traffic, where G is an Any-Source Multicast (ASM) group. If there are no tenant multicast routers, the MEG DR SHOULD perform the "First Hop Router" (FHR) functionality for group G on behalf of the Tenant Domain, as described in [RFC7761]. This means that the MEG DR must know the identity of the Rendezvous Point (RP) for each group, must send Register messages to the Rendezvous Point, etc.

If the MEG DR is to be the FHR for the Tenant Domain, it must see all the multicast traffic that is sourced from within the domain and

destined to an ASM group address. The MEG can ensure this by originating an SBD-SMET route for (*,*). As an optimization, an SBD-SMET route for (*, "any ASM group"), or even (*, "any ASM group that might have MVPN sources") can be defined.

In some deployment scenarios, it may be preferred that the MEG that receives the (S,G) traffic over an AC be the one provides the FHR functionality. In that case, the MEG DR wold not need to provide the FHR functionality for (S,G) traffic that is attached to another MEG.

Other deployment scenarios are also possible. For example, one might want to configure the MEGs to themselves be RPs. In this case, the RPs would have to exchange with each other information about which sources are active. The method exchanging such information is outside the scope of this document.

6.1.2.2.3. Source on Multihomed Segment

Suppose S is attached to a segment that is all-active multi-homed to PEl and PE2. If S is transmitting to two groups, say G1 and G2, it is possible that PE1 will receive the (S,G1) traffic from S while PE2 receives the (S,G2) traffic from S.

This creates an issue for MVPN/EVPN interworking, because there is no way to cause L3VPN/MVPN nodes to select PE1 as the ingress PE for (S,G1) traffic while selecting PE2 as the ingress PE for (S,G2) traffic.

However, the following procedure ensures that the IP multicast traffic will still flow, even if the L3VPN/MVPN nodes picks the "wrong" EVPN-PE as the Upstream PE for (say) the (S,G1) traffic.

Suppose S is on an ethernet segment, belonging to BD1, that is multi-homed to both PE1 and PE2, where PE1 is a MEG. And suppose that IP multicast traffic from S to G travels over the AC that attaches the segment to PE2 . If PE1 receives a C-multicast Source Tree Join (S,G) route, it MUST originate an SMET route for (S,G). Normal OISM procedures will then cause PE2 to send the (S,G) traffic to PE1 on an EVPN IP multicast tunnel. Normal OISM procedures will also cause PE1 to send the (S,G) traffic up its BD1 IRB interface. Normal MVPN procedures will then cause PE1 to forward the traffic on an MVPN tunnel. In this case, the routing is not optimal, but the traffic does flow correctly.

6.1.2.3. Obtaining Optimal Routing of Traffic Between MVPN and EVPN

The routing of IP multicast traffic between MVPN nodes and EVPN nodes will be optimal as long as there is a MEG along the optimal route. There are various deployment strategies that can be used to obtain optimal routing between MVPN and EVPN.

In one such scenario, a Tenant Domain will have a small number of strategically placed MEGs. For example, a Data Center may have a small number of MEGs that connect it to a wide-area network. Then the optimal route into or out of the Data Center would be through the MEGs.

In this scenario, the MEGs do not need to originate VPN-IP host routes for the multicast sources, they only need to originate VPN-IP subnet routes. The internal structure of the EVPN is completely hidden from the MVPN node. EVPN actions such as MAC Mobility and Mass Withdrawal ([RFC7432]) have zero impact on the MVPN control plane.

While this deployment scenario provides the most optimal routing and has the least impact on the installed based of MVPN nodes, it does complicate network planning considerations.

Another way of providing routing that is close to optimal is to turn each EVPN PE into a MEG. Then routing of MVPN-to-EVPN traffic is optimal. However, routing of EVPN-to-MVPN traffic is not guaranteed to be optimal when a source host is on a multi-homed ethernet segment (as discussed in Section 6.1.2.2.)

The obvious disadvantage of this method is that it requires every EVPN PE to be a MEG.

The procedures specified in this document allow an operator to add MEG functionality to any subset of his EVPN OISM PEs. This allows an operator to make whatever trade-offs he deems appropriate between optimal routing and MEG deployment.

6.1.2.4. DR Selection

Each MEG MUST be configured with an "MEG dummy ethernet segment" that has no ACs.

EVPN supports a number of procedures that can be used to select the Designated Forwarder (DF) for a particular BD on a particular ethernet segment. Some of the possible procedures can be found, e.g., in [RFC7432], [EVPN-DF-NEW], and [EVPN-DF-WEIGHTED]. Whatever

procedure is in use in a given deployment can be adapted to select a MEG DR for a given BD, as follows.

Each MEG will originate an Ethernet Segment route for the MEG dummy ethernet segment. It MUST carry a Route Target derived from the corresponding Ethernet Segment Identifier. Thus only MEGs will import the route.

Once the set of MEGs is known, it is also possible to determine the set of BDs supported by each MEG. The DF selection procedure can then be used to choose a MEG DR for the SBD. (The conditions under which the MEG DR changes depends upon the DF selection algorithm that is in use.)

These procedures can also be used to select a DR for each BD.

6.1.3. Interworking with 'Global Table Multicast'

If multicast service to the outside sources and/or receivers is provided via the BGP-based "Global Table Multicast" (GTM) procedures of [RFC7716], the procedures of Section 6.1.2 can easily be adapted for EVPN/GTM interworking. The way to adapt the MVPN procedures to GTM is explained in [RFC7716].

6.1.4. Interworking with PIM

As we have been discussing, there may be receivers in an EVPN tenant domain that are interested in multicast flows whose sources are outside the EVPN Tenant Domain. Or there may be receivers outside an EVPN Tenant Domain that are interested in multicast flows whose sources are inside the Tenant Domain.

If the outside sources and/or receivers are part of an MVPN, interworking procedures are covered in Section 6.1.2.

There are also cases where an external source or receiver are attached via IP, and the layer 3 multicast routing is done via PIM. In this case, the interworking between the "PIM domain" and the EVPN tenant domain is done at L3 Gateways that perform "PIM/EVPN Gateway" (PEG) functionality. A PEG is very similar to a MEG, except that its layer 3 multicast routing is done via PIM rather than via BGP.

If external sources or receivers for a given group are attached to a PEG via a layer 3 interface, that interface should be treated as a VRF interface attached to the Tenant Domain's L3VPN VRF. The layer 3 multicast routing instance for that Tenant Domain will either run PIM on the VRF interface or will listen for IGMP/MLD messages on that interface. If the external receiver is attached elsewhere on an IP

network, the PE has to enable PIM on its interfaces to the backbone network. In both cases, the PE needs to perform PEG functionality, and its IMET routes must carry a flag or EC identifying it as a PEG.

For each BD on which there is a multicast source or receiver, one of the PEGs will becomes the PEG DR. DR selection can be done using the same procedures specified in Section 6.1.2.4.

As long as there are no tenant multicast routers within the EVPN Tenant Domain, the PEGs do not need to run PIM on their IRB interfaces.

6.1.4.1. Source Inside EVPN Domain

If a PEG receives a PIM Join(S,G) from outside the EVPN tenant domain, it may find it necessary to create (S,G) state. The PE needs to determine whether S is within the Tenant Domain. If S is not within the EVPN Tenant Domain, the PE carries out normal layer 3 multicast routing procedures. If S is within the EVPN tenant domain, the IIF of the (S,G) state is set as follows:

- o if S is on a BD that is attached to the PE, the IIF is the PE's IRB interface to that BD;

- o if S is not on a BD that is attached to the PE, the IIF is the PE's IRB interface to the SBD.

When the PE creates such an (S,G) state, it MUST originate (if it hasn't already) an SBD-SMET route for (S,G). This will cause it to pull the (S,G) traffic via layer 2. When the traffic arrives over an EVPN tunnel, it gets sent up an IRB interface where the layer 3 multicast routing determines the packet's disposition. The SBD-SMET route is withdrawn when the (S,G) state no longer exists (unless there is some other reason for not withdrawing it).

If there are no tenant multicast routers with the EVPN tenant domain, there cannot be an RP in the Tenant Domain, so a PEG does not have to handle externally arriving PIM Join(*,G) messages.

The PEG DR for a particular BD MUST act as the a First Hop Router for that BD. It will examine all (S,G) traffic on the BD, and whenever G is an ASM group, the PEG DR will send Register messages to the RP for G. This means that the PEG DR will need to pull all the (S,G) traffic originating on a given BD, by originating an SMET (*,*) route for that BD. If a PEG DR is the DR for all the BDS, in SHOULD originate just an SBD-SMET (*,*) route rather than an SMET (*,*) route for each BD.

The rules for exporting IP routes to multicast sources are the same as those specified for MEGs in Section 6.1.2.2, except that the exported routes will be IP routes rather than VPN-IP routes, and it is not necessary to attach the VRF Route Import EC or the Source AS EC.

When a source is on a multi-homed segment, the same issue discussed in Section 6.1.2.2.3 exists. Suppose S is on an ethernet segment, belonging to BD1, that is multi-homed to both PE1 and PE2, where PE1 is a PEG. And suppose that IP multicast traffic from S to G travels over the AC that attaches the segment to PE2. If PE1 receives an external PIM Join (S,G) route, it MUST originate an SMET route for (S,G). Normal OISM procedures will cause PE2 to send the (S,G) traffic to PE1 on an EVPN IP multicast tunnel. Normal OISM procedures will also cause PE1 to send the (S,G) traffic up its BD1 IRB interface. Normal PIM procedures will then cause PE1 to forward the traffic along a PIM tree. In this case, the routing is not optimal, but the traffic does flow correctly.

6.1.4.2. Source Outside EVPN Domain

By means of normal OISM procedures, a PEG learns whether there are receivers in the Tenant Domain that are interested in receiving (*,G) or (S,G) traffic. The PEG must determine whether S (or the RP for G) is outside the EVPN Tenant Domain. If so, and if there is a receiver on BD1 interested in receiving such traffic, the PEG DR for BD1 is responsible for originating a PIM Join(S,G) or Join(*,G) control message.

An alternative would be to allow any PEG that is directly attached to a receiver to originate the PIM Joins. Then the PEG DR would only have to originate PIM Joins on behalf of receivers that are not attached to a PEG. However, if this is done, it is necessary for the PEGs to run PIM on all their IRB interfaces, so that the PIM Assert procedures can be used to prevent duplicate delivery to a given BD.

The IIF for the layer 3 (S,G) or (*,G) state is determined by normal PIM procedures. If a receiver is on BD1, and the PEG DR is attached to BD1, its IRB interface to BD1 is added to the OIF list. This ensures that any receivers locally attached to the PEG DR will receive the traffic. If there are receivers attached to other EVPN PEs, then whenever (S,G) traffic from an external source matches a (*,G) state, the PEG will create (S,G) state. The IIF will be set to whatever external interface the traffic is expected to arrive on (copied from the (*,G) state), the OIF list is copied from the (*,G) state, and the SBD IRB interface added to the OIF list.

6.2. Interworking with PIM via an External PIM Router

Section 6.1 describes how to use an OISM PE router as the gateway to a non-EVPN multicast domain, when the EVPN tenant domain is not being used as an intermediate transit network for multicast. An alternative approach is to have one or more external PIM routers (perhaps operated by a tenant) on one of the BDs of the tenant domain. We will refer to this BD as the "gateway BD".

In this model:

- The EVPN Tenant Domain is treated as a stub network attached to the external PIM routers.

- The external PIM routers follow normal PIM procedures, and provide the FHR and LHR functionality for the entire Tenant Domain.

- The OISM PEs do not run PIM.

- If an OISM PE not attached to the gateway BD has interest in a given multicast flow, it conveys that interest to the OISM PEs that are attached to the gateway BD. This is done by following normal OISM procedures. As a result, IGMP/MLD messages will seen by the external PIM routers on the gateway BD, and those external PIM routers will send PIM Join messages externally as required. Traffic of the given multicast flow will then be received by one of the external PIM routers, and that traffic will be forwarded by that router to the gateway BD.

  The normal OISM procedures will then cause the given multicast flow to be tunneled to any PEs of the EVPN Tenant Domain that have interest in the flow. PEs attached to the gateway BD will see the flow as originating from the gateway BD, other PEs will see the flow as originating from the SBD.

- An OISM PE attached to a gateway BD MUST set its layer 2 multicast state to indicate that each AC to the gateway BD has interest in all multicast flows. It MUST also originate an SMET route for (*,*). The procedures for originating SMET routes are discussed in Section 2.5.

- This will cause the OISM PEs attached to the gateway BD to receive all the IP multicast traffic that is sourced within the EVPN tenant domain, and to transmit that traffic to the gateway BD, where the external PIM routers will see it. (Of course, if the gateway BD has a multi-homed segment, only the PE that is the DF for that segment will transmit the multicast traffic to the segment.)

7. Using an EVPN Tenant Domain as an Intermediate (Transit) Network for Multicast traffic

In this section, we consider the scenario where one or more BDs of an EVPN Tenant Domain are being used to carry IP multicast traffic for which the source and at least one receiver are not part the tenant domain. That is, one or more BDs of the Tenant Domain are intermediate "links" of a larger multicast tree created by PIM.

We define a "tenant multicast router" as a multicast router, running PIM, that is:

attached to one or more BDs of the Tenant Domain, but

is not an EVPN PE router.

In order an EVPN Tenant Domain to be used as a transit network for IP multicast, one or more of its BDs must have tenant multicast routers, and an OISM PE that attaching to such a BD MUST be provisioned to enable PIM on its IRB interface to that BD. (This is true even if none of the tenant routers is on a segment attached to the PE.) Further, all the OISM PEs (even ones not attached to a BD with tenant multicast routers) MUST be provisioned to enable PIM on their SBD IRB interfaces.

If PIM is enabled on a particular BD, the DR Selection procedure of Section 6.1.2.4 MUST be replaced by the normal PIM DR Election procedure of [RFC7761]. Note that this may result in one of the tenant routers being selected as the DR, rather than one of the OISM PE routers. In this case, First Hop Router and Last Hop Router functionality will not be performed by any of the EVPN PEs.

A PIM control message on a particular BD is considered to be a link-local multicast message, and as such is sent transparently from PE to PE via the BUM tunnel for that BD. This is true whether the control message was received from an AC, or whether it was received from the local layer 3 routing instance via an IRB interface.

A PIM Join/Prune message contains three fields that are relevant to the present discussion:

o Upstream Neighbor

o Group Address (G)

o Source Address (S), omitted in the case of (*,G) Join/Prune messages.

We will generally speak of a PIM Join as a "Join(S,G)" or a "Join(*,G)" message, and will use the term "Join(X,G)" to mean "either Join(S,G) or Join(*,G)". In the context of a Join(X,G), we will use the term "X" to mean "S in the case of (S,G), or G's RP in the case of (*,G)".

Suppose BD1 contains two tenant multicast routers, C1 and C2. Suppose C1 is on a segment attached to PE1, and C2 is on a segment attached to PE2. When C1 sends a PIM Join(X,G) to BD1, the Upstream Neighbor field might be set to either PE1, PE2, or C2. C1 chooses the Upstream Neighbor based on its unicast routing. Typically, it will choose as the Upstream Neighbor the PIM router on BD1 that is "closest" (according to the unicast routing) to X. Note that this will not necessarily be PE1. PE1 may not even be visible to the unicast routing algorithm used by the tenant routers. Even if it is, it is unlikely to be the PIM router that is closest to X. So we need to consider the following two cases:

C1 sends a PIM Join(X,G) to BD1, with PE1 as the Upstream Neighbor.

PE1's PIM routing instance will see the Join arrive on the BD1 IRB interface. If X is not within the Tenant Domain, PE1 handles the Join according to normal PIM procedures. This will generally result in PE1 selecting an Upstream Neighbor and sending it a Join(X,G).

If X is within the Tenant Domain, but is attached to some other PE, PE1 sends (if it hasn't already) an SBD-SMET route for (X,G). The IIF of the layer 3 (X,G) state will be the SBD IRB interface, and the OIF list will include the IRB interface to BD1.

The SBD-SMET route will pull the (X,G) traffic to PE1, and the (X,G) state will result in the (X,G) traffic being forwarded to C1.

If X is within the Tenant Domain, but is attached to PE1 itself, no SBD-SMET route is sent. The IIF of the layer 3 (X,G) state will be the IRB interface to X's BD, and the OIF list will include the IRB interface to BD1.

C1 sends a PIM Join(X,G) to BD1, with either PE2 or C2 as the Upstream Neighbor.

PE1's PIM routing instance will see the Join arrive on the BD1 IRB interface. If neither X nor Upstream Neighbor is within the

tenant domain, PE1 handles the Join according to normal PIM procedures. This will NOT result in PE1 sending a Join(X,G).

If either X or Upstream Neighbor is within the Tenant Domain, PE1 sends (if it hasn't already) an SBD-SMET route for (X,G). The IIF of the layer 3 (X,G) state will be the SBD IRB interface, and the OIF list will include the IRB interface to BD1.

The SBD-SMET route will pull the (X,G) traffic to PE1, and the (X,G) state will result in the (X,G) traffic being forwarded to C1.

8. IANA Considerations

To be supplied.

9. Security Considerations

This document uses protocols and procedures defined in the normative references, and inherits the security considerations of those references.

This document adds flags or Extended Communities (ECs) to a number of BGP routes, in order to signal that particular nodes support the OISM, IPMG, MEG, and/or PEG functionalities that are defined in this document. Incorrect addition, removal, or modification of those flags and/or ECs will cause the procedures defined herein to malfunction, in which case loss or diversion of data traffic is possible.

10. Acknowledgements

The authors thank Vikram Nagarajan and Princy Elizabeth for their work on Section 6.2. The authors also benefited tremendously from discussions with Aldrin Isaac on EVPN multicast optimizations.

11. References

11.1. Normative References

[EVPN-AR] Rabadan, J., Ed., "Optimized Ingress Replication solution for EVPN", internet-draft ietf-bess-evpn-optimized-ir-02.txt, August 2017.

[EVPN-BUM]
Zhang, Z., Lin, W., Rabadan, J., and K. Patel, "Updates on EVPN BUM Procedures", internet-draft ietf-bess-evpn-bum-procedure-updates-02.txt, September 2017.

[EVPN-IRB]
Sajassi, A., Salam, S., Thoria, S., Drake, J., Rabadan, J., and L. Yong, "Integrated Routing and Bridging in EVPN", internet-draft draft-ietf-bess-evpn-inter-subnet-forwarding-03.txt, February 2017.

[EVPN_IP_Prefix]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", internet-draft ietf-bess-evpn-prefix-advertisement-09.txt, November 2017.

[IGMP-Proxy]
Sajassi, A., Thoria, S., Patel, K., Yeung, D., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", internet-draft draft-ietf-bess-evpn-igmp-mld-proxy-00.txt, March 2017.

[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>.

[RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, DOI 10.17487/RFC2236, November 1997, <https://www.rfc-editor.org/info/rfc2236>.

[RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, DOI 10.17487/RFC2710, October 1999, <https://www.rfc-editor.org/info/rfc2710>.

[RFC6625] Rosen, E., Ed., Rekhter, Y., Ed., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, DOI 10.17487/RFC6625, May 2012, <https://www.rfc-editor.org/info/rfc6625>.

[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>.

11.2. Informative References

[EVPN-BIER]
Zhang, Z., Przygienda, A., Sajassi, A., and J. Rabadan, "EVPN BUM Using BIER", internet-draft ietf-bier-evpn-00.txt, August 2017.

[EVPN-DF-NEW]
Mohanty, S., Patel, K., Sajassi, A., Drake, J., and T. Przygienda, "A new Designated Forwarder Election for the EVPN", internet-draft ietf-bess-evpn-df-election-03.txt, October 2017.

[EVPN-DF-WEIGHTED]
Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., and S. Mohanty, "Preference-based EVPN DF Election", internet-draft ietf-bess-evpn-pref-df-00.txt, June 2017.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[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>.

[RFC7716] Zhang, J., Giuliano, L., Rosen, E., Ed., Subramanian, K., and D. Pacella, "Global Table Multicast with BGP Multicast VPN (BGP-MVPN) Procedures", RFC 7716, DOI 10.17487/RFC7716, December 2015, <https://www.rfc-editor.org/info/rfc7716>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

Appendix A. Integrated Routing and Bridging

This Appendix provides a short tutorial on the interaction of routing and bridging. First it shows the traditional model, where bridging and routing are performed in separate boxes. Then it shows the model specified in [EVPN-IRB], where a single box contains both routing and bridging functions. The latter model is presupposed in the body of this document.

Figure 1 shows a "traditional" router that only does routing and has no L2 bridging capabilities. There are two LANs, LAN1 and LAN2. LAN1 is realized by switch1, LAN2 by switch2. The router has an interface, "lan1" that attaches to LAN1 (via switch1) and an interface "lan2" that attachs to LAN2 (via switch2). Each intreface is configured, as an IP interface, with an IP address and a subnet mask.

```
            +-------+        +--------+        +-------+
            |       |    lan1|        |lan2    |       |
    H1 -----+Switch1+--------+ Router1+--------+Switch2+------H3
            |       |        |        |        |       |
    H2 -----|       |        |        |        |       |
            +-------+        +--------+        +-------+
        |_________________|              |__________________|
            LAN1                                 LAN2
```

Figure 1: Conventional Router with LAN Interfaces

IP traffic (unicast or multicast) that remains within a single subnet never reaches the router. For instance, if H1 emits an ethernet frame with H2's MAC address in the ethernet destination address field, the frame will go from H1 to Switch1 to H2, without ever reaching the router. Since the frame is never seen by a router, the IP datagram within the frame remains entirely unchanged; e.g., its TTL is not decremented. The ethernet Source and Destination MAC addresses are not changed either.

If H1 wants to send a unicast IP datagram to H3, which is on a different subnet, H1 has to be configured with the IP address of a

"default router". Let's assume that H1 is configured with an IP address of Router1 as its default router address. H1 compares H3's IP address with its own IP address and IP subnet mask, and determines that H3 is on a different subnet. So the packet has to be routed. H1 uses ARP to map Router1's IP address to a MAC address on LAN1. H1 then encapsulates the datagram in an ethernet frame, using router1's MAC address as the destination MAC address, and sends the frame to Router1.

Router1 then receives the frame over its lan1 interface. Router1 sees that the frame is addressed to it, so it removes the ethernet encapsulation and processes the IP datagram. The datagram is not addressed to Router1, so it must be forwarded further. Router1 does a lookup of the datagram's IP destination field, and determines that the destination (H3) can be reached via Router1's lan2 interface. Router1 now performs the IP processing of the datagram: it decrements the IP TTL, adjusts the IP header checksum (if present), may fragment the packet is necessary, etc. Then the datagram (or its fragments) are encapsulated in an ethernet header, with Router1's MAC address on LAN2 as the MAC Source Address, and H3's MAC address on LAN2 (which Router1 determines via ARP) as the MAC Destination Address. Finally the packet is sent out the lan2 interface.

If H1 has an IP multicast datagram to send (i.e., an IP datagram whose Destination Address field is an IP Multicast Address), it encapsulates it in an ethernet frame whose MAC Destination Address is computed from the IP Destination Address.

If H2 is a receiver for that multicast address, H2 will receive a copy of the frame, unchanged, from H1. The MAC Source Address in the ethernet encapsulation does not change, the IP TTL field does not get decremented, etc.

If H3 is a receiver for that multicast address, the datagram must be routed to H3. In order for this to happen, Router1 must be configured as a multicast router, and it must accept traffic sent to ethernet multicast addresses. Router1 will receive H1's multicast frame on its lan1 interface, will remove the ethernet encapsulation, and will determine how to dispatch the IP datagram based on Router1's multicast forwarding states. If Router1 knows that there is a receiver for the multicast datagram on LAN2, makes a copy of the datagram, decrements the TTL (and performs any other necessary IP processing), then encapsulates the datagram in ethernet frame for LAN2. The MAC Source Address for this frame will be Router1's MAC Source Address on LAN2. The MAC Destination Address is computed from the IP Destination Address. Finally, the frame is sent out Router1's LAN2 interface.

Figure 2 shows an Integrated Router/Bridge that supports the routing/ bridging integration model of [EVPN-IRB].

```
              +-------------------------------------------+
              |          Integrated Router/Bridge         |
              +-------+        +---------+        +-------+
              |       |    IRB1|   L3    |IRB2    |       |
     H1 -----+  BD1   +--------+Routing +--------+  BD2  +------H3
              |       |        |Instance|        |       |
     H2 -----|       |        |        |        |       |
              +-------+        +---------+        +-------+
          |___________________|           |____________________|
                   LAN1                             LAN2
```

Figure 2: Integrated Router/Bridge

In Figure 2, a single box consists of one or more "L3 Routing Instances". The routing/forwarding tables of a given routing instance is known as an IP-VRF ([EVPN-IRB]). In the context of EVPN, it is convenient to think of each routing instance as representing the routing of a particular tenant. Each IP-VRF is attached to one or more interfaces.

When several EVPN PEs have a routing instance of the same tenant domain, those PEs advertise IP routes to the attached hosts. This is done as specified in [EVPN-IRB].

The integrated router/bridge shown in Figure 2 also attaches to a number of "Broadcast Domains" (BDs). Each BD performs the functions that are performed by the bridges in Figure 1. To the L3 routing instance, each BD appears to be a LAN. The interface attaching a particular BD to a particular IP-VRF is known as an "IRB Interface". From the perspective of L3 routing, each BD is a subnet. Thus each IRB interface is configured with a MAC address (which is the router's

MAC address on the corresponding LAN), as well as an IP address and subnet mask.

The integrated router/bridge shown in Figure 2 may have multiple ACs to each BD. These ACs are visible only to the bridging function, not to the routing instance. To the L3 routing instance, there is just one "interface" to each BD.

If the L3 routing instance represents the IP routing of a particular tenant, the BDs attached to that routing instance are BDs belonging to that same tenant.

Bridging and routing now proceed exactly as in the case of Figure 1, except that BD1 replaces Switch1, BD2 replaces Switch2, interface IRB1 replaces interface lan1, and interface IRB2 replaces interface lan2.

It is important to understand that an IRB interface connects an L3 routing instance to a BD, NOT to a "MAC-VRF". (See [RFC7432] for the definition of "MAC-VRF".) A MAC-VRF may contain several BDs, as long as no MAC address appears in more than one BD. From the perspective of the L3 routing instance, each individual BD is an individual IP subnet; whether each BD has its own MAC-VRF or not is irrelevant to the L3 routing instance.

Figure 3 illustrates IRB when a pair of BDs (subnets) are attached to two different PE routers. In this example, each BD has two segments, and one segment of each BD is attached to one PE router.

```
                +------------------------------------------+
                |        Integrated Router/Bridges         |

                +-------+        +--------+        +-------+
                |       |    IRB1|        |IRB2    |       |
        H1 -----+  BD1  +--------+   PE1  +--------+  BD2  +------H3
                |(Seg-1)|        |(L3 Rtg)|        |(Seg-1)|
        H2 -----|       |        |        |        |       |
                +-------+        +--------+        +-------+
           |___________________|     |       |____________________|
                      LAN1           |                   LAN2
                                     |
                                     |
                +-------+        +--------+        +-------+
                |       |    IRB1|        |IRB2    |       |
        H4 -----+  BD1  +--------+   PE2  +--------+  BD2  +------H5
                |(Seg-2)|        |(L3 Rtg)|        |(Seg-2)|
                |       |        |        |        |       |
                +-------+        +--------+        +-------+
```

Figure 3: Integrated Router/Bridges with Distributed Subnet

If H1 needs to send an IP packet to H4, it determines from its IP address and subnet mask that H4 is on the same subnet as H1. Although H1 and H4 are not attached to the same PE router, EVPN provides ethernet communication among all hosts that are on the same BD. H1 thus uses ARP to find H4's MAC address, and sends an ethernet frame with H4's MAC address in the Destination MAC address field. The frame is received at PE1, but since the Destination MAC address

is not PE1's MAC address, PE1 assumes that the frame is to remain on BD1. Therefore the packet inside the frame is NOT decapsulated, and is NOT send up the IRB interface to PE1's routing instance. Rather, standard EVPN intra-subnet procedures (as detailed in [RFC7432] are used to deliver the frame to PE2, which then sends it to H4.

If H1 needs to send an IP packet to H5, it determines from its IP address and subnet mask that H5 is NOT on the same subnet as H1. Assuming that H1 has been configured with the IP address of PE1 as its default router, H1 sends the packet in an ethernet frame with PE1's MAC address in its Destination MAC Address field. PE1 receives the frame, and sees that the frame is addressed to it. PE1 thus sends the frame up its IRB1 interface to the L3 routing instance. Appropriate IP processing is done (e.g., TTL decrement). The L3 routing instance determines that the "next hop" for H5 is PE2, so the packet is encapsulated (e.g., in MPLS) and sent across the backbone to PE2's routing instance. PE2 will see that the packet's destination, H5, is on BD2 segment-2, and will send the packet down its IRB2 interface. This causes the IP packet to be encapsulated in an ethernet frame with PE2's MAC address (on BD2) in the Source Address field and H5's MAC address in the Destination Address field.

Note that if H1 has an IP packet to send to H3, the forwarding of the packet is handled entirely within PE1. PE1's routing instance sees the packet arrive on its IRB1 interface, and then transmits the packet by sending it down its IRB2 interface.

Often, all the hosts in a particular Tenant Domain will be provisioned with the same value of the default router IP address. This IP address can be assigned, as an "anycast address", to all the EVPN PEs attached to that Tenant Domain. Thus although all hosts are provisioned with the same "default router address", the actual default router for a given host will be one of the PEs that is attached to the same ethernet segment as the host. This provisioning method ensures that IP packets from a given host are handled by the closest EVPN PE that supports IRB.

In the topology of Figure 3, one could imagine that H1 is configured with a default router address that belongs to PE2 but not to PE1. Inter-subnet routing would still work, but IP packets from H1 to H3 would then follow the non-optimal path H1-->PE1-->PE2-->PE1-->H3. Sending traffic on this sort of path, where it leaves a router and then comes back to the same router, is sometimes known as "hairpinning". Similarly, if PE2 supports IRB but PE1 dos not, the same non-optimal path from H1 to H3 would have to be followed. To avoid hairpinning, each EVPN PE needs to support IRB.

It is worth pointing out the way IRB interfaces interact with multicast traffic. Referring again to Figure 3, suppose PE1 and PE2 are functioning as IP multicast routers. Suppose also that H3 transmits a multicast packet, and both H1 and H4 are interested in receiving that packet. PE1 will receive the packet from H3 via its IRB2 interface. The ethernet encapsulation from BD2 is removed, the IP header processing is done, and the packet is then reencapsulated for BD1, with PE1's MAC address in the MAC Source Address field. Then the packet is sent down the IRB1 interface. Layer 2 procedures (as defined in [RFC7432] would then be used to deliver a copy of the packet locally to H1, and remotely to H4.

Please be aware that his document modifies the semantics, described in the previous paragraph, of sending/receiving multicast traffic on an IRB interface. This is explained in Section 1.5.1 and subsequent sections.

Authors' Addresses

Wen Lin
Juniper Networks, Inc.

EMail: wlin@juniper.net

Zhaohui Zhang
Juniper Networks, Inc.

EMail: zzhang@juniper.net

John Drake
Juniper Networks, Inc.

EMail: jdrake@juniper.net

Eric C. Rosen (editor)
Juniper Networks, Inc.

EMail: erosen@juniper.net

Jorge Rabadan
Nokia

EMail: jorge.rabadan@nokia.com

Ali Sajassi
Cisco Systems

EMail: sajassi@cisco.com