BESS
Internet-Draft
Intended status: Standards Track
Expires: April 27, 2018

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

October 24, 2017

EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding
draft-lin-bess-evpn-irb-mcast-04

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

Copyright Notice

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", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

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:

- 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.

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

- 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:

- 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.

- 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.

- 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:

- 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.

- 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.

- o 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.

- o 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:

- o 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.

- o 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.)

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

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

- (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.

- 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.

- 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.

- 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.

- 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.

- 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]).

- 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]).

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

- 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:

- PE1 is interested in receiving the flow;

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

- 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:

o 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.

o 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:

o 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.

o 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

- 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.

- 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:

- 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.

- 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:

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

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

- 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:

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

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

- 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:

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

- 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.

- 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.

- 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:

- 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;

- 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;

- 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:

- 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.

- 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.)

- 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:

- 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.

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

- 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:

- 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.

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

- 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:

- o 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.)

- o 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:

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

- 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-01.txt, December 2016.

[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-05.txt, July 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>.

11.2. Informative References

[EVPN-BIER]
Zhang, Z., Przygienda, A., Sajassi, A., and J. Rabadan, "Updates on EVPN BUM Procedures", internet-draft ietf-zzhang-bier-evpn-00.txt, June 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-02.txt, April 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

BESS Working Group   Y. Liu
Internet Draft   F. Guo
Intended status: Standards Track   Huawei Technologies
Expires: March 19, 2018   X. Liu
                                                                 Jabil
                                                             R. Kebler
                                                      Juniper Networks
                                                          M. Sivakumar
                                                                 Cisco
                                                          Sep 19, 2017

Yang Data Model for Multicast in MPLS/BGP IP VPNs
draft-liu-bess-mvpn-yang-05

Status of this Memo

Copyright Notice

Abstract

This document defines a YANG data model that can be used to configure and manage multicast in MPLS/BGP IP VPNs.

Table of Contents

1. Introduction

YANG [RFC6020] [RFC7950] is a data definition language that was introduced to define the contents of a conceptual data store that allows networked devices to be managed using NETCONF [RFC6241]. YANG is proving relevant beyond its initial confines, as bindings to other interfaces (e.g. REST) and encoding other than XML (e.g. JSON) are being defined. Furthermore, YANG data models can be used as the basis of implementation for other interface, such as CLI and Programmatic APIs.

This document defines a YANG data model that can be used to configure and manage Multicast in MPLS/BGP IP VPN (MVPN). It includes Cisco systems’ solution [RFC6037], BGP MVPN [RFC6513] [RFC6514] etc. Currently this model is incomplete, but it will support the core MVPN protocols, as well as many other features mentioned in separate MVPN RFCs. In addition, Non-core features described in MVPN standards other than mentioned above RFC in future version.

1.1. Requirements Language

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

1.2. Terminology

The terminology for describing YANG data models is found in [RFC6020].

This draft employs YANG tree diagrams, which are explained in [I-D.ietf-netmod-rfc6087bis].

2. Design of Data model

2.1. Scope of model

The model covers Rosen MVPN [RFC6037], BGP MVPN [RFC6513] [RFC6514]. The representation of some of extension features is not completely specified in this draft of the data model. This model is being circulated in its current form for early oversight and review of the basic hierarchy.

The operational state fields of this model are also incomplete, though the structure of what has been written may be taken as representative of the structure of the model when complete.

This model does not cover other MVPN related protocols such as MVPN Extranet [RFC7900] or MVPN MLDP In-band signaling [RFC7246] etc., these will be covered by future Internet Drafts.

2.2. Optional capabilities

This model is designed to represent the capabilities of MVPN devices with various specifications, including some with basic subsets of the MVPN protocols. The main design goals of this draft are that any major now-existing implementation may be said to support the basic model, and that the configuration of all implementations meeting the specification is easy to express through some combination of the features in the basic model and simple vendor augmentations.

On the other hand, operational state parameters are not so widely designated as features, as there are many cases where the defaulting of an operational state parameter would not cause any harm to the system, and it is much more likely that an implementation without native support for a piece of operational state would be able to derive a suitable value for a state variable that is not natively supported.

For the same reason, wide constant ranges (for example, timer maximum and minimum) will be used in the model. It is expected that vendors will augment the model with any specific restrictions that might be required. Vendors may also extend the features list with proprietary extensions.

2.3. Position of address family in hierarchy

The current draft contains MVPN IPv4 and IPv6 as separate schema branches in the structure. The reason for this is to inherit l3vpn yang model structure and make it easier for implementations which may optionally choose to support specific address families. And the names of objects may be different between the IPv4 and IPv6 address families.

3. Module Structure

The MVPN YANG model follows the Guidelines for YANG Module Authors (NMDA) [draft-dsdt-nmda-guidelines-01].The MVPN modules define the network-instance-wide configuration and operational state options in a two-level hierarchy as listed below:

Instance level: Only including configuration data nodes now. MVPN configuration attributes for the entire routing instance, including route-target, I-PMSI tunnel and S-PMSI number, common timer etc.

PMSI tunnel level: MVPN configuration attributes applicable to the I-PMSI and per S-PMSI tunnel configuration attributes, including tunnel mode, tunnel specific parameters and threshold etc. MVPN PMSI tunnel operational state attributes applicable to the I-PMSI and per S-PMSI tunnel operational state attributes, including tunnel mode, tunnel role, tunnel specific parameters and referenced private source and group address etc.

Where fields are not genuinely essential to protocol operation, they are marked as optional. Some fields will be essential but have a default specified, so that they need not be configured explicitly.

We define the MVPN model as a network-instance-centric model, and the MVPN model will augment "/ni:network-instances/ni:network-instance:" in [I-D.ietf-rtgwg-ni-model] and will allow a single mvpn instance per VRF.

```
  augment /ni:network-instances/ni:network-instance:
    +--rw mvpn
```

```
       +--rw mvpnv4
       |  +--rw signaling-mode?                enumeration
       |  +--rw auto-discovery-mode?           enumeration
       |  +--rw config-type?                   enumeration
       |  +--rw is-sender-site?                boolean
       |  +--rw rpt-spt-mode?                  enumeration
       |  +--rw mvpn-route-targets
       |  |  +--rw mvpn-route-target* [rt-type rt-value]
       |  |     +--rw rt-type     enumeration
       |  |     +--rw rt-value    string
       |  +--rw mvpn-ipmsi-tunnel
       |  |  +--rw tunnel-type?                enumeration
       |  |  +--rw (ipmsi-tunnel-attribute)?
       |  |     +--:(p2mp-te)
       |  |     |  +--rw te-p2mp-template?           string
       |  |     +--:(p2mp-mldp)
       |  |     +--:(pim-ssm)
       |  |     |  +--rw ssm-default-group-addr?     inet:ip-address
       |  |     +--:(pim-sm)
       |  |     |  +--rw sm-default-group-addr?      inet:ip-address
       |  |     +--:(bidir-pim)
       |  |     |  +--rw bidir-default-group-addr?   inet:ip-address
       |  |     +--:(ingress-replication)
       |  |     +--:(mp2mp-mldp)
       |  +--rw mvpn-spmsi-tunnels
       |  |  +--rw switch-delay-time?           uint8
       |  |  +--rw switch-back-holddown-time?   uint16
       |  |  +--rw tunnel-limit?                uint16
       |  |  +--rw mvpn-spmsi-tunnel* [tunnel-type]
       |  |     +--rw tunnel-type                     enumeration
       |  |     +--rw (spmsi-tunnel-attribute)?
       |  |     |  +--:(p2mp-te)
       |  |     |  |  +--rw te-p2mp-template?               string
       |  |     |  +--:(p2mp-mldp)
       |  |     |  +--:(pim-ssm)
       |  |     |  |  +--rw ssm-group-pool-addr?            inet:ip-address
       |  |     |  |  +--rw ssm-group-pool-masklength?      uint8
       |  |     |  +--:(pim-sm)
       |  |     |  |  +--rw sm-group-pool-addr?             inet:ip-address
       |  |     |  |  +--rw sm-group-pool-masklength?       uint8
       |  |     |  +--:(bidir-pim)
       |  |     |  |  +--rw bidir-group-pool-addr?          inet:ip-address
       |  |     |  |  +--rw bidir-group-pool-masklength?    uint8
       |  |     |  +--:(ingress-replication)
       |  |     |  +--:(mp2mp-mldp)
       |  |     +--rw switch-threshold?               uint32
       |  |     +--rw switch-wildcard-mode?           enumeration
       |  |     +--rw (address-mask-or-acl)?
       |  |        +--:(address-mask)
       |  |        |  +--rw ipv4-group-addr?               inet:ipv4-address
       |  |        |  +--rw ipv4-group-masklength?         uint8
       |  |        |  +--rw ipv4-source-addr?              inet:ipv4-address
       |  |        |  +--rw ipv4-source-masklength?        uint8
       |  |        +--:(acl)
       |  |           +--rw group-acl-ipv4?                string
       |  +--ro mvpn-ipmsi-tunnel-info
       |  |  +--ro tunnel-type?                     enumeration
       |  |  +--ro (pmsi-tunnel-attribute)?
       |  |  |  +--:(p2mp-te)
       |  |  |  |  +--ro te-p2mp-id?                      uint16
       |  |  |  |  +--ro te-tunnel-id?                    uint16
       |  |  |  |  +--ro te-extend-tunnel-id?             uint16
       |  |  |  +--:(p2mp-mldp)
       |  |  |  |  +--ro mldp-root-addr?                  inet:ip-address
       |  |  |  |  +--ro mldp-lsp-id?                     string
       |  |  |  +--:(pim-ssm)
       |  |  |  |  +--ro ssm-group-addr?                  inet:ip-address
       |  |  |  +--:(pim-sm)
       |  |  |  |  +--ro sm-group-addr?                   inet:ip-address
       |  |  |  +--:(bidir-pim)
       |  |  |  |  +--ro bidir-group-addr?                inet:ip-address
       |  |  |  +--:(ingress-replication)
       |  |  |  +--:(mp2mp-mldp)
       |  |  +--ro tunnel-role?                     enumeration
       |  |  +--ro mvpn-pmsi-ipv4-ref-sg-entries
       |  |     +--ro mvpn-pmsi-ipv4-ref-sg-entries* [ipv4-source-address ipv4-g
roup-address]
       |  |        +--ro ipv4-source-address    inet:ipv4-address
       |  |        +--ro ipv4-group-address     inet:ipv4-address
       |  +--ro mvpn-spmsi-tunnel-ipv4-info
       |     +--ro mvpn-spmsi-tunnel-ipv4-info* [tunnel-type]
       |        +--ro tunnel-type                      enumeration
       |        +--ro (pmsi-tunnel-attribute)?
       |        |  +--:(p2mp-te)
       |        |  |  +--ro te-p2mp-id?                      uint16
       |        |  |  +--ro te-tunnel-id?                    uint16
       |        |  |  +--ro te-extend-tunnel-id?             uint16
       |        |  +--:(p2mp-mldp)
       |        |  |  +--ro mldp-root-addr?                  inet:ip-address
       |        |  |  +--ro mldp-lsp-id?                     string
       |        |  +--:(pim-ssm)
       |        |  |  +--ro ssm-group-addr?                  inet:ip-address
       |        |  +--:(pim-sm)
       |        |  |  +--ro sm-group-addr?                   inet:ip-address
       |        |  +--:(bidir-pim)
       |        |  |  +--ro bidir-group-addr?                inet:ip-address
       |        |  +--:(ingress-replication)
       |        |  +--:(mp2mp-mldp)
       |        +--ro tunnel-role?                     enumeration
       |        +--ro mvpn-pmsi-ipv4-ref-sg-entries
       |           +--ro mvpn-pmsi-ipv4-ref-sg-entries* [ipv4-source-address ipv
4-group-address]
       |              +--ro ipv4-source-address    inet:ipv4-address
       |              +--ro ipv4-group-address     inet:ipv4-address
       +--rw mvpnv6
          +--rw signaling-mode?                enumeration
          +--rw auto-discovery-mode?           enumeration
          +--rw config-type?                   enumeration
          +--rw is-sender-site?                boolean
          +--rw rpt-spt-mode?                  enumeration
          +--rw mvpn-route-targets
          |  +--rw mvpn-route-target* [rt-type rt-value]
          |     +--rw rt-type     enumeration
          |     +--rw rt-value    string
          +--rw mvpn-ipmsi-tunnel
          |  +--rw tunnel-type?                enumeration
          |  +--rw (ipmsi-tunnel-attribute)?
          |     +--:(p2mp-te)
          |     |  +--rw te-p2mp-template?           string
          |     +--:(p2mp-mldp)
          |     +--:(pim-ssm)
          |     |  +--rw ssm-default-group-addr?     inet:ip-address
          |     +--:(pim-sm)
          |     |  +--rw sm-default-group-addr?      inet:ip-address
          |     +--:(bidir-pim)
          |     |  +--rw bidir-default-group-addr?   inet:ip-address
          |     +--:(ingress-replication)
          |     +--:(mp2mp-mldp)
          +--rw mvpn-spmsi-tunnels
          |  +--rw switch-delay-time?           uint8
          |  +--rw switch-back-holddown-time?   uint16
          |  +--rw tunnel-limit?                uint16
          |  +--rw mvpn-spmsi-tunnel* [tunnel-type]
          |     +--rw tunnel-type                    enumeration
          |     +--rw (spmsi-tunnel-attribute)?
          |     |  +--:(p2mp-te)
          |     |  |  +--rw te-p2mp-template?              string
          |     |  +--:(p2mp-mldp)
          |     |  +--:(pim-ssm)
          |     |  |  +--rw ssm-group-pool-addr?           inet:ip-address
          |     |  |  +--rw ssm-group-pool-masklength?     uint8
          |     |  +--:(pim-sm)
          |     |  |  +--rw sm-group-pool-addr?            inet:ip-address
          |     |  |  +--rw sm-group-pool-masklength?      uint8
          |     |  +--:(bidir-pim)
          |     |  |  +--rw bidir-group-pool-addr?         inet:ip-address
          |     |  |  +--rw bidir-group-pool-masklength?   uint8
          |     |  +--:(ingress-replication)
          |     |  +--:(mp2mp-mldp)
          |     +--rw switch-threshold?              uint32
          |     +--rw switch-wildcard-mode?          enumeration
          |     +--rw (address-mask-or-acl)?
          |        +--:(address-mask)
          |        |  +--rw ipv6-group-addr?               inet:ipv6-address
          |        |  +--rw ipv6-groupmasklength?          uint8
          |        |  +--rw ipv6-source-addr?              inet:ipv6-address
          |        |  +--rw ipv6-source-masklength?        uint8
          |        +--:(acl)
          |           +--rw group-acl-ipv6?                string
          +--ro mvpn-ipmsi-tunnel-info
          |  +--ro tunnel-type?                     enumeration
          |  +--ro (pmsi-tunnel-attribute)?
          |  |  +--:(p2mp-te)
          |  |  |  +--ro te-p2mp-id?                      uint16
          |  |  |  +--ro te-tunnel-id?                    uint16
          |  |  |  +--ro te-extend-tunnel-id?             uint16
          |  |  +--:(p2mp-mldp)
          |  |  |  +--ro mldp-root-addr?                  inet:ip-address
          |  |  |  +--ro mldp-lsp-id?                     string
          |  |  +--:(pim-ssm)
          |  |  |  +--ro ssm-group-addr?                  inet:ip-address
          |  |  +--:(pim-sm)
          |  |  |  +--ro sm-group-addr?                   inet:ip-address
          |  |  +--:(bidir-pim)
          |  |  |  +--ro bidir-group-addr?                inet:ip-address
          |  |  +--:(ingress-replication)
          |  |  +--:(mp2mp-mldp)
          |  +--ro tunnel-role?                     enumeration
          |  +--ro mvpn-pmsi-ipv6-ref-sg-entries
          |     +--ro mvpn-pmsi-ipv6-ref-sg-entries* [ipv6-source-address ipv6-g
roup-address]
          |        +--ro ipv6-source-address    inet:ipv6-address
          |        +--ro ipv6-group-address     inet:ipv6-address
          +--ro mvpn-spmsi-tunnel-ipv6-info
             +--ro mvpn-spmsi-tunnel-ipv6-info* [tunnel-type]
                +--ro tunnel-type                      enumeration
                +--ro (pmsi-tunnel-attribute)?
                |  +--:(p2mp-te)
                |  |  +--ro te-p2mp-id?                      uint16
                |  |  +--ro te-tunnel-id?                    uint16
                |  |  +--ro te-extend-tunnel-id?             uint16
                |  +--:(p2mp-mldp)
                |  |  +--ro mldp-root-addr?                  inet:ip-address
                |  |  +--ro mldp-lsp-id?                     string
                |  +--:(pim-ssm)
                |  |  +--ro ssm-group-addr?                  inet:ip-address
                |  +--:(pim-sm)
                |  |  +--ro sm-group-addr?                   inet:ip-address
                |  +--:(bidir-pim)
                |  |  +--ro bidir-group-addr?                 inet:ip-address
                |  +--:(ingress-replication)
                |  +--:(mp2mp-mldp)
                +--ro tunnel-role?                      enumeration
                +--ro mvpn-pmsi-ipv6-ref-sg-entries
                   +--ro mvpn-pmsi-ipv6-ref-sg-entries* [ipv6-source-address ipv
6-group-address]
                      +--ro ipv6-source-address     inet:ipv6-address
                      +--ro ipv6-group-address      inet:ipv6-address
```

4. MVPN YANG Modules

```
   <CODE BEGINS> file "ietf-mvpn@2017-09-15.yang"
   module ietf-mvpn {
      namespace "urn:ietf:params:xml:ns:yang:ietf-mvpn";
      prefix mvpn;

      import ietf-network-instance {
        prefix ni;
      }

      import ietf-inet-types {
        prefix inet;
      }


      organization
        "IETF BESS(BGP Enabled Services) Working Group";
      contact
        "
        Yisong Liu
        <mailto:liuyisong@huawei.com>
        Feng Guo
        <mailto:guofeng@huawei.com>
        Xufeng Liu
        <mailto:Xufeng_Liu@jabil.com>
        Robert Kebler
        <mailto:rkebler@juniper.net>
        Mahesh Sivakumar
        <mailto:masivaku@cisco.com>";
      description
        "This YANG module defines the generic configuration
         and operational state data for mvpn, which is common across
         all of the vendor implementations of the protocol. It is
         intended that the module will be extended by vendors to
         define vendor-specific mvpn parameters.";

      revision 2017-09-15 {
        description
```

```
          "Update for NMDA version and errata.";
        reference
          "RFC XXXX: A YANG Data Model for MVPN";
      }
      revision 2017-07-03 {
        description
          "Update S-PMSI configuration and errata.";
        reference
          "RFC XXXX: A YANG Data Model for MVPN";
      }
      revision 2016-10-28 {
        description
          "Initial revision.";
        reference
          "RFC XXXX: A YANG Data Model for MVPN";
      }

      grouping mvpn-instance-config {
        description "Mvpn basic configuration per instance.";

        leaf signaling-mode {
          type enumeration {
            enum invalid {
              value "0";
              description "invalid";
            }
            enum bgp {
              value "1";
              description "bgp";
            }
            enum pim {
              value "2";
              description "pim";
            }
            enum mldp {
              value "3";
              description "mldp";
            }
          }
          default "invalid";
          description "Signaling mode for C-multicast route.";
        }
        leaf auto-discovery-mode {
          type enumeration {
            enum none {
              value "0";
              description "none";
            }
            enum ad {
              value "1";
              description "auto-discovery by BGP";
            }
          }
          default "none";
          description "Auto discovery mode.";
        }
        leaf config-type {
          type enumeration {
            enum md {
              value "0";
              description "md(rosen)";
            }
            enum ng {
              value "1";
              description "ng";
            }
          }
          default "md";
          description "Mvpn type, which can be md(rosen) mvpn or ng mvpn.";
        }
        leaf is-sender-site {
          type boolean;
          default "false";
          description "Configure the current PE as a sender PE.";
        }
        leaf rpt-spt-mode {
          type enumeration {
            enum spt-only {
              value "0";
              description
                "Only spt mode for crossing public net.";
            }
            enum rpt-spt {
              value "1";
              description
                "Both rpt and spt mode for corssing public net.";
            }
          }
          default "spt-only";
          description
            "ASM mode in multicast private net for crossing public net.";
        }

      }

      grouping mvpn-vpn-targets {
        description "May be different from l3vpn unicast route-targets";
        container mvpn-route-targets{
          description "Multicast vpn route-targets";
          list mvpn-route-target {
            key "rt-type rt-value" ;
            description
              "List of multicast route-targets" ;
            leaf rt-type {
               type enumeration {
                 enum export-extcommunity {
                   value "0";
                   description "export-extcommunity";
                 }
                 enum import-extcommunity {
                   value "1";
                   description "import-extcommunity";
                 }
               }
               mandatory "true";
               description
                 "rt types are as follows:
                  export-extcommunity: specifies the value of
                  the extended community attribute of the
                  route from an outbound interface to the
                  destination vpn.
                  import-extcommunity: receives routes that
                  carry the specified extended community
                  attribute";
             }
             leaf rt-value {
               type string {
                 length "3..21";
               }
               description
                 "the available mvpn target formats are as
                  follows:
                  - 16-bit as number:32-bit user-defined
                  number, for example, 1:3. an as number
                  ranges from 0 to 65535, and a user-defined
                  number ranges from 0 to 4294967295. The as
                  number and user-defined number cannot be
                  both 0s. That is, a vpn target cannot be 0:0.
                  - 32-bit ip address:16-bit user-defined
                  number, for example, 192.168.122.15:1.
                  The ip address ranges from 0.0.0.0 to
                  255.255.255.255, and the user-defined
                  number ranges from 0 to 65535.";
             }
           }
         }
       }
```

```
      grouping mvpn-ipmsi-tunnel-config {
        description "Default mdt for rosen mvpn and I-PMSI for ng mvpn";

        container mvpn-ipmsi-tunnel {
          description "I-PMSI tunnel configuraton";
          leaf tunnel-type {
            type enumeration {
              enum invalid {
                value "0";
                description "invalid";
              }
              enum p2mp-te {
                value "1";
                description "p2mp-te";
              }
              enum p2mp-mldp {
                value "2";
                description "p2mp-mldp";
              }
              enum pim-ssm {
                value "3";
                description "pim-ssm";
              }
              enum pim-sm {
                value "4";
                description "pim-sm";
              }
              enum bidir-pim {
                value "5";
                description "bidir-pim";
              }
              enum ingress-replication {
                value "6";
                description "ingress-replication";
              }
              enum mp2mp-mldp {
                value "7";
                description "mp2mp-mldp";
              }
            }
            description "I-PMSI tunnel type.";
          }
          choice ipmsi-tunnel-attribute {
            description "I-PMSI tunnel attributes configuration";
            case p2mp-te {
              description "P2mp TE tunnel";
              leaf te-p2mp-template {
                type string {
                  length "1..31";
                }
                description "P2mp te tunnel template";
              }
            }
            case p2mp-mldp {
              description "Mldp tunnel";
            }
            case pim-ssm {
              description "Pim ssm tunnel";
              leaf ssm-default-group-addr {
                type inet:ip-address;
                description "Default mdt or I-PMSI group address.";
              }
            }
            case pim-sm {
              description "Pim sm tunnel";
              leaf sm-default-group-addr {
                type inet:ip-address;
                description "Default mdt or I-PMSI group address.";
              }
            }
            case bidir-pim {
              description "Bidir pim tunnel";
              leaf bidir-default-group-addr {
                type inet:ip-address;
                description "Default mdt or I-PMSI group address.";
              }
            }
            case ingress-replication {
              description "Ingress replication p2p tunnel";
            }
            case mp2mp-mldp {
              description "Mp2mp mldp tunnel";
            }
          }
        }
      }

      grouping mvpn-spmsi-tunnel-basic-config {
        description "S-PMSI tunnel basic configuration";
        leaf tunnel-type {
          type enumeration {
            enum invalid {
              value "0";
              description "invalid";
            }
            enum p2mp-te {
              value "1";
              description "p2mp-te";
            }
            enum p2mp-mldp {
              value "2";
              description "p2mp-mldp";
            }
            enum pim-ssm {
              value "3";
              description "pim-ssm";
            }
            enum pim-sm {
              value "4";
              description "pim-sm";
            }
            enum bidir-pim {
              value "5";
              description "bidir-pim";
            }
            enum ingress-replication {
              value "6";
              description "ingress-replication";
            }
            enum mp2mp-mldp {
              value "7";
              description "mp2mp-mldp";
            }
          }
          description "S-PMSI tunnel type.";
        }
        choice spmsi-tunnel-attribute {
          description "S-PMSI tunnel attributes configuration";
          case p2mp-te {
            description "P2mp te tunnel";
            leaf te-p2mp-template {
              type string {
                length "1..31";
              }
              description "P2mp te tunnel template";
            }
          }
          case p2mp-mldp {
            description "Mldp tunnel";
          }
          case pim-ssm {
            description "Pim ssm tunnel";
            leaf ssm-group-pool-addr {
              type inet:ip-address;
              description "Group pool address for data mdt or pim s-pmsi.";
            }
            leaf ssm-group-pool-masklength {
              type uint8 {
                range "8..128";
              }
              description "Group pool mask for data mdt or pim s-pmsi";
            }
          }
          case pim-sm {
            description "Pim sm tunnel";
            leaf sm-group-pool-addr {
              type inet:ip-address;
              description "Group pool address for data mdt or pim s-pmsi.";
            }
            leaf sm-group-pool-masklength {
              type uint8 {
                range "8..128";
              }
              description "Group pool mask for data mdt or pim s-pmsi";
            }
          }
          case bidir-pim {
            description "Bidir pim tunnel";
            leaf bidir-group-pool-addr {
              type inet:ip-address;
              description "Group pool address for data mdt or pim s-pmsi.";
            }
            leaf bidir-group-pool-masklength {
              type uint8 {
                range "8..128";
              }
              description "Group pool mask for data mdt or pim s-pmsi";
            }
          }
          case ingress-replication {
            description "Ingress replication p2p tunnel";
          }
          case mp2mp-mldp {
            description "Mp2mp mldp tunnel";
          }
        }
        leaf switch-threshold {
          type uint32 {
            range "0..4194304";
          }
          default "0";
          description
            "Multicast packet rate threshold for
             triggering the switching from the
             I-PMSI to the S-PMSI. The value is
```

```
               an integer ranging from 0 to 4194304, in
               kbit/s. The default value is 0.";
          }
          leaf switch-wildcard-mode {
            type enumeration {
              enum source-group {
                value "0";
                description
                  "Wildcard neither for source or group address.";
              }
              enum star-star {
                value "1";
                description
                  "Wildcard for both source and group address.";
              }
              enum star-group {
                value "2";
                description
                  "Wildcard only for source address.";
              }
              enum source-star {
                value "3";
                description
                  "Wildcard only for group address.";
              }
            }
            default "source-group";
            description
              "I-PMSI switching to S-PMSI mode for private net
              wildcard mode, which including (*,*), (*,G), (S,*),
              (S,G) four modes.";
          }
        }

        grouping mvpn-spmsi-tunnel-config-ipv4 {
          description
            "Data mdt for rosen mvpn or S-PMSI for ng mvpn in
             IPv4 private network";

          container mvpn-spmsi-tunnels {
            description "S-PMSI tunnel configuration";
            leaf switch-delay-time {
              type uint8 {
                range "3..60";
              }
              units seconds;
              default "5";
              description
              "Delay for switching from the I-PMSI to
```

```
                the S-PMSI. The value is an integer
                ranging from 3 to 60, in seconds. ";
          }
          leaf switch-back-holddown-time {
            type uint16 {
              range "0..512";
            }
            units seconds;
            default "60";
            description
              "Delay for switching back from the S-PMSI
               to the I-PMSI. The value is an integer
               ranging from 0 to 512, in seconds. ";
          }
          leaf tunnel-limit {
            type uint16 {
              range "1..1024";
            }
            description
              "Maximum number of s-pmsi tunnels allowed.";
          }

          list mvpn-spmsi-tunnel {
            key "tunnel-type";
            description "S-PMSI tunnel attributes configuration";

            uses mvpn-spmsi-tunnel-basic-config;

            choice address-mask-or-acl {
              description
                "Type of definition of private net multicast address range";
              case address-mask {
                description "Use the type of address and mask";
                leaf ipv4-group-addr {
                  type inet:ipv4-address;
                  description
                    "Start and end ipv4 addresses of the group
                     address in private net. ";
                }
                leaf ipv4-group-masklength {
                  type uint8 {
                    range "4..32";
                  }
                  description
                    "Group mask length for ipv4 addresses in
                     the group address pool in private net.";
                }
                leaf ipv4-source-addr {
                  type inet:ipv4-address;
                  description
                    "Start and end ipv4 addresses of the source
                     address in private net.";
                }
                leaf ipv4-source-masklength {
                  type uint8 {
                    range "0..32";
                  }
                  description
                    "Source mask length for ipv4 addresses in
                     the group address pool in private net.";
                }
              }
              case acl {
                description "Use the type of acl";
                leaf group-acl-ipv4 {
                  type string {
                    length "1..32";
                  }
                  description
                    "Specify the (s, g) entry on which the
                     S-PMSI tunnel takes effect.
                     The value is an integer ranging from 3000
                     to 3999 or a string of 32 case-sensitive
                     characters. If no value is specified, the
                     switch-group address pool takes effect on
                     all (s, g).";
                }
              }
            }
          }
        }
      }

      grouping mvpn-spmsi-tunnel-config-ipv6 {
        description
          "Data mdt for rosen mvpn or S-PMSI for ng mvpn in
           IPv6 private network";

        container mvpn-spmsi-tunnels {
          description "S-PMSI tunnel configuration";
          leaf switch-delay-time {
            type uint8 {
              range "3..60";
            }
            units seconds;
            default "5";
            description
            "Delay for switching from the I-PMSI to
```

```
             the S-PMSI. The value is an integer
             ranging from 3 to 60, in seconds. ";
          }
          leaf switch-back-holddown-time {
            type uint16 {
              range "0..512";
            }
            units seconds;
            default "60";
            description
              "Delay for switching back from the S-PMSI
               to the I-PMSI. The value is an integer
               ranging from 0 to 512, in seconds. ";
          }
          leaf tunnel-limit {
            type uint16 {
              range "1..1024";
            }
            description
              "Maximum number of s-pmsi tunnels allowed.";
          }

          list mvpn-spmsi-tunnel {
            key "tunnel-type";
            description "S-PMSI tunnel parameter configuration";

            uses mvpn-spmsi-tunnel-basic-config;

            choice address-mask-or-acl {
              description
                "Type of definition of private net multicast address range";
              case address-mask {
                description "Use the type of address and mask";
                leaf ipv6-group-addr {
                  type inet:ipv6-address;
                  description
                    "Start and end ipv6 addresses of the group
                     address in private net.";
                }
                leaf ipv6-groupmasklength {
                  type uint8 {
                    range "8..128";
                  }
                  description
                    "Group mask length for ipv6 addresses in
                     the group address pool in private net.";
                }
                leaf ipv6-source-addr {
                  type inet:ipv6-address;
                  description
                    "Start and end ipv6 addresses of the source
                     address in private net.";
                }
                leaf ipv6-source-masklength {
                  type uint8 {
                    range "0..128";
                  }
                  description
                    "Source mask length for ipv6 addresses in
                     the group address pool in private net.";
                }
              }
              case acl {
                description "Use the type of acl";
                leaf group-acl-ipv6 {
                  type string {
                    length "1..32";
                  }
                  description
                    "Specify the (s, g) entry on which the
                     S-PMSI tunnel takes effect.
                     The value is an integer ranging from 3000
                     to 3999 or a string of 32 case-sensitive
                     characters. If no value is specified, the
                     switch-group address pool takes effect on
                     all (s, g).";
                }
              }
            }
          }
        }
      }

      grouping mvpn-pmsi-state {
        description "PMSI tunnel operational state information";
        leaf tunnel-type {
          type enumeration {
            enum invalid {
              value "0";
              description "invalid";
            }
            enum p2mp-te {
              value "1";
              description "p2mp-te";
            }
            enum p2mp-mldp {
              value "2";
              description "p2mp-mldp";
            }
            enum pim-ssm {
              value "3";
              description "pim-ssm";
            }
            enum pim-sm {
              value "4";
              description "pim-sm";
            }
            enum bidir-pim {
              value "5";
              description "bidir-pim";
            }
            enum ingress-replication {
              value "6";
              description "ingress-replication";
            }
            enum mp2mp-mldp {
              value "7";
              description "mp2mp-mldp";
            }
          }
          description "PMSI tunnel type.";
        }
        choice pmsi-tunnel-attribute {
          description "PMSI tunnel operational state information for each type";
          case p2mp-te {
            description "P2mp te tunnel";
            leaf te-p2mp-id {
              type uint16 {
                range "0..65535";
              }
              default "0";
              description "P2mp id of the p2mp tunnel.";
            }
            leaf te-tunnel-id {
              type uint16 {
                range "1..65535";
              }
              description "Id of the p2mp tunnel.";
            }
            leaf te-extend-tunnel-id {
              type uint16 {
                range "1..65535";
              }
              description "P2mp extended tunnel interface id.";
            }
          }
          case p2mp-mldp {
            description "P2mp mldp tunnel";
            leaf mldp-root-addr {
              type inet:ip-address;
              description "Ip address of the root of a p2mp ldp lsp.";
            }
            leaf mldp-lsp-id {
              type string {
                length "1..256";
              }
              description "P2mp ldp lsp id.";
            }
          }
          case pim-ssm {
            description "Pim ssm tunnel";
            leaf ssm-group-addr {
              type inet:ip-address;
              description "Group address for pim ssm";
            }
          }
          case pim-sm {
            description "Pim sm tunnel";
            leaf sm-group-addr {
              type inet:ip-address;
              description "Group address for pim sm";
            }
          }
          case bidir-pim {
            description "Bidir pim tunnel";
            leaf bidir-group-addr {
              type inet:ip-address;
              description "Group address for bidir-pim";
            }
          }
          case ingress-replication {
            description "Ingress replication p2p tunnel";
          }
          case mp2mp-mldp {
            description "mp2mp mldp tunnel";
          }
        }
        leaf tunnel-role {
          type enumeration {
            enum none {
              value "0";
              description "none";
            }
            enum root {
              value "1";
              description "root";
            }
            enum leaf {
              value "2";
              description "leaf";
            }
            enum root-and-leaf {
              value "3";
              description "root-and-leaf";
            }
          }
          description "Role of a tunnel node.";
        }
      }

      grouping mvpn-pmsi-ipv4-entry {
        description
          "Multicast entries in ipv4 mvpn referenced the pmsi tunnel";
        container mvpn-pmsi-ipv4-ref-sg-entries {
          description
            "Multicast entries in ipv4 mvpn referenced the pmsi tunnel";
          list mvpn-pmsi-ipv4-ref-sg-entries {
            key "ipv4-source-address ipv4-group-address";
            description
              "IPv4 source and group address of private network entry";
            leaf ipv4-source-address {
              type inet:ipv4-address;
              description
                "IPv4 source address of private network entry
                 in I-PMSI or S-PMSI.";
            }
            leaf ipv4-group-address {
              type inet:ipv4-address;
              description
                "IPv4 group address of private network entry
                 in I-PMSI or S-PMSI.";
            }
          }
        }
      }


      grouping mvpn-pmsi-ipv6-entry {
        description
          "Multicast entries in ipv6 mvpn referenced the pmsi tunnel";
        container mvpn-pmsi-ipv6-ref-sg-entries {
          description
            "Multicast entries in ipv6 mvpn referenced the pmsi tunnel";
          list mvpn-pmsi-ipv6-ref-sg-entries {
            key "ipv6-source-address ipv6-group-address";
            description
              "IPv6 source and group address of private network entry";
            leaf ipv6-source-address {
              type inet:ipv6-address;
              description
                "IPv6 source address of private network entry
                 in I-PMSI or S-PMSI.";
            }
            leaf ipv6-group-address {
              type inet:ipv6-address;
              description
                "IPv6 group address of private network entry
                 in I-PMSI or S-PMSI.";
            }
          }
        }
      }

      grouping mvpn-ipmsi-tunnel-state-ipv4 {
        description
          "Default mdt or I-PMSI operational state information";
        container mvpn-ipmsi-tunnel-info {
          config false;
          description
            "Default mdt or I-PMSI operational state information";
          uses mvpn-pmsi-state;
          uses mvpn-pmsi-ipv4-entry;
        }
      }

      grouping mvpn-ipmsi-tunnel-state-ipv6 {
        description
          "Default mdt or I-PMSI operational state information";
        container mvpn-ipmsi-tunnel-info {
          config false;
          description
            "Default mdt or I-PMSI operational state information";
          uses mvpn-pmsi-state;
          uses mvpn-pmsi-ipv6-entry;
        }
      }

      grouping mvpn-spmsi-tunnel-state-ipv4 {
        description
          "Data mdt or S-PMSI operational state information";
        container mvpn-spmsi-tunnel-ipv4-info {
          config false;
          description
            "Data mdt or S-PMSI operational state information";
          list mvpn-spmsi-tunnel-ipv4-info {
            key "tunnel-type";
            description
              "Data mdt or S-PMSI operational state information";
            uses mvpn-pmsi-state;
            uses mvpn-pmsi-ipv4-entry;
          }
        }
      }

      grouping mvpn-spmsi-tunnel-state-ipv6 {
        description
          "Data mdt or S-PMSI operational state information";
        container mvpn-spmsi-tunnel-ipv6-info {
          config false;
          description
            "Data mdt or S-PMSI operational state information";
          list mvpn-spmsi-tunnel-ipv6-info {
            key "tunnel-type";
            description
              "Data mdt or S-PMSI operational state information";
            uses mvpn-pmsi-state;
            uses mvpn-pmsi-ipv6-entry;
          }
        }
      }

      augment "/ni:network-instances/ni:network-instance" {
        description
          "Augment network instance container for per multicast VRF
           configuration and operational state.";
        container mvpn {
          description
            "Mvpn configuration and operational state information.";
          container mvpnv4 {
            description
              "Configuration of multicast IPv4 vpn specific parameters and
               operational state of multicast IPv4 vpn specific parameters";
            uses mvpn-instance-config;
            uses mvpn-vpn-targets;
            uses mvpn-ipmsi-tunnel-config;
            uses mvpn-spmsi-tunnel-config-ipv4;
            uses mvpn-ipmsi-tunnel-state-ipv4;
            uses mvpn-spmsi-tunnel-state-ipv4;
          }
          container mvpnv6 {
            description
              "Configuration of multicast IPv6 vpn specific parameters and
               operational state of multicast IPv6 vpn specific parameters";
            uses mvpn-instance-config;
            uses mvpn-vpn-targets;
            uses mvpn-ipmsi-tunnel-config;
            uses mvpn-spmsi-tunnel-config-ipv6;
            uses mvpn-ipmsi-tunnel-state-ipv6;
            uses mvpn-spmsi-tunnel-state-ipv6;
          }
        }
      }
   }
   <CODE ENDS>
```

5. Security Considerations

The data model defined does not introduce any security implications. This draft does not change any underlying security issues inherent in [RFC8022].

6. IANA Considerations

TBD

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, October 2010

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, June 2011

[RFC7950] Bjorklund, M., Ed., "The YANG 1.1 Data Modeling Language", RFC 7950, August 2016

[I-D.ietf-netmod-rfc6087bis] Bierman, A., "Guidelines for Authors and Reviewers of YANG Data Model Documents", draft-ietf-netmod-rfc6087bis-14, September 2017.

[I-D.dsdt-nmda-guidelines] M. Bjorklund, J. Schoenwaelder, P. Shafer, K. Watsen, R. Wilton, "Guidelines for YANG Module Authors (NMDA)", draft-dsdt-nmda-guidelines-01, May 2017

7.2. Informative References

[RFC6037] Rosen, E., Cai, Y., and IJ. Wijnands, "Cisco Systems' Solution for Multicast in BGP/MPLS IP VPNs", RFC 6037, October 2010.

[RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

[RFC7246] IJ. Wijnands, P. Hitchen, N. Leymann, W. Henderickx, A. Gulko and J. Tantsura, " Multipoint Label Distribution Protocol In-Band Signaling in a Virtual Routing and Forwarding (VRF) Table Context ", RFC 7246, June 2014.

[RFC7900] Y. Rekhter, E. Rosen, R. Aggarwal, Arktan, Y. Cai and T. Morin, " Extranet Multicast in BGP/IP MPLS VPNs ", RFC 7900, June 2016.

[I-D.ietf-rtgwg-ni-model] Berger, L., Hopps, C., Lindem, A., and D. Bogdanovic, X. Liu, "Network Instance Model", draft-ietf-rtgwg-ni-model-03, July 2017.

[I-D.ietf-bess-l3vpn-yang] D. Jain, K. Patel, P. Brissette, Z. Li, S. Zhuang, X. Liu, J. Haas, S. Esale and B. Wen, "Yang Data Model for BGP/MPLS L3 VPNs", draft-ietf-bess-l3vpn-yang-01, April 2017.

8. Acknowledgments

The authors would like to thank Anish Peter, Stig Venaas for their valuable contributions.

Authors' Addresses

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

Email: liuyisong@huawei.com

Feng Guo
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: guofeng@huawei.com

Xufeng Liu
Jabil
8281 Greensboro Drive, Suite 200
McLean VA 22102
USA

Email: Xufeng_Liu@jabil.com

Robert Kebler
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: rkebler@juniper.net

Mahesh Sivakumar
Cisco Systems, Inc
510 McCarthy Blvd
Milpitas, California 95035
USA

Email: masivaku@cisco.com

BESS Working Group
Internet Draft
Intended status: Standards Track
Expires: May 08, 2019

Y. Liu
F. Guo
Huawei Technologies
S. Litkowski
Orange
X. Liu
Volta Networks
R. Kebler
M. Sivakumar
Juniper Networks

November 08, 2018

Yang Data Model for Multicast in MPLS/BGP IP VPNs
draft-liu-bess-mvpn-yang-07

Abstract

This document defines a YANG data model that can be used to configure and manage multicast in MPLS/BGP IP VPNs.

Table of Contents

1. Introduction

YANG [RFC6020] [RFC7950] is a data definition language that was introduced to define the contents of a conceptual data store that allows networked devices to be managed using NETCONF [RFC6241]. YANG is proving relevant beyond its initial confines, as bindings to other interfaces (e.g. REST) and encoding other than XML (e.g. JSON) are being defined. Furthermore, YANG data models can be used as the basis of implementation for other interface, such as CLI and Programmatic APIs.

This document defines a YANG data model that can be used to configure and manage Multicast in MPLS/BGP IP VPN (MVPN). It includes Cisco systems’ solution [RFC6037], BGP MVPN [RFC6513] [RFC6514] etc. This model will support the core MVPN protocols, as well as many other features mentioned in separate MVPN RFCs. In addition, Non-core features described in MVPN standards other than mentioned above RFC in separate documents.

1.1. Terminology

The terminology for describing YANG data models is found in [RFC6020] & [RFC7950].

The following abbreviations are used in this document and the defined model:

MVPN:

Multicast Virtual Private Network [RFC6513].

PMSI:

P-Multicast Service Interface [RFC6513].

PIM:

Protocol Independent Multicast [RFC7761].

SM:

Sparse Mode [RFC7761].

SSM:

Source Specific Multicast [RFC4607].

BIDIR-PIM:

Bidirectional Protocol Independent Multicast [RFC5015].

MLDP:

Multipoint Label Distribution Protocol [RFC6388].

P2MP TE:

Point to Multipoint Traffic Engineering [RFC4875].

1.2. Tree Diagrams

Tree diagrams used in this document follow the notation defined in [RFC8340].

1.3. Prefixes in Data Node Names

In this document, names of data nodes, actions, and other data model objects are often used without a prefix, as long as it is clear from the context in which YANG module each name is defined. Otherwise, names are prefixed using the standard prefix associated with the

Prefix	YANG module	Reference
ni	ietf-network-instance	[I-D.ietf-ni-model]
l3vpn	ietf-bgp-l3vpn	[I-D.ietf-l3vpn-yang]
inet	ietf-inet-types	[RFC6991]
rt-types	ietf-routing-types	[RFC8294]
acl	ietf-access-control-list	[I-D.ietf-acl-yang]

Table 1: Prefixes and Corresponding YANG Modules

2. Design of Data model

2.1. Scope of model

The model covers Rosen MVPN [RFC6037], BGP MVPN [RFC6513] [RFC6514]. The configuration of MVPN features, and the operational state fields and RPC definitions are not all included in this document of the data model. This model can be extended, though the structure of what has been written may be taken as representative of the structure of the whole model.

This model does not cover other MVPN related protocols such as MVPN Extranet [RFC7900] or MVPN MLDP In-band signaling [RFC7246] etc., these will be specified in separate documents.

2.2. Optional capabilities

This model is designed to represent the capabilities of MVPN devices with various specifications, including some with basic subsets of the MVPN protocols. The main design goals of this document are that any major now-existing implementation may be said to support the basic model, and that the configuration of all implementations meeting the specification is easy to express through some combination of the features in the basic model and simple vendor augmentations.

On the other hand, operational state parameters are not so widely designated as features, as there are many cases where the defaulting of an operational state parameter would not cause any harm to the system, and it is much more likely that an implementation without native support for a piece of operational state would be able to derive a suitable value for a state variable that is not natively supported.

For the same reason, wide constant ranges (for example, timer maximum and minimum) will be used in the model. It is expected that vendors will augment the model with any specific restrictions that might be required. Vendors may also extend the features list with proprietary extensions.

2.3. Position of address family in hierarchy

The current draft contains MVPN IPv4 and IPv6 as separate schema branches in the structure. The reason for this is to inherit l3vpn yang model structure and make it easier for implementations which may optionally choose to support specific address families. And the names of some objects may be different between the IPv4 and IPv6 address families.

3. Module Structure

The MVPN YANG model follows the Guidelines for YANG Module Authors (NMDA) [RFC8342]. The operational state data is combined with the associated configuration data in the same hierarchy [I-D.ietf-netmod-rfc6087bis]. The MVPN modules define for both IPv4 and IPv6 in a two-level hierarchy as listed below:

Instance level: Only including configuration data nodes now. MVPN configuration attributes for the entire routing instance, including route-target, I-PMSI tunnel and S-PMSI number, common timer etc.

PMSI tunnel level: MVPN configuration attributes applicable to the I-PMSI and per S-PMSI tunnel configuration attributes, including tunnel mode, tunnel specific parameters and threshold etc. MVPN PMSI tunnel operational state attributes applicable to the I-PMSI and per S-PMSI tunnel operational state attributes, including tunnel mode, tunnel role, tunnel specific parameters and referenced private source and group address etc.

Where fields are not genuinely essential to protocol operation, they are marked as optional. Some fields will be essential but have a default specified, so that they need not be configured explicitly.

This MVPN model augments "/ni:network-instances/ni:network-instance/ni:ni-type/l3vpn:l3vpn/l3vpn:l3vpn/l3vpn:ipv4:" for IPv4 MVPN service and "/ni:network-instances/ni:network-instance/ni:ni-type/l3vpn:l3vpn/l3vpn:l3vpn/l3vpn:ipv6" for IPv6 MVPN service specified in [I-D.ietf-l3vpn-yang].

```
  augment /ni:network-instances/ni:network-instance/ni:ni-type/l3vpn:l3vpn/l3vpn
:l3vpn/l3vpn:ipv4:
    +--rw multicast
       +--rw signaling-mode?            enumeration
       +--rw auto-discovery-mode?       enumeration
       +--rw mvpn-type?                 enumeration
       +--rw is-sender-site?            boolean {mvpn-sender}?
       +--rw rpt-spt-mode?              enumeration
       +--rw mvpn-route-targets {mvpn-separate-rt}?
       |  +--rw mvpn-route-target* [mvpn-rt-type mvpn-rt-value]
       |     +--rw mvpn-rt-type     enumeration
       |     +--rw mvpn-rt-value    string
       +--rw mvpn-ipmsi-tunnel-ipv4
       |  +--rw tunnel-type?                     enumeration
       |  +--rw (ipmsi-tunnel-attribute)?
       |  |  +--:(p2mp-te)
       |  |  |  +--rw te-p2mp-template?                string
       |  |  +--:(p2mp-mldp)
       |  |  +--:(pim-ssm)
       |  |  |  +--rw ssm-default-group-addr?          rt-types:ip-multicast-gro
up-address
       |  |  +--:(pim-sm)
       |  |  |  +--rw sm-default-group-addr?           rt-types:ip-multicast-gro
up-address
       |  |  +--:(bidir-pim)
       |  |  |  +--rw bidir-default-group-addr?        rt-types:ip-multicast-gro
up-address
       |  |  +--:(ingress-replication)
       |  |  +--:(mp2mp-mldp)
       |  +--ro (pmsi-tunnel-state-attribute)?
       |  |  +--:(p2mp-te)
       |  |  |  +--ro te-p2mp-id?                      uint16
       |  |  |  +--ro te-tunnel-id?                    uint16
       |  |  |  +--ro te-extend-tunnel-id?             uint16
       |  |  +--:(p2mp-mldp)
       |  |  |  +--ro mldp-root-addr?                  inet:ip-address
       |  |  |  +--ro mldp-lsp-id?                     string
       |  |  +--:(pim-ssm)
       |  |  |  +--ro ssm-group-addr?                  rt-types:ip-multicast-gro
up-address
       |  |  +--:(pim-sm)
       |  |  |  +--ro sm-group-addr?                   rt-types:ip-multicast-gro
up-address
       |  |  +--:(bidir-pim)
       |  |  |  +--ro bidir-group-addr?                rt-types:ip-multicast-gro
up-address
       |  |  +--:(ingress-replication)
       |  |  +--:(mp2mp-mldp)
       |  +--ro tunnel-role?                     enumeration
       |  +--ro mvpn-pmsi-ipv4-ref-sg-entries
       |     +--ro mvpn-pmsi-ipv4-ref-sg-entries* [ipv4-source-address ipv4-grou
p-address]
       |        +--ro ipv4-source-address    inet:ipv4-address
       |        +--ro ipv4-group-address     rt-types:ipv4-multicast-group-addre
ss
       +--rw mvpn-spmsi-tunnels-ipv4
          +--rw switch-delay-time?           uint8
          +--rw switch-back-holddown-time?   uint16
          +--rw tunnel-limit?                uint16
          +--rw mvpn-spmsi-tunnel-ipv4* [tunnel-type]
             +--rw tunnel-type                      enumeration
             +--rw (spmsi-tunnel-attribute)?
             |  +--:(p2mp-te)
             |  |  +--rw te-p2mp-template?                string
             |  +--:(p2mp-mldp)
             |  +--:(pim-ssm)
             |  |  +--rw ssm-group-pool-addr?             rt-types:ip-multicast-
group-address
             |  |  +--rw ssm-group-pool-masklength?       uint8
             |  +--:(pim-sm)
             |  |  +--rw sm-group-pool-addr?              rt-types:ip-multicast-
group-address
             |  |  +--rw sm-group-pool-masklength?        uint8
             |  +--:(bidir-pim)
             |  |  +--rw bidir-group-pool-addr?           rt-types:ip-multicast-
group-address
             |  |  +--rw bidir-group-pool-masklength?     uint8
             |  +--:(ingress-replication)
             |  +--:(mp2mp-mldp)
             +--rw switch-threshold?                uint32
             +--rw per-item-tunnel-limit?           uint16
             +--rw switch-wildcard-mode?            enumeration {mvpn-switch-wil
dcard-mode}?
             +--rw (address-mask-or-acl)?
             |  +--:(address-mask)
             |  |  +--rw ipv4-group-addr?                 rt-types:ipv4-multicas
t-group-address
             |  |  +--rw ipv4-group-masklength?           uint8
             |  |  +--rw ipv4-source-addr?                inet:ipv4-address
             |  |  +--rw ipv4-source-masklength?          uint8
             |  +--:(acl-name)
             |     +--rw group-acl-ipv4?                  -> /acl:acls/acl/name
             +--ro (pmsi-tunnel-state-attribute)?
             |  +--:(p2mp-te)
             |  |  +--ro te-p2mp-id?                      uint16
             |  |  +--ro te-tunnel-id?                    uint16
             |  |  +--ro te-extend-tunnel-id?             uint16
             |  +--:(p2mp-mldp)
             |  |  +--ro mldp-root-addr?                  inet:ip-address
             |  |  +--ro mldp-lsp-id?                     string
             |  +--:(pim-ssm)
             |  |  +--ro ssm-group-addr?                  rt-types:ip-multicast-
group-address
             |  +--:(pim-sm)
             |  |  +--ro sm-group-addr?                   rt-types:ip-multicast-
group-address
             |  +--:(bidir-pim)
             |  |  +--ro bidir-group-addr?                rt-types:ip-multicast-
group-address
```

```
             |  +--:(ingress-replication)
             |  +--:(mp2mp-mldp)
             +--ro tunnel-role?                       enumeration
             +--ro mvpn-pmsi-ipv4-ref-sg-entries
                +--ro mvpn-pmsi-ipv4-ref-sg-entries* [ipv4-source-address ipv4-g
roup-address]
                   +--ro ipv4-source-address    inet:ipv4-address
                   +--ro ipv4-group-address     rt-types:ipv4-multicast-group-ad
dress
  augment /ni:network-instances/ni:network-instance/ni:ni-type/l3vpn:l3vpn/l3vpn
:l3vpn/l3vpn:ipv6:
    +--rw multicast
       +--rw signaling-mode?            enumeration
       +--rw auto-discovery-mode?       enumeration
       +--rw mvpn-type?                 enumeration
       +--rw is-sender-site?            boolean {mvpn-sender}?
       +--rw rpt-spt-mode?              enumeration
       +--rw mvpn-route-targets {mvpn-separate-rt}?
       |  +--rw mvpn-route-target* [mvpn-rt-type mvpn-rt-value]
       |     +--rw mvpn-rt-type     enumeration
       |     +--rw mvpn-rt-value    string
       +--ro mvpn-ipmsi-tunnel-ipv6
       |  +--ro tunnel-type?                        enumeration
       |  +--ro (ipmsi-tunnel-attribute)?
       |  |  +--:(p2mp-te)
       |  |  |  +--ro te-p2mp-template?                string
       |  |  +--:(p2mp-mldp)
       |  |  +--:(pim-ssm)
       |  |  |  +--ro ssm-default-group-addr?          rt-types:ip-multicast-gro
up-address
       |  |  +--:(pim-sm)
       |  |  |  +--ro sm-default-group-addr?           rt-types:ip-multicast-gro
up-address
       |  |  +--:(bidir-pim)
       |  |  |  +--ro bidir-default-group-addr?        rt-types:ip-multicast-gro
up-address
       |  |  +--:(ingress-replication)
       |  |  +--:(mp2mp-mldp)
       |  +--ro (pmsi-tunnel-state-attribute)?
       |  |  +--:(p2mp-te)
       |  |  |  +--ro te-p2mp-id?                      uint16
       |  |  |  +--ro te-tunnel-id?                    uint16
       |  |  |  +--ro te-extend-tunnel-id?             uint16
       |  |  +--:(p2mp-mldp)
       |  |  |  +--ro mldp-root-addr?                  inet:ip-address
       |  |  |  +--ro mldp-lsp-id?                     string
       |  |  +--:(pim-ssm)
       |  |  |  +--ro ssm-group-addr?                  rt-types:ip-multicast-gro
up-address
       |  |  +--:(pim-sm)
       |  |  |  +--ro sm-group-addr?                   rt-types:ip-multicast-gro
up-address
       |  |  +--:(bidir-pim)
       |  |  |  +--ro bidir-group-addr?                rt-types:ip-multicast-gro
up-address
       |  |  +--:(ingress-replication)
       |  |  +--:(mp2mp-mldp)
       |  +--ro tunnel-role?                        enumeration
       |  +--ro mvpn-pmsi-ipv6-ref-sg-entries
       |     +--ro mvpn-pmsi-ipv6-ref-sg-entries* [ipv6-source-address ipv6-grou
p-address]
       |        +--ro ipv6-source-address    inet:ipv6-address
       |        +--ro ipv6-group-address     rt-types:ipv6-multicast-group-addre
ss
       +--rw mvpn-spmsi-tunnels-ipv6
          +--rw switch-delay-time?           uint8
          +--rw switch-back-holddown-time?   uint16
          +--rw tunnel-limit?                uint16
          +--rw mvpn-spmsi-tunnel-ipv6* [tunnel-type]
             +--rw tunnel-type                      enumeration
             +--rw (spmsi-tunnel-attribute)?
             |  +--:(p2mp-te)
             |  |  +--rw te-p2mp-template?                string
             |  +--:(p2mp-mldp)
             |  +--:(pim-ssm)
             |  |  +--rw ssm-group-pool-addr?             rt-types:ip-multicast-
group-address
             |  |  +--rw ssm-group-pool-masklength?       uint8
             |  +--:(pim-sm)
             |  |  +--rw sm-group-pool-addr?              rt-types:ip-multicast-
group-address
             |  |  +--rw sm-group-pool-masklength?        uint8
             |  +--:(bidir-pim)
             |  |  +--rw bidir-group-pool-addr?           rt-types:ip-multicast-
group-address
             |  |  +--rw bidir-group-pool-masklength?     uint8
             |  +--:(ingress-replication)
             |  +--:(mp2mp-mldp)
             +--rw switch-threshold?                uint32
             +--rw per-item-tunnel-limit?           uint16
             +--rw switch-wildcard-mode?            enumeration {mvpn-switch-wil
dcard-mode}?
             +--rw (address-mask-or-acl)?
             |  +--:(address-mask)
             |  |  +--rw ipv6-group-addr?                 rt-types:ipv6-multicas
t-group-address
             |  |  +--rw ipv6-groupmasklength?            uint8
             |  |  +--rw ipv6-source-addr?                inet:ipv6-address
             |  |  +--rw ipv6-source-masklength?          uint8
             |  +--:(acl-name)
             |     +--rw group-acl-ipv6?                  -> /acl:acls/acl/name
             +--ro (pmsi-tunnel-state-attribute)?
             |  +--:(p2mp-te)
             |  |  +--ro te-p2mp-id?                      uint16
             |  |  +--ro te-tunnel-id?                    uint16
             |  |  +--ro te-extend-tunnel-id?             uint16
             |  +--:(p2mp-mldp)
             |  |  +--ro mldp-root-addr?                  inet:ip-address
             |  |  +--ro mldp-lsp-id?                     string
             |  +--:(pim-ssm)
             |  |  +--ro ssm-group-addr?                  rt-types:ip-multicast-
group-address
             |  +--:(pim-sm)
             |  |  +--ro sm-group-addr?                   rt-types:ip-multicast-
group-address
             |  +--:(bidir-pim)
             |  |  +--ro bidir-group-addr?               rt-types:ip-multicast-
group-address
             |  +--:(ingress-replication)
             |  +--:(mp2mp-mldp)
             +--ro tunnel-role?                    enumeration
             +--ro mvpn-pmsi-ipv6-ref-sg-entries
                +--ro mvpn-pmsi-ipv6-ref-sg-entries* [ipv6-source-address ipv6-g
roup-address]
                   +--ro ipv6-source-address    inet:ipv6-address
                   +--ro ipv6-group-address     rt-types:ipv6-multicast-group-ad
dress
```

4. MVPN YANG Modules

```
   <CODE BEGINS> file ietf-mvpn@2018-11-08.yang
   module ietf-mvpn {
      yang-version 1.1;
      namespace "urn:ietf:params:xml:ns:yang:ietf-mvpn";
      prefix mvpn;

      import ietf-network-instance {
        prefix ni;
      }

      import ietf-bgp-l3vpn {
        prefix l3vpn;
      }

      import ietf-inet-types {
        prefix inet;
      }

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

      import ietf-access-control-list {
        prefix acl;
      }

      organization
        "IETF BESS(BGP Enabled Services) Working Group";
      contact
        "
        Yisong Liu
        <mailto:liuyisong@huawei.com>
        Stephane Litkowski
        <mailto:stephane.litkowski@orange.com>
        Feng Guo
        <mailto:guofeng@huawei.com>
        Xufeng Liu
```

```
          <mailto:xufeng.liu.ietf@gmail.com>
          Robert Kebler
          <mailto:rkebler@juniper.net>
          Mahesh Sivakumar
          <mailto:sivakumar.mahesh@gmail.com>";
        description
          "This YANG module defines the generic configuration
           and operational state data for mvpn, which is common across
           all of the vendor implementations of the protocol. It is
           intended that the module will be extended by vendors to
           define vendor-specific mvpn parameters.";

        revision 2018-11-08 {
          description
            "Update for leaf type and reference.";
          reference
            "RFC XXXX: A YANG Data Model for MVPN";
        }
        revision 2018-05-10 {
          description
            "Update for Model structure and errata.";
          reference
            "RFC XXXX: A YANG Data Model for MVPN";
        }
        revision 2017-09-15 {
          description
            "Update for NMDA version and errata.";
          reference
            "RFC XXXX: A YANG Data Model for MVPN";
        }
        revision 2017-07-03 {
          description
            "Update S-PMSI configuration and errata.";
          reference
            "RFC XXXX: A YANG Data Model for MVPN";
        }
        revision 2016-10-28 {
          description
            "Initial revision.";
          reference
            "RFC XXXX: A YANG Data Model for MVPN";
        }

        /* Features */
        feature mvpn-sender {
          description
            "Support configuration to specify the current PE as the sender PE";
        }
        feature mvpn-separate-rt {
```

```
        description
          "Support route-targets configuration of MVPN when they are
           different from the route-targets of unicast L3VPN.";
      }
      feature mvpn-switch-wildcard-mode {
        description
          "Support configuration to use wildcard mode when multicast
           packets switch from I-PMSI to S-PMSI.";
      }

      grouping mvpn-instance-config {
        description "Mvpn basic configuration per instance.";

        leaf signaling-mode {
          type enumeration {
            enum invalid {
              value "0";
              description "invalid";
            }
            enum bgp {
              value "1";
              description "bgp";
            }
            enum pim {
              value "2";
              description "pim";
            }
            enum mldp {
              value "3";
              description "mldp";
            }
          }
          default "invalid";
          description "Signaling mode for C-multicast route.";
        }
        leaf auto-discovery-mode {
          type enumeration {
            enum invalid {
              value "0";
              description "no auto-discovery";
            }
            enum pim {
              value "1";
              description "auto-discovery by PIM signaling";
            }
            enum bgp {
              value "2";
              description "auto-discovery by BGP signaling";
            }
          }
          default "invalid";
          description "Auto discovery mode.";
        }
        leaf mvpn-type {
          type enumeration {
            enum rosen-mvpn {
              value "0";
              description "Rosen mvpn mode referenced RFC6037";
            }
            enum ng-mvpn {
              value "1";
              description "BGP/MPLS mvpn mode referenced RFC6513&RFC6514";
            }
          }
          default "ng-mvpn";
          description
            "Mvpn type, which can be rosen mvpn mode or ng mvpn mode.";
        }
        leaf is-sender-site {
          if-feature mvpn-sender;
          type boolean;
          default false;
          description "Configure the current PE as a sender PE.";
        }
        leaf rpt-spt-mode {
          type enumeration {
            enum spt-only {
              value "0";
              description
                "Only spt mode for crossing public net.";
            }
            enum rpt-spt {
              value "1";
              description
                "Both rpt and spt mode for corssing public net.";
            }
          }
          description
            "ASM mode in multicast private net for crossing public net.";
        }
      }/* mvpn-instance-config */

      grouping mvpn-rts {
        description "May be different from l3vpn unicast route-targets";
        container mvpn-route-targets{
          if-feature mvpn-separate-rt;
          description "Multicast vpn route-targets";
          list mvpn-route-target {
            key "mvpn-rt-type mvpn-rt-value" ;
            description
              "List of multicast route-targets" ;
            leaf mvpn-rt-type {
              type enumeration {
                 enum export-extcommunity {
                   value "0";
                   description "export-extcommunity";
                 }
                 enum import-extcommunity {
                   value "1";
                   description "import-extcommunity";
                 }
              }
              description
                 "rt types are as follows:
                  export-extcommunity: specifies the value of
                  the extended community attribute of the
                  route from an outbound interface to the
                  destination vpn.
                  import-extcommunity: receives routes that
                  carry the specified extended community
                  attribute";
            }
            leaf mvpn-rt-value {
              type string {
                length "3..21";
              }
              description
                 "the available mvpn target formats are as
                  follows:
                  - 16-bit as number:32-bit user-defined
                  number, for example, 1:3. an as number
                  ranges from 0 to 65535, and a user-defined
                  number ranges from 0 to 4294967295. The as
                  number and user-defined number cannot be
                  both 0s. That is, a vpn target cannot be 0:0.
                  - 32-bit ip address:16-bit user-defined
                  number, for example, 192.168.122.15:1.
                  The ip address ranges from 0.0.0.0 to
                  255.255.255.255, and the user-defined
                  number ranges from 0 to 65535.";
            }
          }
        }
      }

      grouping mvpn-ipmsi-tunnel-config {
        description
```

```
          "Configuration of default mdt for rosen mvpn
           and I-PMSI for ng mvpn";

        leaf tunnel-type {
          type enumeration {
            enum no-tunnel {
              value "0";
              description "no tunnel information present";
            }
            enum p2mp-te {
              value "1";
              description "p2mp-te";
            }
            enum p2mp-mldp {
              value "2";
              description "p2mp-mldp";
            }
            enum pim-ssm {
              value "3";
              description "pim-ssm";
            }
            enum pim-sm {
              value "4";
              description "pim-sm";
            }
            enum bidir-pim {
              value "5";
              description "bidir-pim";
            }
            enum ingress-replication {
              value "6";
              description "ingress-replication";
            }
            enum mp2mp-mldp {
              value "7";
              description "mp2mp-mldp";
            }
          }
          description "I-PMSI tunnel type.";
        }
        choice ipmsi-tunnel-attribute {
          description "I-PMSI tunnel attributes configuration";
          case p2mp-te {
            description "P2mp TE tunnel";
            leaf te-p2mp-template {
              type string {
                length "1..31";
              }
              description "P2mp te tunnel template";
            }
          }
          case p2mp-mldp {
            description "Mldp tunnel";
          }
          case pim-ssm {
            description "Pim ssm tunnel";
            leaf ssm-default-group-addr {
              type rt-types:ip-multicast-group-address;
              description "Default mdt or I-PMSI group address.";
            }
          }
          case pim-sm {
            description "Pim sm tunnel";
            leaf sm-default-group-addr {
              type rt-types:ip-multicast-group-address;
              description "Default mdt or I-PMSI group address.";
            }
          }
          case bidir-pim {
            description "Bidir pim tunnel";
            leaf bidir-default-group-addr {
              type rt-types:ip-multicast-group-address;
              description "Default mdt or I-PMSI group address.";
            }
          }
          case ingress-replication {
            description "Ingress replication p2p tunnel";
          }
          case mp2mp-mldp {
            description "Mp2mp mldp tunnel";
          }
        }
      }/* mvpn-ipmsi-tunnel-config */

      grouping mvpn-spmsi-tunnel-per-item-config {
        description "S-PMSI tunnel basic configuration";
        leaf tunnel-type {
          type enumeration {
            enum no-tunnel {
              value "0";
              description "no tunnel information present";
            }
            enum p2mp-te {
              value "1";
              description "p2mp-te";
            }
            enum p2mp-mldp {
              value "2";
              description "p2mp-mldp";
            }
            enum pim-ssm {
              value "3";
              description "pim-ssm";
            }
            enum pim-sm {
              value "4";
              description "pim-sm";
            }
            enum bidir-pim {
              value "5";
              description "bidir-pim";
            }
            enum ingress-replication {
              value "6";
              description "ingress-replication";
            }
            enum mp2mp-mldp {
              value "7";
              description "mp2mp-mldp";
            }
          }
          description "S-PMSI tunnel type.";
        }
        choice spmsi-tunnel-attribute {
          description "S-PMSI tunnel attributes configuration";
          case p2mp-te {
            description "P2mp te tunnel";
            leaf te-p2mp-template {
              type string {
                length "1..31";
              }
              description "P2mp te tunnel template";
            }
          }
          case p2mp-mldp {
            description "Mldp tunnel";
          }
          case pim-ssm {
            description "Pim ssm tunnel";
            leaf ssm-group-pool-addr {
              type rt-types:ip-multicast-group-address;
              description "Group pool address for data mdt or pim s-pmsi.";
            }
            leaf ssm-group-pool-masklength {
              type uint8 {
                range "8..128";
              }
              description "Group pool mask for data mdt or pim s-pmsi";
            }
          }
          case pim-sm {
            description "Pim sm tunnel";
            leaf sm-group-pool-addr {
              type rt-types:ip-multicast-group-address;
              description "Group pool address for data mdt or pim s-pmsi.";
            }
            leaf sm-group-pool-masklength {
              type uint8 {
                range "8..128";
              }
              description "Group pool mask for data mdt or pim s-pmsi";
            }
          }
          case bidir-pim {
            description "Bidir pim tunnel";
            leaf bidir-group-pool-addr {
              type rt-types:ip-multicast-group-address;
              description "Group pool address for data mdt or pim s-pmsi.";
            }
            leaf bidir-group-pool-masklength {
              type uint8 {
                range "8..128";
              }
              description "Group pool mask for data mdt or pim s-pmsi";
            }
          }
          case ingress-replication {
            description "Ingress replication p2p tunnel";
          }
          case mp2mp-mldp {
            description "Mp2mp mldp tunnel";
          }
        }
        leaf switch-threshold {
          type uint32 {
            range "0..4194304";
          }
          units "kbps";
          default "0";
          description
            "Multicast packet rate threshold for
             triggering the switching from the
             I-PMSI to the S-PMSI. The value is
             an integer ranging from 0 to 4194304, in
             kbps. The default value is 0.";
        }
        leaf per-item-tunnel-limit {
          type uint16 {
            range "1..1024";
          }
          description
            "Maximum number of S-PMSI tunnels allowed
             per S-PMSI configuration item per mvpn instance.";
        }
        leaf switch-wildcard-mode {
          if-feature mvpn-switch-wildcard-mode;
          type enumeration {
            enum source-group {
              value "0";
              description
                "Wildcard neither for source or group address.";
            }
            enum star-star {
              value "1";
              description
                "Wildcard for both source and group address.";
            }
            enum star-group {
              value "2";
              description
                "Wildcard only for source address.";
            }
            enum source-star {
              value "3";
              description
                "Wildcard only for group address.";
            }
          }
          default "source-group";
          description
            "I-PMSI switching to S-PMSI mode for private net
            wildcard mode, which including (*,*), (*,G), (S,*),
            (S,G) four modes.";
        }
      }/* mvpn-spmsi-tunnel-per-item-config */

      grouping mvpn-spmsi-tunnel-common-config {
        description
          "Data mdt for rosen mvpn or S-PMSI for ng mvpn configuration
           attributes for both IPv4 and IPv6 private network";
        leaf switch-delay-time {
          type uint8 {
            range "3..60";
          }
          units seconds;
          default "5";
          description
            "Delay for switching from the I-PMSI to
             the S-PMSI. The value is an integer
             ranging from 3 to 60, in seconds. ";
        }
        leaf switch-back-holddown-time {
          type uint16 {
            range "0..512";
          }
          units seconds;
          default "60";
          description
            "Delay for switching back from the S-PMSI
             to the I-PMSI. The value is an integer
             ranging from 0 to 512, in seconds. ";
        }
        leaf tunnel-limit {
          type uint16 {
            range "1..8192";
          }
          description
            "Maximum number of s-pmsi tunnels allowed
             per mvpn instance.";
        }
      }/* mvpn-spmsi-tunnel-common-config */

      grouping mvpn-pmsi-state {
        description "PMSI tunnel operational state information";

        choice pmsi-tunnel-state-attribute {
          config false;
          description
            "PMSI tunnel operational state information for each type";
          case p2mp-te {
            description "P2mp te tunnel";
            leaf te-p2mp-id {
              type uint16 {
                range "0..65535";
              }
              default "0";
              description "P2mp id of the p2mp tunnel.";
            }
            leaf te-tunnel-id {
              type uint16 {
                range "1..65535";
              }
              description "Id of the p2mp tunnel.";
            }
            leaf te-extend-tunnel-id {
              type uint16 {
                range "1..65535";
              }
              description "P2mp extended tunnel interface id.";
            }
          }
          case p2mp-mldp {
            description "P2mp mldp tunnel";
            leaf mldp-root-addr {
              type inet:ip-address;
              description "Ip address of the root of a p2mp ldp lsp.";
            }
            leaf mldp-lsp-id {
              type string {
                length "1..256";
              }
              description "P2mp ldp lsp id.";
            }
          }
          case pim-ssm {
            description "Pim ssm tunnel";
            leaf ssm-group-addr {
              type rt-types:ip-multicast-group-address;
              description "Group address for pim ssm";
            }
          }
          case pim-sm {
            description "Pim sm tunnel";
            leaf sm-group-addr {
              type rt-types:ip-multicast-group-address;
              description "Group address for pim sm";
            }
          }
          case bidir-pim {
            description "Bidir pim tunnel";
            leaf bidir-group-addr {
              type rt-types:ip-multicast-group-address;
              description "Group address for bidir-pim";
            }
          }
          case ingress-replication {
            description "Ingress replication p2p tunnel";
          }
          case mp2mp-mldp {
            description "mp2mp mldp tunnel";
          }
        }
        leaf tunnel-role {
          type enumeration {
            enum none {
              value "0";
              description "none";
            }
            enum root {
              value "1";
              description "root";
            }
            enum leaf {
              value "2";
              description "leaf";
            }
            enum root-and-leaf {
              value "3";
              description "root-and-leaf";
            }
          }
          config false;
          description "Role of a node for a p-tunnel.";
        }
      }/* mvpn-pmsi-state */

      grouping mvpn-pmsi-ipv4-entry {
        description
          "Multicast entries in ipv4 mvpn referenced the pmsi tunnel";
        container mvpn-pmsi-ipv4-ref-sg-entries {
          config false;
          description
            "Multicast entries in ipv4 mvpn referenced the pmsi tunnel";
          list mvpn-pmsi-ipv4-ref-sg-entries {
            key "ipv4-source-address ipv4-group-address";
            description
              "IPv4 source and group address of private network entry";
            leaf ipv4-source-address {
              type inet:ipv4-address;
              description
                "IPv4 source address of private network entry
                 in I-PMSI or S-PMSI.";
            }
            leaf ipv4-group-address {
              type rt-types:ipv4-multicast-group-address;
              description
                "IPv4 group address of private network entry
                 in I-PMSI or S-PMSI.";
            }
          }
        }
      }/* mvpn-pmsi-ipv4-entry */
      grouping mvpn-pmsi-ipv6-entry {
        description
          "Multicast entries in ipv6 mvpn referenced the pmsi tunnel";
        container mvpn-pmsi-ipv6-ref-sg-entries {
          config false;
          description
            "Multicast entries in ipv6 mvpn referenced the pmsi tunnel";
          list mvpn-pmsi-ipv6-ref-sg-entries {
            key "ipv6-source-address ipv6-group-address";
            description
              "IPv6 source and group address of private network entry";
            leaf ipv6-source-address {
              type inet:ipv6-address;
              description
                "IPv6 source address of private network entry
                 in I-PMSI or S-PMSI.";
            }
            leaf ipv6-group-address {
              type rt-types:ipv6-multicast-group-address;
              description
                "IPv6 group address of private network entry
                 in I-PMSI or S-PMSI.";
            }
          }
        }
      }/* mvpn-pmsi-ipv6-entry */

      grouping mvpn-ipmsi-tunnel-info-ipv4 {
        description
          "Default mdt or I-PMSI configuration and
           operational state information";
        container mvpn-ipmsi-tunnel-ipv4 {
          description
            "Default mdt or I-PMSI configuration and
             operational state information";
          uses mvpn-ipmsi-tunnel-config;
          uses mvpn-pmsi-state;
          uses mvpn-pmsi-ipv4-entry;
        }
      }

      grouping mvpn-ipmsi-tunnel-info-ipv6 {
        description
          "Default mdt or I-PMSI configuration and
           operational state information";
        container mvpn-ipmsi-tunnel-ipv6 {
          config false;
          description
            "Default mdt or I-PMSI configuration and
             operational state information";
          uses mvpn-ipmsi-tunnel-config;
          uses mvpn-pmsi-state;
          uses mvpn-pmsi-ipv6-entry;
        }
      }

      grouping mvpn-spmsi-tunnel-info-ipv4 {
        description
          "Data mdt for rosen mvpn or S-PMSI for ng mvpn in
           IPv4 private network";

        container mvpn-spmsi-tunnels-ipv4 {
          description
            "S-PMSI tunnel configuration and
             operational state information.";
          uses mvpn-spmsi-tunnel-common-config;

          list mvpn-spmsi-tunnel-ipv4 {
            key "tunnel-type";
            description
              "S-PMSI tunnel attributes configuration and
               operational state information.";

            uses mvpn-spmsi-tunnel-per-item-config;
            choice address-mask-or-acl {
              description
                "Type of definition of private net multicast address range";
              case address-mask {
                description "Use the type of address and mask";
                leaf ipv4-group-addr {
                  type rt-types:ipv4-multicast-group-address;
                  description
                    "Start address of the IPv4 group
                     address range in private net. ";
                }
                leaf ipv4-group-masklength {
                  type uint8 {
                    range "4..32";
                  }
                  description
                    "Group mask length for the IPv4
                     group address range in private net.";
                }
                leaf ipv4-source-addr {
                  type inet:ipv4-address;
                  description
```

```
                    "Start address of the IPv4 source
                     address range in private net.";
                }
                leaf ipv4-source-masklength {
                  type uint8 {
                    range "0..32";
                  }
                  description
                    "Source mask length for the IPv4
                     source address range in private net.";
                }
              }
              case acl-name {
                description "Use the type of acl";
                leaf group-acl-ipv4 {
                  type leafref {
                    path "/acl:acls/acl:acl/acl:name";
                  }
                  description
                    "Specify the (s, g) entry on which the
                     S-PMSI tunnel takes effect.
                     The value is an integer ranging from 3000
                     to 3999 or a string of 32 case-sensitive
                     characters. If no value is specified, the
                     switch-group address pool takes effect on
                     all (s, g).";
                }
              }
            }
            uses mvpn-pmsi-state;
            uses mvpn-pmsi-ipv4-entry;
          }/* list mvpn-spmsi-tunnel-ipv4 */
        }/* container mvpn-spmsi-tunnels-ipv4 */
      }/* grouping mvpn-spmsi-tunnel-info-ipv4 */

      grouping mvpn-spmsi-tunnel-info-ipv6 {
        description
          "Data mdt for rosen mvpn or S-PMSI for ng mvpn in
           IPv6 private network";

        container mvpn-spmsi-tunnels-ipv6 {
          description
            "S-PMSI tunnel configuration and
             operational state information.";
          uses mvpn-spmsi-tunnel-common-config;

          list mvpn-spmsi-tunnel-ipv6 {
            key "tunnel-type";
            description
```

```
              "S-PMSI tunnel attributes configuration and
               operational state information.";
            uses mvpn-spmsi-tunnel-per-item-config;

            choice address-mask-or-acl {
              description
                "Type of definition of private net multicast address range";
              case address-mask {
                description "Use the type of address and mask";

                leaf ipv6-group-addr {
                  type rt-types:ipv6-multicast-group-address;
                  description
                    "Start address of the IPv6 group
                     address range in private net. ";
                }
                leaf ipv6-groupmasklength {
                  type uint8 {
                    range "8..128";
                  }
                  description
                    "Group mask length for the IPv6
                     group address range in private net.";
                }
                leaf ipv6-source-addr {
                  type inet:ipv6-address;
                  description
                    "Start address of the IPv6 source
                     address range in private net.";
                }
                leaf ipv6-source-masklength {
                  type uint8 {
                    range "0..128";
                  }
                  description
                    "Source mask length for the IPv6
                     source address range in private net.";
                }
              }
              case acl-name {
                description "Use the type of acl";
                leaf group-acl-ipv6 {
                  type leafref {
                    path "/acl:acls/acl:acl/acl:name";
                  }
                  description
                    "Specify the (s, g) entry on which the
                     S-PMSI tunnel takes effect.
                     The value is an integer ranging from 3000
                     to 3999 or a string of 32 case-sensitive
                     characters. If no value is specified, the
                     switch-group address pool takes effect on
                     all (s, g).";
                }
              }
            }
            uses mvpn-pmsi-state;
            uses mvpn-pmsi-ipv6-entry;
          }/* list mvpn-spmsi-tunnel-ipv6 */
        }/* container mvpn-spmsi-tunnels-ipv6 */
      }/* grouping mvpn-spmsi-tunnel-info-ipv6 */

      augment "/ni:network-instances/ni:network-instance/ni:ni-type/"
              +"l3vpn:l3vpn/l3vpn:l3vpn/l3vpn:ipv4" {
        description
          "Augment l3vpn ipv4 container for per multicast VRF
           configuration and operational state.";
        container multicast {
          description
            "Configuration of multicast IPv4 vpn specific parameters and
             operational state of multicast IPv4 vpn specific parameters";
          uses mvpn-instance-config;
          uses mvpn-rts;
          uses mvpn-ipmsi-tunnel-info-ipv4;
          uses mvpn-spmsi-tunnel-info-ipv4;
        }
      }

      augment "/ni:network-instances/ni:network-instance/ni:ni-type/"
              +"l3vpn:l3vpn/l3vpn:l3vpn/l3vpn:ipv6" {
        description
          "Augment l3vpn ipv6 container for per multicast VRF
           configuration and operational state.";
        container multicast {
          description
            "Configuration of multicast IPv6 vpn specific parameters and
             operational state of multicast IPv6 vpn specific parameters";
          uses mvpn-instance-config;
          uses mvpn-rts;
          uses mvpn-ipmsi-tunnel-info-ipv6;
          uses mvpn-spmsi-tunnel-info-ipv6;
        }
      }
   }
   <CODE ENDS>
```

5. Security Considerations

TBD

6. IANA Considerations

TBD

7. References

7.1. Normative References

[RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, October 2010

[RFC6037] Rosen, E., Cai, Y., and IJ. Wijnands, "Cisco Systems' Solution for Multicast in BGP/MPLS IP VPNs", RFC 6037, October 2010.

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, June 2011

[RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

[RFC6991] Schoenwaelder, J., Ed., "Common YANG Data Types", RFC 6991, July 2013

[RFC7246] IJ. Wijnands, P. Hitchen, N. Leymann, W. Henderickx, A. Gulko and J. Tantsura, " Multipoint Label Distribution Protocol In-Band Signaling in a Virtual Routing and Forwarding (VRF) Table Context ", RFC 7246, June 2014.

[RFC7900] Y. Rekhter, E. Rosen, R. Aggarwal, Arktan, Y. Cai and T. Morin, " Extranet Multicast in BGP/IP MPLS VPNs ", RFC 7900, June 2016.

[RFC7950] Bjorklund, M., Ed., "The YANG 1.1 Data Modeling Language", RFC 7950, August 2016

[RFC8294] Liu, X., Qu, Y., Lindem, A., Hopps, C., and L. Berger, "Common YANG Data Types for the Routing Area", RFC 8294, December 2017

[RFC8342] Bjorklund, M., Schoenwaelder, J., Shafer, P., Watsen, K., and R. Wilton, "Network Management Datastore Architecture (NMDA)", RFC 8342, March 2018

[I-D.ietf-acl-yang] M. Jethanandani, L. Huang, S. Agarwal and D. Blair, "Network Access Control List (ACL) YANG Data Model", draft-ietf-netmod-acl-model-19(work in progress), April 2018

[I-D.ietf-ni-model] Berger, L., Hopps, C., Lindem, A., and D. Bogdanovic, X. Liu, "Network Instance Model", draft-ietf-rtgwg-ni-model-12(work in progress), March 2018.

[I-D.ietf-l3vpn-yang] D. Jain, K. Patel, P. Brissette, Z. Li, S. Zhuang, X. Liu, J. Haas, S. Esale and B. Wen, "Yang Data Model for BGP/MPLS L3 VPNs", draft-ietf-bess-l3vpn-yang-04(work in progress), October 2018.

7.2. Informative References

[RFC8340] Bjorklund, M. and L. Berger, Ed., "YANG Tree Diagrams", BCP 215, RFC 8340, March 2018

[I-D.ietf-netmod-rfc6087bis] Bierman, A., "Guidelines for Authors and Reviewers of YANG Data Model Documents", draft-ietf-netmod-rfc6087bis-20(work in progress), March 2018

8. Acknowledgments

The authors would like to thank Anish Peter, Stig Venaas for their valuable contributions.

Authors' Addresses

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

Email: liuyisong@huawei.com

Feng Guo
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: guofeng@huawei.com

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Xufeng Liu
Volta Networks

Email: xufeng.liu.ietf@gmail.com

Robert Kebler
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: rkebler@juniper.net

Mahesh Sivakumar
Juniper Networks
1133 Innovation Way
Sunnyvale, California
USA

Email: sivakumar.mahesh@gmail.com

INTERNET-DRAFT

Intended Status: Proposed Standard

N. Malhotra, Ed.
A. Sajassi
A. Pattekar
(Cisco)
A. Lingala
(AT&T)
J. Rabadan
(Nokia)
J. Drake
(Juniper Networks)

Expires: April 29, 2018
October 26, 2017

Extended Mobility Procedures for EVPN-IRB
draft-malhotra-bess-evpn-irb-extended-mobility-01

Abstract

The procedure to handle host mobility in a layer 2 Network with EVPN control plane is defined as part of RFC 7432. EVPN has since evolved to find wider applicability across various IRB use cases that include distributing both MAC and IP reachability via a common EVPN control plane. MAC Mobility procedures defined in RFC 7432 are extensible to IRB use cases if a fixed 1:1 mapping between VM IP and MAC is assumed across VM moves. Generic mobility support for IP and MAC that allows these bindings to change across moves is required to support a broader set of EVPN IRB use cases, and requires further consideration. EVPN all-active multi-homing further introduces scenarios that require additional consideration from mobility perspective. Intent of this draft is to enumerate a set of design considerations applicable to mobility across EVPN IRB use cases and define generic sequence number assignment procedures to address these IRB use cases.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

EVPN-IRB enables capability to advertise both MAC and IP routes via a single MAC+IP RT-2 advertisement. MAC is imported into local bridge MAC table and enables L2 bridged traffic across the network overlay. IP is imported into the local ARP table in an asymmetric IRB design OR imported into the IP routing table in a symmetric IRB design, and enables routed traffic across the layer 2 network overlay. Please refer to [EVPN-INTER-SUBNET] more background on EVPN IRB forwarding modes.

To support EVPN mobility procedure, a single sequence number mobility attribute is advertised with the combined MAC+IP route. A single sequence number advertised with the combined MAC+IP route to resolve both MAC and IP reachability implicitly assumes a 1:1 fixed mapping between IP and MAC. While a fixed 1:1 mapping between IP and MAC is a common use case that could be addressed via existing MAC mobility procedure, additional IRB scenarios need to be considered, that don’t necessarily adhere to this assumption. Following IRB mobility scenarios are considered:

o VM move results in VM IP and MAC moving together

o VM move results in VM IP moving to a new MAC association

o VM move results in VM MAC moving to a new IP association

While existing MAC mobility procedure can be leveraged for MAC+IP move in the first scenario, subsequent scenarios result in a new MAC-IP association. As a result, a single sequence number assigned independently per-[MAC, IP] is not sufficient to determine most recent reachability for both MAC and IP, unless the sequence number assignment algorithm is designed to allow for changing MAC-IP bindings across moves.

Purpose of this draft is to define additional sequence number assignment and handling procedures to adequately address generic mobility support across EVPN-IRB overlay use cases that allow MAC-IP bindings to change across VM moves and can support mobility for both MAC and IP components carried in an EVPN RT-2 for these use cases.

In addition, for hosts on an ESI multi-homed to multiple GW devices, additional procedure is proposed to ensure synchronized sequence number assignments across the multi-homing devices.

Content presented in this draft is independent of data plane encapsulation used in the overlay being MPLS or NVO Tunnels. It is also largely independent of the EVPN IRB solution being based on symmetric OR asymmetric IRB design as defined in [EVPN-INTER-SUBNET]. In addition to symmetric and asymmetric IRB, mobility solution for a routed overlay, where traffic to an end host in the overlay is always IP routed using EVPN RT-5 is also presented in section 8.

To summarize, this draft covers mobility mobility for the following independent of the overlay encapsulation being MPLS or an NVO Tunnel:

o Symmetric EVPN IRB overlay

o Asymmetric EVPN IRB overlay

o Routed EVPN overlay

1.1 Terminology

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

o ARP is widely referred to in this document. This is simply for ease of reading, and as such, these references are equally applicable to ND (neighbor discovery) as well.

o GW: used widely in the document refers to an IRB GW that is doing routing and bridging between an access network and an EVPN enabled overlay network.

o RT-2: EVPN route type 2 carrying both MAC and IP reachability

o RT-5: EVPN route type 5 carrying IP prefix reachability

o ES: EVPN Ethernet Segment

o MAC-IP: IP association for a MAC, referred to in this document may be IPv4, IPv6 or both.

2. Optional MAC only RT-2

In an EVPN IRB scenario, where a single MAC+IP RT-2 advertisement carries both IP and MAC routes, a MAC only RT-2 advertisement is redundant for host MACs that are advertised via MAC+IP RT-2. As a result, a MAC only RT-2 is an optional route that may not be advertised from or received at an IRB GW. This is an important consideration for mobility scenarios discussed in subsequent sections.

MAC only RT-2 may still be advertised for non-IP host MACs that are not advertised via MAC+IP RT-2.

3. Mobility Use Cases

This section describes the IRB mobility use cases considered in this document. Procedures to address them are covered later in section 6 and section 7.

- o VM move results in VM IP and MAC moving together
- o VM move results in VM IP moving to a new MAC association
- o VM move results in VM MAC moving to a new IP association

3.1 VM MAC+IP Move

This is the baseline case, wherein a VM move results in both VM MAC and IP moving together with no change in MAC-IP binding across a move. Existing MAC mobility defined in RFC 7432 may be leveraged to apply to corresponding MAC+IP route to support this mobility scenario.

3.2 VM IP Move to new MAC

This is the case, where a VM move results in VM IP moving to a new MAC binding.

3.2.1 VM Reload

A VM reload or an orchestrated VM move that results in VM being re-spawned at a new location may result in VM getting a new MAC assignment, while maintaining existing IP address. This results in a VM IP move to a new MAC binding:

IP-a, MAC-a ---> IP-a, MAC-b

3.2.2 MAC Sharing

This takes into account scenarios, where multiple hosts, each with a unique IP, may share a common MAC binding, and a host move results in a new MAC binding for the host IP.

As an example, host VMs running on a single physical server, each with a unique IP, may share the same physical server MAC. In yet another scenario, an L2 access network may be behind a firewall, such that all hosts IPs on the access network are learnt with a common firewall MAC. In all such "shared MAC" use cases, multiple local MAC-

IP ARP entries may be learnt with the same MAC. A VM IP move, in such scenarios (for e.g., to a new physical server), could result in new MAC association for the VM IP.

3.2.3 Problem

In both of the above scenarios, a combined MAC+IP EVPN RT-2 advertised with a single sequence number attribute implicitly assumes a fixed IP to MAC mapping. A host IP move to a new MAC breaks this assumption and results in a new MAC+IP route. If this new MAC+IP route is independently assigned a new sequence number, the sequence number can no longer be used to determine most recent host IP reachability in a symmetric EVPN-IRB design OR the most recent IP to MAC binding in an asymmetric EVPN-IRB design.

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+

   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
   | GW1 |   | GW2 |      | GW3 |   | GW4 |      | GW5 |   | GW6 |
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
      \         /            \         /            \         /
       \ ESI-1 /              \ ESI-2 /              \ ESI-3 /
        \     /               \     /               \     /
        +\---/+                +\---/+                +\---/+
        | \ / |                | \ / |                | \ / |
        +--+--+                +--+--+                +--+--+
           |                      |                      |
      Server-MAC1            Server-MAC2            Server-MAC3
           |                      |                      |
   [VM-IP1, VM-IP2]       [VM-IP3, VM-IP4]       [VM-IP5, VM-IP6]
```

Figure 1

As an example, consider a topology shown in Figure 1, with host VMs sharing the physical server MAC. In steady state, [IP1, MAC1] route is learnt at [GW1, GW2] and advertised to remote GWs with a sequence number N. Now, VM-IP1 is moved to Server-MAC2. ARP or ND based local learning at [GW3, GW4] would now result in a new [IP1, MAC2] route being learnt. If route [IP1, MAC2] is learnt as a new MAC+IP route and assigned a new sequence number of say 0, mobility procedure for VM-IP1 will not trigger across the overlay network.

A clear sequence number assignment procedure needs to be defined to unambiguously determine the most recent IP reachability, IP to MAC binding, and MAC reachability for such a MAC sharing scenario.

3.3 VM MAC move to new IP

This is a scenario where host move or re-provisioning behind a new gateway location may result in the same VM MAC getting a new IP address assigned.

3.3.1 Problem

Complication with this scenario is that MAC reachability could be carried via a combined MAC+IP route while a MAC only route may not be advertised at all. A single sequence number association with the MAC+IP route again implicitly assumes a fixed mapping between MAC and IP. A MAC move resulting in a new IP association for the host MAC breaks this assumption and results in a new MAC+IP route. If this new MAC+IP route independently assumes a new sequence number, this mobility attribute can no longer be used to determine most recent host MAC reachability as opposed to the older existing MAC reachability.

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
   | GW1 |   | GW2 |      | GW3 |   | GW4 |      | GW5 |   | GW6 |
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
      \         /            \         /            \         /
       \ ESI-1 /              \ ESI-2 /              \ ESI-3 /
        \     /                \     /                \     /
        +\---/+                +\---/+                +\---/+
        | \ / |                | \ / |                | \ / |
        +--+--+                +--+--+                +--+--+
           |                      |                      |
        Server1                Server2                Server3
           |                      |                      |
   [VM-IP1-M1, VM-IP2-M2]  [VM-IP3-M3, VM-IP4-M4] [VM-IP5-M5, VM-IP6-M6]
```

As an example, IP1-M1 is learnt locally at [GW1, GW2] and currently advertised to remote hosts with a sequence number N. Consider a scenario where a VM with MAC M1 is re-provisioned at server 2, however, as part of this re-provisioning, assigned a different IP address say IP7. [IP7, M1] is learnt as a new route at [GW3, GW4] and advertised to remote GWs with a sequence number of 0. As a result, L3 reachability to IP7 would be established across the overlay, however, MAC mobility procedure for MAC1 will not trigger as a result of this MAC-IP route advertisement. If an optional MAC only route is also advertised, sequence number associated with the MAC only route would trigger MAC mobility as per [RFC7432]. However, in the absence of an additional MAC only route advertisement, a single sequence number advertised with a combined MAC+IP route would not be sufficient to update MAC reachability across the overlay.

A MAC-IP sequence number assignment procedure needs to be defined to unambiguously determine the most recent MAC reachability in such a scenario without a MAC only route being advertised.

Further, GW1/GW2, on learning new reachability for [IP7, M1] via GW3/GW4 MUST probe and delete any local IPs associated with MAC M1, such as [IP1, M1] in the above example.

Arguably, MAC mobility sequence number defined in [RFC7432], could be interpreted to apply only to the MAC part of MAC-IP route, and would hence cover this scenario. It could hence be interpreted as a clarification to [RFC7432] and one of the considerations for a common sequence number assignment procedure across all MAC-IP mobility scenarios detailed in this document.

4. EVPN All Active multi-homed ES

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+

   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
   | GW1 |   | GW2 |      | GW3 |   | GW4 |      | GW5 |   | GW6 |
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
      \         /            \         /            \         /
       \ ESI-1 /             \ ESI-2 /             \ ESI-3 /
        \     /               \     /               \     /
        +\---/+                +\---/+                +\---/+
        | \ / |                | \ / |                | \ / |
        +--+--+                +--+--+                +--+--+
           |                      |                      |
        Server-1               Server-2               Server-3
```

Figure 2

Consider an EVPN-IRB overlay network shown in Figure 2, with hosts multi-homed to two or more leaf GW devices via an all-active multi-homed ES. MAC and ARP entries learnt on a local ESI may also be synchronized across the multi-homing GW devices sharing this ESI. This MAC and ARP SYNC enables local switching of intra and inter subnet ECMP traffic flows from remote hosts. In other words, local MAC and ARP entries on a given Ethernet segment (ES) may be learnt via local learning and / or sync from another GW device sharing the same ES.

For a host that is multi-homed to multiple GW devices via an all-active ES interface, local learning of host MAC and MAC-IP at each GW device is an independent asynchronous event, that is dependent on traffic flow and or ARP / ND response from the host hashing to a directly connected GW on the MC-LAG interface. As a result, sequence number mobility attribute value assigned to a locally learnt MAC or MAC-IP route (as per RFC 7432) at each device may not always be the same, depending on transient states on the device at the time of local learning.

As an example, consider a host VM that is deleted from ESI-2 and moved to ESI-1. It is possible for host to be learnt on say, GW1 following deletion of the remote route from [GW3, GW4], while being learnt on GW2 prior to deletion of remote route from [GW3, GW4]. If so, GW1 would process local host route learning as a new route and assign a sequence number of 0, while GW2 would process local host route learning as a remote to local move and assign a sequence number of N+1, N being the existing sequence number assigned at [GW3, GW4]. Inconsistent sequence numbers advertised from multi-homing devices introduces ambiguity with respect to sequence number based mobility procedures across the overlay.

- o Ambiguity with respect to how the remote ToRs should handle paths with same ESI and different sequence numbers. A remote ToR may not program ECMP paths if it receives routes with different sequence numbers from a set of multi-homing GWs sharing the same ESI.

- o Breaks consistent route versioning across the network overlay that is needed for EVPN mobility procedures to work.

As an example, in this inconsistent state, GW2 would drop a remote route received for the same host with sequence number N (as its local sequence number is N+1), while GW1 would install it as the best route (as its local sequence number is 0).

There is need for a mechanism to ensure consistency of sequence numbers advertised from a set of multi-homing devices for EVPN mobility to work reliably.

In order to support mobility for multi-homed hosts using the sequence number mobility attribute, local MAC and MAC-IP routes MUST be advertised with the same sequence number by all GW devices that the ESI is multi-homed to. In other words, there is need for a mechanism to ensure consistency of sequence numbers advertised from a set of multi-homing devices for EVPN mobility to work reliably.

5. Design Considerations

To summarize, sequence number assignment scheme and implementation must take following considerations into account:

- o MAC+IP may be learnt on an ESI multi-homed to multiple GW devices, hence requires sequence numbers to be synchronized across multi-homing GW devices.

- o MAC only RT-2 is optional in an IRB scenario and may not necessarily be advertised in addition to MAC+IP RT-2

- o Single MAC may be associated with multiple IPs, i.e., multiple host IPs may share a common MAC

- o Host IP move could result in host moving to a new MAC, resulting in a new IP to MAC association and a new MAC+IP route.

- o Host MAC move to a new location could result in host MAC being associated with a different IP address, resulting in a new MAC to IP association and a new MAC+IP route

- o LOCAL MAC-IP learn via ARP would always accompanied by a LOCAL MAC learn event resulting from the ARP packet. MAC and MAC-IP learning, however, could happen in any order

- o Use cases discussed earlier that do not maintain a constant 1:1 MAC-IP mapping across moves could potentially be addressed by using separate sequence numbers associated with MAC and IP components of MAC+IP route. Maintaining two separate sequence numbers however adds significant overhead with respect to complexity, debugability, and backward compatibility. It is therefore goal of solution presented here to address these requirements via a single sequence number attribute.

6. Solution Components

This section goes over main components of the EVPN IRB mobility solution proposed in this draft. Later sections will go over exact sequence number assignment procedures resulting from concepts described in this section.

6.1 Sequence Number Inheritance

Main idea presented here is to view a LOCAL MAC-IP route as a child of the corresponding LOCAL MAC only route that inherits the sequence number attribute from the parent LOCAL MAC only route:

```
  Mx-IPx -----> Mx (seq# = N)
```

As a result, both parent MAC and child MAC-IP routes share one common sequence number associated with the parent MAC route. Doing so ensures that a single sequence number attribute carried in a combined MAC+IP route represents sequence number for both a MAC only route as well as a MAC+IP route, and hence makes the MAC only route truly optional. As a result, optional MAC only route with its own sequence number is not required to establish most recent reachability for a MAC in the overlay network. Specifically, this enables a MAC to assume a different IP address on a move, and still be able to establish most recent reachability to the MAC across the overlay network via mobility attribute associated with the MAC+IP route advertisement. As an example, when Mx moves to a new location, it would result in LOCAL Mx being assigned a higher sequence number at its new location as per RFC 7432. If this move results in Mx assuming a different IP address, IPz, LOCAL Mx+IPz route would inherit the new sequence number from Mx.

LOCAL MAC and LOCAL MAC-IP routes would typically be sourced from data plane learning and ARP learning respectively, and could get learnt in control plane in any order. Implementation could either replicate inherited sequence number in each MAC-IP entry OR maintain a single attribute in the parent MAC by creating a forward reference LOCAL MAC object for cases where a LOCAL MAC-IP is learnt before the LOCAL MAC.

Arguably, this inheritance may be assumed from RFC 7432, in which case, the above may be interpreted as a clarification with respect to interpretation of a MAC sequence number in a MAC-IP route.

6.2 MAC Sharing

Further, for the shared MAC scenario, this would result in multiple LOCAL MAC-IP siblings inheriting sequence number attribute from a common parent MAC route:

```
  Mx-IP1 -----
   |         |
  Mx-IP2 -----
    .        |
    .        +---> Mx (seq# = N)
    .        |
  Mx-IPw -----
    |        |
  Mx-IPx -----
```

In such a case, a host-IP move to a different physical server would result in IP moving to a new MAC binding. A new MAC-IP route resulting from this move must now be advertised with a sequence number that is higher than the previous MAC-IP route for this IP, advertised from the prior location. As an example, consider a route Mx-IPx that is currently advertised with sequence number N from GW1. IPx moving to a new physical server behind GW2 results in IPx being associated with MAC Mz. A new local Mz-IPx route resulting from this move at GW2 must now be advertised with a sequence number higher than N. This is so that GW devices, including GW1, GW2, and other remote GW devices that are part of the overlay can clearly determine and program the most recent MAC binding and reachability for the IP. GW1, on receiving this new Mz-IPx route with sequence number say, N+1, for symmetric IRB case, would update IPx reachability via GW2 in forwarding, for asymmetric IRB case, would update IPx’s ARP binding to Mz. In addition, GW1 would clear and withdraw the stale Mx-IPx route with the lower sequence number.

This also implies that sequence number associated with local MAC Mz and all local MAC-IP children of Mz at GW2 must now be incremented to N+1, and re-advertised across the overlay. While this re-advertisement of all local MAC-IP children routes affected by the parent MAC route is an overhead, it avoids the need for two separate sequence number attributes to be maintained and advertised for IP and MAC components of MAC+IP RT-2. Implementation would need to be able to lookup MAC-IP routes for a given IP and update sequence number for it’s parent MAC and its MAC-IP children.

6.3 Multi-homing Mobility Synchronization

In order to support mobility for multi-homed hosts, local MAC and MAC-IP routes learnt on the shared ESI MUST be advertised with the same sequence number by all GW devices that the ESI is multi-homed to. This also applies to local MAC only routes. LOCAL MAC and MAC-IP may be learnt natively via data plane and ARP/ND respectively as well as via SYNC from another multi-homing GW to achieve local switching. Local and SYNC route learning can happen in any order. Local MAC-IP routes advertised by all multi-homing GW devices sharing the ESI must carry the same sequence number, independent of the order in which they are learnt. This implies:

- On local or sync MAC-IP route learning, sequence number for the local MAC-IP route MUST be compared and updated to the higher value.

- On local or sync MAC route learning, sequence number for the local MAC route MUST be compared and updated to the higher value.

If an update to local MAC-IP sequence number is required as a result of above comparison with sync MAC-IP route, it would essentially amount to a sequence number update on the parent local MAC, resulting in the inherited sequence number update on the MAC-IP route.

7. Requirements for Sequence Number Assignment

Following sections summarize sequence number assignment procedure needed on local and sync MAC and MAC-IP route learning events in order to accomplish the above.

7.1 LOCAL MAC-IP learning

A local Mx-IPx learning via ARP or ND should result in computation OR re-computation of parent MAC Mx’s sequence number, following which the MAC-IP route Mx-IPx would simply inherit parent MAC’s sequence number. Parent MAC Mx Sequence number should be computed as follows:

- o MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.

- o MUST be at least equal to corresponding SYNC MAC sequence number if one is present.

- o If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number

Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

7.2 LOCAL MAC learning

Local MAC Mx Sequence number should be computed as follows:

- o MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.

- o MUST be at least equal to corresponding SYNC MAC sequence number if one is present.

- o Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

Note that the local MAC sequence number might already be present if there was a local MAC-IP learnt prior to the local MAC, in which case the above may not result in any change in local MAC’s sequence number.

7.3 Remote MAC OR MAC-IP Update

On receiving a remote MAC OR MAC-IP route update associated with a MAC Mx with a sequence number that is higher than a LOCAL route for MAC Mx:

- o GW MUST trigger probe and deletion procedure for all LOCAL IPs associated with MAC Mx

- o GW MUST trigger deletion procedure for LOCAL MAC route for Mx

7.4 REMOTE (SYNC) MAC update

Corresponding local MAC Mx (if present) Sequence number should be re-computed as follows:

- o If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.

- o If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

7.5 REMOTE (SYNC) MAC-IP update

If this is a SYNCed MAC-IP on a local ESI, it would also result in a derived SYNC MAC Mx route entry, as MAC only RT-2 advertisement is optional. Corresponding local MAC Mx (if present) Sequence number should be re-computed as follows:

- o If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.

- o If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

7.6 Inter-op

In general, if all GW nodes in the overlay network follow the above sequence number assignment procedure, and the GW is advertising both MAC+IP and MAC routes, sequence number advertised with the MAC and MAC+IP routes with the same MAC would always be the same. However, an inter-op scenario with a different implementation could arise, where a GW implementation non-compliant with this document or with RFC 7432 assigns and advertises independent sequence numbers to MAC and MAC+IP routes. To handle this case, if different sequence numbers are received for remote MAC+IP and corresponding remote MAC routes from a remote GW, sequence number associated with the remote MAC route should be computed as:

- o Highest of the all received sequence numbers with remote MAC+IP and MAC routes with the same MAC.

- o MAC sequence number would be re-computed on a MAC or MAC+IP route withdraw as per above.

A MAC and / or IP move to the local GW would now result in the MAC (and hence all MAC-IP) sequence numbers incremented from the above computed remote MAC sequence number.

8. Routed Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

- o A host MAC is never advertised in EVPN overlay control plane
- o Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI
- o An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged
- o Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC 7814] for more details.

Host mobility within the stretched subnet would still need to be supported for this use. In the absence of any host MAC routes, sequence number mobility EXT-COMM specified in [RFC7432], section 7.7 may be associated with a /32 OR /128 host IP prefix advertised via EVPN route type 5. MAC mobility procedures defined in RFC 7432 can now be applied as is to host IP prefixes:

- o On LOCAL learning of a host IP, on a new ESI, host IP MUST be advertised with a sequence number attribute that is higher than what is currently advertised with the old ESI
- o on receiving a host IP route advertisement with a higher sequence number, a PE MUST trigger ARP/ND probe and deletion procedure on any LOCAL route for that IP with a lower sequence number. A PE would essentially move the forwarding entry to point to the remote route with a higher sequence number and send an ARP/ND PROBE for the local IP route. If the IP has indeed moved, PROBE would timeout and the local IP host route would be deleted.

Note that there is still only one sequence number associated with a host route at any time. For earlier use cases where a host MAC is advertised along with the host IP, a sequence number is only associated with a MAC. Only if the MAC is not advertised at all, as in this use case, is a sequence number associated with a host IP.

Note that this mobility procedure would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

9. Duplicate Host Detection

Duplicate host detection scenarios across EVPN IRB can be classified as follows:

- o Scenario A: where two hosts have the same MAC (host IPs may or may not be duplicate)

- o Scenario B: where two hosts have the same IP but different MACs

- o Scenario C: where two hosts have the same IP and host MAC is not advertised at all

Duplicate detection procedures for scenario B and C would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

9.1 Scenario A

For all use cases where duplicate hosts have the same MAC, MAC is detected as duplicate via duplicate MAC detection procedure described in RFC 7432. Corresponding MAC-IP routes with the same MAC do not require duplicate detection and MUST simply inherit the DUPLICATE property from the corresponding MAC route. In other words, if a MAC route is in DUPLICATE state, all corresponding MAC-IP routes MUST also be treated as DUPLICATE. Duplicate detection procedure need only be applied to MAC routes.

9.2 Scenario B

Due to misconfiguration, a situation may arise where hosts with different MACs are configured with the same IP. This scenario would not be detected by existing duplicate MAC detection procedure and would result in incorrect forwarding of routed traffic destined to this IP.

Such a situation, on LOCAL MAC-IP learning, would be detected as a move scenario via the following local MAC sequence number computation procedure described earlier in section 5.1:

- o If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number

Such a move that results in sequence number increment on local MAC because of a remote MAC-IP route associated with a different MAC MUST be counted as an "IP move" against the "IP" independent of MAC. Duplicate detection procedure described in RFC 7432 can now be applied to an "IP" entity independent of MAC. Once an IP is detected as DUPLICATE, corresponding MAC-IP route should be treated as DUPLICATE. Associated MAC routes and any other MAC-IP routes associated with this MAC should not be affected.

9.2.1 Duplicate IP Detection Procedure for Scenario B

Duplicate IP detection procedure for such a scenario is specified in [EVPN-PROXY-ARP]. What counts as an "IP move" in this scenario is further clarified as follows:

- o On learning a LOCAL MAC-IP route Mx-IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different MAC association, say, Mz-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC

- o On learning a REMOTE MAC-IP route Mz-IPx, check if there is an existing LOCAL route for IPx with a different MAC association, say, Mx-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC

A MAC-IP route SHOULD be treated as DUPLICATE if either of the following two conditions are met:

- o Corresponding MAC route is marked as DUPLICATE via existing duplicate detection procedure

- o Corresponding IP is marked as DUPLICATE via extended procedure described above

9.3 Scenario C

For a purely routed overlay scenario described in section 8, where only a host IP is advertised via EVPN RT-5, together with a sequence number mobility attribute, duplicate MAC detection procedures specified in RFC 7432 can be intuitively applied to IP only host routes for the purpose of duplicate IP detection.

- o On learning a LOCAL host IP route IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx.

- o On learning a REMOTE host IP route IPx, check if there is an existing LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx

- o With configurable parameters "N" and "M", If "N" IP moves are detected within "M" seconds for IPx, treat IPx as DUPLICATE

9.4 Duplicate Host Recovery

Once a MAC or IP is marked as DUPLICATE and FROZEN, corrective action must be taken to un-provision one of the duplicate MAC or IP. Un-provisioning a duplicate MAC or IP in this context refers to a corrective action taken on the host side. Once one of the duplicate MAC or IP is un-provisioned, normal operation would not resume until the duplicate MAC or IP ages out, following this correction, unless additional action is taken to speed up recovery.

This section lists possible additional corrective actions that could be taken to achieve faster recovery to normal operation.

9.4.1 Route Un-freezing Configuration

Unfreezing the DUPLICATE OR FROZEN MAC or IP via a CLI can be leveraged to recover from DUPLICATE and FROZEN state following corrective un-provisioning of the duplicate MAC or IP.

Unfreezing the frozen MAC or IP via a CLI at a GW should result in that MAC OR IP being advertised with a sequence number that is higher than the sequence number advertised from the other location of that MAC or IP.

Two possible corrective un-provisioning scenarios exist:

- Scenario A: A duplicate MAC or IP may have been un-provisioned at the location where it was NOT marked as DUPLICATE and FROZEN

- Scenario B: A duplicate MAC or IP may have been un-provisioned at the location where it was marked as DUPLICATE and FROZEN

Unfreezing the DUPLICATE and FROZEN MAC or IP, following the above corrective un-provisioning scenarios would result in recovery to steady state as follows:

- Scenario A: If the duplicate MAC or IP was un-provisioned at the location where it was NOT marked as DUPLICATE, unfreezing the route at the FROZEN location will result in the route being advertised with a higher sequence number. This would in-turn result in automatic clearing of local route at the GW location, where the host was un-provisioned via ARP/ND PROBE and DELETE procedure specified earlier in section 8 and in [RFC 7432].

- Scenario B: If the duplicate host is un-provisioned at the location where it was marked as DUPLICATE, unfreezing the route will trigger an advertisement with a higher sequence number to the other location. This would in-turn trigger re-learning of local route at the remote location, resulting in another advertisement with a higher sequence number from the remote location. Route at the local location would now be cleared on receiving this remote route advertisement, following the ARP/ND PROBE.

9.4.2 Route Clearing Configuration

In addition to the above, route clearing CLIs may also be leveraged to clear the local MAC or IP route, to be executed AFTER the duplicate host is un-provisioned:

- o clear mac CLI: A clear MAC CLI can be leveraged to clear a DUPLICATE MAC route, to recover from a duplicate MAC scenario

- o clear ARP/ND: A clear ARP/ND CLI may be leveraged to clear a DUPLICATE IP route to recover from a duplicate IP scenario

Note that the route unfreeze CLI may still need to be run if the route was un-provisioned and cleared from the NON-DUPLICATE / NON-FROZEN location. Given that unfreezing of the route via the un-freeze CLI would any ways result in auto-clearing of the route from the "un-provisioned" location, as explained in the prior section, need for a route clearing CLI for recovery from DUPLICATE / FROZEN state is truly optional.

10. Security Considerations

11. IANA Considerations

12. References

12.1 Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[EVPN-PROXY-ARP] Rabadan et al., "Operational Aspects of Proxy-ARP/ND in EVPN Networks", draft-ietf-bess-evpn-proxy-arp-nd-02, work in progress, April 2017, <https://tools.ietf.org/html/draft-ietf-bess-evpn-proxy-arp-nd-02>.

[EVPN-INTER-SUBNET] Sajassi et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-03, work in progress, Feb 2017, <https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-03>.

[RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., Fee, B., "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension Solution", RFC 7814, March 2016, <https://tools.ietf.org/html/rfc7814>.

12.2 Informative References

13. Acknowledgements

Authors would like to thank Vibov Bhan and Patrice Brisset for feedback and comments through the process.

Authors’ Addresses

Neeraj Malhotra (Editor)
Cisco
EMail: nmalhotr@cisco.com

Ali Sajassi
Cisco
EMail: sajassi@cisco.com

Aparna Pattekar
Cisco
Email: apjoshi@cisco.com

Avinash Lingala
AT&T
Email: ar977m@att.com

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

John Drake
Juniper Networks
EMail: jdrake@juniper.net

Appendix A

An alternative approach considered was to associate two independent sequence number attributes with MAC and IP components of a MAC-IP route. However, the approach of enabling IRB mobility procedures using a single sequence number associated with a MAC, as specified in this document was preferred for the following reasons:

- o Procedural overhead and complexity associated with maintaining two separate sequence numbers all the time, only to address scenarios with changing MAC-IP bindings is a big overhead for topologies where MAC-IP bindings never change.

- o Using a single sequence number associated with MAC is much simpler and adds no overhead for topologies where MAC-IP bindings never change.

- o Using a single sequence number associated with MAC is aligned with existing MAC mobility implementations. On other words, it is an easier implementation extension to existing MAC mobility procedure.

INTERNET-DRAFT   N. Malhotra, Ed.
                                                           (Arrcus)
                                                         A. Sajassi
                                                        A. Pattekar
Intended Status: Proposed Standard   (Cisco)
                                                         A. Lingala
                                                             (AT&T)
                                                         J. Rabadan
                                                            (Nokia)
                                                           J. Drake
                                                (Juniper Networks)

Expires: Jul 19, 2019   Jan 15, 2019

Extended Mobility Procedures for EVPN-IRB
draft-malhotra-bess-evpn-irb-extended-mobility-04

Abstract

The procedure to handle host mobility in a layer 2 Network with EVPN control plane is defined as part of RFC 7432. EVPN has since evolved to find wider applicability across various IRB use cases that include distributing both MAC and IP reachability via a common EVPN control plane. MAC Mobility procedures defined in RFC 7432 are extensible to IRB use cases if a fixed 1:1 mapping between VM IP and MAC is assumed across VM moves. Generic mobility support for IP and MAC that allows these bindings to change across moves is required to support a broader set of EVPN IRB use cases, and requires further consideration. EVPN all-active multi-homing further introduces scenarios that require additional consideration from mobility perspective. Intent of this draft is to enumerate a set of design considerations applicable to mobility across EVPN IRB use cases and define generic sequence number assignment procedures to address these IRB use cases.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

EVPN-IRB enables capability to advertise both MAC and IP routes via a single MAC+IP RT-2 advertisement. MAC is imported into local bridge MAC table and enables L2 bridged traffic across the network overlay. IP is imported into the local ARP table in an asymmetric IRB design OR imported into the IP routing table in a symmetric IRB design, and enables routed traffic across the layer 2 network overlay. Please refer to [EVPN-INTER-SUBNET] more background on EVPN IRB forwarding modes.

To support EVPN mobility procedure, a single sequence number mobility attribute is advertised with the combined MAC+IP route. A single sequence number advertised with the combined MAC+IP route to resolve both MAC and IP reachability implicitly assumes a 1:1 fixed mapping between IP and MAC. While a fixed 1:1 mapping between IP and MAC is a common use case that could be addressed via existing MAC mobility procedure, additional IRB scenarios need to be considered, that don’t necessarily adhere to this assumption. Following IRB mobility scenarios are considered:

o VM move results in VM IP and MAC moving together

o VM move results in VM IP moving to a new MAC association

o VM move results in VM MAC moving to a new IP association

While existing MAC mobility procedure can be leveraged for MAC+IP move in the first scenario, subsequent scenarios result in a new MAC-IP association. As a result, a single sequence number assigned independently per-[MAC, IP] is not sufficient to determine most recent reachability for both MAC and IP, unless the sequence number assignment algorithm is designed to allow for changing MAC-IP bindings across moves.

Purpose of this draft is to define additional sequence number assignment and handling procedures to adequately address generic mobility support across EVPN-IRB overlay use cases that allow MAC-IP bindings to change across VM moves and can support mobility for both MAC and IP components carried in an EVPN RT-2 for these use cases.

In addition, for hosts on an ESI multi-homed to multiple GW devices, additional procedure is proposed to ensure synchronized sequence number assignments across the multi-homing devices.

Content presented in this draft is independent of data plane encapsulation used in the overlay being MPLS or NVO Tunnels. It is also largely independent of the EVPN IRB solution being based on symmetric OR asymmetric IRB design as defined in [EVPN-INTER-SUBNET]. In addition to symmetric and asymmetric IRB, mobility solution for a routed overlay, where traffic to an end host in the overlay is always IP routed using EVPN RT-5 is also presented in section 8.

To summarize, this draft covers mobility mobility for the following independent of the overlay encapsulation being MPLS or an NVO Tunnel:

o Symmetric EVPN IRB overlay

o Asymmetric EVPN IRB overlay

o Routed EVPN overlay

1.1 Terminology

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

o ARP is widely referred to in this document. This is simply for ease of reading, and as such, these references are equally applicable to ND (neighbor discovery) as well.

o GW: used widely in the document refers to an IRB GW that is doing routing and bridging between an access network and an EVPN enabled overlay network.

o RT-2: EVPN route type 2 carrying both MAC and IP reachability

o RT-5: EVPN route type 5 carrying IP prefix reachability

o ES: EVPN Ethernet Segment

o MAC-IP: IP association for a MAC, referred to in this document may be IPv4, IPv6 or both.

2. Optional MAC only RT-2

In an EVPN IRB scenario, where a single MAC+IP RT-2 advertisement carries both IP and MAC routes, a MAC only RT-2 advertisement is redundant for host MACs that are advertised via MAC+IP RT-2. As a result, a MAC only RT-2 is an optional route that may not be advertised from or received at an IRB GW. This is an important consideration for mobility scenarios discussed in subsequent sections.

MAC only RT-2 may still be advertised for non-IP host MACs that are not advertised via MAC+IP RT-2.

3. Mobility Use Cases

This section describes the IRB mobility use cases considered in this document. Procedures to address them are covered later in section 6 and section 7.

- VM move results in VM IP and MAC moving together

- VM move results in VM IP moving to a new MAC association

- VM move results in VM MAC moving to a new IP association

3.1 VM MAC+IP Move

This is the baseline case, wherein a VM move results in both VM MAC and IP moving together with no change in MAC-IP binding across a move. Existing MAC mobility defined in RFC 7432 may be leveraged to apply to corresponding MAC+IP route to support this mobility scenario.

3.2 VM IP Move to new MAC

This is the case, where a VM move results in VM IP moving to a new MAC binding.

3.2.1 VM Reload

A VM reload or an orchestrated VM move that results in VM being re-spawned at a new location may result in VM getting a new MAC assignment, while maintaining existing IP address. This results in a VM IP move to a new MAC binding:

IP-a, MAC-a ---> IP-a, MAC-b

3.2.2 MAC Sharing

This takes into account scenarios, where multiple hosts, each with a unique IP, may share a common MAC binding, and a host move results in a new MAC binding for the host IP.

As an example, host VMs running on a single physical server, each with a unique IP, may share the same physical server MAC. In yet another scenario, an L2 access network may be behind a firewall, such that all hosts IPs on the access network are learnt with a common firewall MAC. In all such "shared MAC" use cases, multiple local MAC-

IP ARP entries may be learnt with the same MAC. A VM IP move, in such scenarios (for e.g., to a new physical server), could result in new MAC association for the VM IP.

3.2.3 Problem

In both of the above scenarios, a combined MAC+IP EVPN RT-2 advertised with a single sequence number attribute implicitly assumes a fixed IP to MAC mapping. A host IP move to a new MAC breaks this assumption and results in a new MAC+IP route. If this new MAC+IP route is independently assigned a new sequence number, the sequence number can no longer be used to determine most recent host IP reachability in a symmetric EVPN-IRB design OR the most recent IP to MAC binding in an asymmetric EVPN-IRB design.

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+

   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
   | GW1 |   | GW2 |      | GW3 |   | GW4 |      | GW5 |   | GW6 |
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
      \         /            \         /            \         /
       \ ESI-1 /              \ ESI-2 /              \ ESI-3 /
        \     /                \     /                \     /
        +\---/+                +\---/+                +\---/+
        | \ / |                | \ / |                | \ / |
        +--+--+                +--+--+                +--+--+
           |                      |                      |
      Server-MAC1            Server-MAC2            Server-MAC3
           |                      |                      |
   [VM-IP1, VM-IP2]       [VM-IP3, VM-IP4]       [VM-IP5, VM-IP6]
```

Figure 1

As an example, consider a topology shown in Figure 1, with host VMs sharing the physical server MAC. In steady state, [IP1, MAC1] route is learnt at [GW1, GW2] and advertised to remote GWs with a sequence number N. Now, VM-IP1 is moved to Server-MAC2. ARP or ND based local learning at [GW3, GW4] would now result in a new [IP1, MAC2] route being learnt. If route [IP1, MAC2] is learnt as a new MAC+IP route and assigned a new sequence number of say 0, mobility procedure for VM-IP1 will not trigger across the overlay network.

A clear sequence number assignment procedure needs to be defined to unambiguously determine the most recent IP reachability, IP to MAC binding, and MAC reachability for such a MAC sharing scenario.

3.3 VM MAC move to new IP

This is a scenario where host move or re-provisioning behind a new gateway location may result in the same VM MAC getting a new IP address assigned.

3.3.1 Problem

Complication with this scenario is that MAC reachability could be carried via a combined MAC+IP route while a MAC only route may not be advertised at all. A single sequence number association with the MAC+IP route again implicitly assumes a fixed mapping between MAC and IP. A MAC move resulting in a new IP association for the host MAC breaks this assumption and results in a new MAC+IP route. If this new MAC+IP route independently assumes a new sequence number, this mobility attribute can no longer be used to determine most recent host MAC reachability as opposed to the older existing MAC reachability.

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
   | GW1 |   | GW2 |      | GW3 |   | GW4 |      | GW5 |   | GW6 |
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
      \         /            \         /            \         /
       \ ESI-1 /              \ ESI-2 /              \ ESI-3 /
        \     /                \     /                \     /
        +\---/+                +\---/+                +\---/+
        | \ / |                | \ / |                | \ / |
        +--+--+                +--+--+                +--+--+
           |                      |                      |
        Server1                Server2                Server3
           |                      |                      |
   [VM-IP1-M1, VM-IP2-M2]  [VM-IP3-M3, VM-IP4-M4] [VM-IP5-M5, VM-IP6-M6]
```

As an example, IP1-M1 is learnt locally at [GW1, GW2] and currently advertised to remote hosts with a sequence number N. Consider a scenario where a VM with MAC M1 is re-provisioned at server 2, however, as part of this re-provisioning, assigned a different IP address say IP7. [IP7, M1] is learnt as a new route at [GW3, GW4] and advertised to remote GWs with a sequence number of 0. As a result, L3 reachability to IP7 would be established across the overlay, however, MAC mobility procedure for MAC1 will not trigger as a result of this MAC-IP route advertisement. If an optional MAC only route is also advertised, sequence number associated with the MAC only route would trigger MAC mobility as per [RFC7432]. However, in the absence of an additional MAC only route advertisement, a single sequence number advertised with a combined MAC+IP route would not be sufficient to update MAC reachability across the overlay.

A MAC-IP sequence number assignment procedure needs to be defined to unambiguously determine the most recent MAC reachability in such a scenario without a MAC only route being advertised.

Further, GW1/GW2, on learning new reachability for [IP7, M1] via GW3/GW4 MUST probe and delete any local IPs associated with MAC M1, such as [IP1, M1] in the above example.

Arguably, MAC mobility sequence number defined in [RFC7432], could be interpreted to apply only to the MAC part of MAC-IP route, and would hence cover this scenario. It could hence be interpreted as a clarification to [RFC7432] and one of the considerations for a common sequence number assignment procedure across all MAC-IP mobility scenarios detailed in this document.

4. EVPN All Active multi-homed ES

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+

   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
   | GW1 |   | GW2 |      | GW3 |   | GW4 |      | GW5 |   | GW6 |
   +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
      \         /            \         /            \         /
       \ ESI-1 /              \ ESI-2 /              \ ESI-3 /
        \     /                \     /                \     /
        +\---/+                +\---/+                +\---/+
        | \ / |                | \ / |                | \ / |
        +--+--+                +--+--+                +--+--+
           |                      |                      |
        Server-1               Server-2               Server-3
```

Figure 2

Consider an EVPN-IRB overlay network shown in Figure 2, with hosts multi-homed to two or more leaf GW devices via an all-active multi-homed ES. MAC and ARP entries learnt on a local ESI may also be synchronized across the multi-homing GW devices sharing this ESI. This MAC and ARP SYNC enables local switching of intra and inter subnet ECMP traffic flows from remote hosts. In other words, local MAC and ARP entries on a given Ethernet segment (ES) may be learnt via local learning and / or sync from another GW device sharing the same ES.

For a host that is multi-homed to multiple GW devices via an all-active ES interface, local learning of host MAC and MAC-IP at each GW device is an independent asynchronous event, that is dependent on traffic flow and or ARP / ND response from the host hashing to a directly connected GW on the MC-LAG interface. As a result, sequence number mobility attribute value assigned to a locally learnt MAC or MAC-IP route (as per RFC 7432) at each device may not always be the same, depending on transient states on the device at the time of local learning.

As an example, consider a host VM that is deleted from ESI-2 and moved to ESI-1. It is possible for host to be learnt on say, GW1 following deletion of the remote route from [GW3, GW4], while being learnt on GW2 prior to deletion of remote route from [GW3, GW4]. If so, GW1 would process local host route learning as a new route and assign a sequence number of 0, while GW2 would process local host route learning as a remote to local move and assign a sequence number of N+1, N being the existing sequence number assigned at [GW3, GW4]. Inconsistent sequence numbers advertised from multi-homing devices introduces ambiguity with respect to sequence number based mobility procedures across the overlay.

- o Ambiguity with respect to how the remote ToRs should handle paths with same ESI and different sequence numbers. A remote ToR may not program ECMP paths if it receives routes with different sequence numbers from a set of multi-homing GWs sharing the same ESI.

- o Breaks consistent route versioning across the network overlay that is needed for EVPN mobility procedures to work.

As an example, in this inconsistent state, GW2 would drop a remote route received for the same host with sequence number N (as its local sequence number is N+1), while GW1 would install it as the best route (as its local sequence number is 0).

There is need for a mechanism to ensure consistency of sequence numbers advertised from a set of multi-homing devices for EVPN mobility to work reliably.

In order to support mobility for multi-homed hosts using the sequence number mobility attribute, local MAC and MAC-IP routes MUST be advertised with the same sequence number by all GW devices that the ESI is multi-homed to. In other words, there is need for a mechanism to ensure consistency of sequence numbers advertised from a set of multi-homing devices for EVPN mobility to work reliably.

5. Design Considerations

To summarize, sequence number assignment scheme and implementation must take following considerations into account:

- o MAC+IP may be learnt on an ESI multi-homed to multiple GW devices, hence requires sequence numbers to be synchronized across multi-homing GW devices.

- o MAC only RT-2 is optional in an IRB scenario and may not necessarily be advertised in addition to MAC+IP RT-2

- o Single MAC may be associated with multiple IPs, i.e., multiple host IPs may share a common MAC

- o Host IP move could result in host moving to a new MAC, resulting in a new IP to MAC association and a new MAC+IP route.

- Host MAC move to a new location could result in host MAC being associated with a different IP address, resulting in a new MAC to IP association and a new MAC+IP route

- LOCAL MAC-IP learn via ARP would always accompanied by a LOCAL MAC learn event resulting from the ARP packet. MAC and MAC-IP learning, however, could happen in any order

- Use cases discussed earlier that do not maintain a constant 1:1 MAC-IP mapping across moves could potentially be addressed by using separate sequence numbers associated with MAC and IP components of MAC+IP route. Maintaining two separate sequence numbers however adds significant overhead with respect to complexity, debugability, and backward compatibility. It is therefore goal of solution presented here to address these requirements via a single sequence number attribute.

6. Solution Components

This section goes over main components of the EVPN IRB mobility solution proposed in this draft. Later sections will go over exact sequence number assignment procedures resulting from concepts described in this section.

6.1 Sequence Number Inheritance

Main idea presented here is to view a LOCAL MAC-IP route as a child of the corresponding LOCAL MAC only route that inherits the sequence number attribute from the parent LOCAL MAC only route:

```
Mx-IPx -----> Mx (seq# = N)
```

As a result, both parent MAC and child MAC-IP routes share one common sequence number associated with the parent MAC route. Doing so ensures that a single sequence number attribute carried in a combined MAC+IP route represents sequence number for both a MAC only route as well as a MAC+IP route, and hence makes the MAC only route truly optional. As a result, optional MAC only route with its own sequence number is not required to establish most recent reachability for a MAC in the overlay network. Specifically, this enables a MAC to assume a different IP address on a move, and still be able to establish most recent reachability to the MAC across the overlay network via mobility attribute associated with the MAC+IP route advertisement. As an example, when Mx moves to a new location, it would result in LOCAL Mx being assigned a higher sequence number at its new location as per RFC 7432. If this move results in Mx assuming a different IP address, IPz, LOCAL Mx+IPz route would inherit the new sequence number from Mx.

LOCAL MAC and LOCAL MAC-IP routes would typically be sourced from data plane learning and ARP learning respectively, and could get learnt in control plane in any order. Implementation could either replicate inherited sequence number in each MAC-IP entry OR maintain a single attribute in the parent MAC by creating a forward reference LOCAL MAC object for cases where a LOCAL MAC-IP is learnt before the LOCAL MAC.

Arguably, this inheritance may be assumed from RFC 7432, in which case, the above may be interpreted as a clarification with respect to interpretation of a MAC sequence number in a MAC-IP route.

6.2 MAC Sharing

Further, for the shared MAC scenario, this would result in multiple LOCAL MAC-IP siblings inheriting sequence number attribute from a common parent MAC route:

```
  Mx-IP1 -----
   |          |
  Mx-IP2 -----
    .         |
    .         +---> Mx (seq# = N)
    .         |
  Mx-IPw -----
    |         |
  Mx-IPx -----
```

In such a case, a host-IP move to a different physical server would result in IP moving to a new MAC binding. A new MAC-IP route resulting from this move must now be advertised with a sequence number that is higher than the previous MAC-IP route for this IP, advertised from the prior location. As an example, consider a route Mx-IPx that is currently advertised with sequence number N from GW1. IPx moving to a new physical server behind GW2 results in IPx being associated with MAC Mz. A new local Mz-IPx route resulting from this move at GW2 must now be advertised with a sequence number higher than N. This is so that GW devices, including GW1, GW2, and other remote GW devices that are part of the overlay can clearly determine and program the most recent MAC binding and reachability for the IP. GW1, on receiving this new Mz-IPx route with sequence number say, N+1, for symmetric IRB case, would update IPx reachability via GW2 in forwarding, for asymmetric IRB case, would update IPx's ARP binding to Mz. In addition, GW1 would clear and withdraw the stale Mx-IPx route with the lower sequence number.

This also implies that sequence number associated with local MAC Mz and all local MAC-IP children of Mz at GW2 must now be incremented to N+1, and re-advertised across the overlay. While this re-advertisement of all local MAC-IP children routes affected by the parent MAC route is an overhead, it avoids the need for two separate sequence number attributes to be maintained and advertised for IP and MAC components of MAC+IP RT-2. Implementation would need to be able to lookup MAC-IP routes for a given IP and update sequence number for it’s parent MAC and its MAC-IP children.

6.3 Multi-homing Mobility Synchronization

In order to support mobility for multi-homed hosts, local MAC and MAC-IP routes learnt on the shared ESI MUST be advertised with the same sequence number by all GW devices that the ESI is multi-homed to. This also applies to local MAC only routes. LOCAL MAC and MAC-IP may be learnt natively via data plane and ARP/ND respectively as well as via SYNC from another multi-homing GW to achieve local switching. Local and SYNC route learning can happen in any order. Local MAC-IP routes advertised by all multi-homing GW devices sharing the ESI must carry the same sequence number, independent of the order in which they are learnt. This implies:

- On local or sync MAC-IP route learning, sequence number for the local MAC-IP route MUST be compared and updated to the higher value.

- On local or sync MAC route learning, sequence number for the local MAC route MUST be compared and updated to the higher value.

If an update to local MAC-IP sequence number is required as a result of above comparison with sync MAC-IP route, it would essentially amount to a sequence number update on the parent local MAC, resulting in the inherited sequence number update on the MAC-IP route.

7. Requirements for Sequence Number Assignment

Following sections summarize sequence number assignment procedure needed on local and sync MAC and MAC-IP route learning events in order to accomplish the above.

7.1 LOCAL MAC-IP learning

A local Mx-IPx learning via ARP or ND should result in computation OR re-computation of parent MAC Mx’s sequence number, following which the MAC-IP route Mx-IPx would simply inherit parent MAC’s sequence number. Parent MAC Mx Sequence number should be computed as follows:

- o MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.
- o MUST be at least equal to corresponding SYNC MAC sequence number if one is present.
- o If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number

Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

7.2 LOCAL MAC learning

Local MAC Mx Sequence number should be computed as follows:

- o MUST be higher than any existing remote MAC route for Mx, as per RFC 7432.
- o MUST be at least equal to corresponding SYNC MAC sequence number if one is present.
- o Once new sequence number for MAC route Mx is computed as per above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

Note that the local MAC sequence number might already be present if there was a local MAC-IP learnt prior to the local MAC, in which case the above may not result in any change in local MAC’s sequence number.

7.3 Remote MAC OR MAC-IP Update

On receiving a remote MAC OR MAC-IP route update associated with a MAC Mx with a sequence number that is higher than a LOCAL route for MAC Mx:

- o GW MUST trigger probe and deletion procedure for all LOCAL IPs associated with MAC Mx
- o GW MUST trigger deletion procedure for LOCAL MAC route for Mx

7.4 REMOTE (SYNC) MAC update

Corresponding local MAC Mx (if present) Sequence number should be re-computed as follows:

- o If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.

- o If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

7.5 REMOTE (SYNC) MAC-IP update

If this is a SYNCed MAC-IP on a local ESI, it would also result in a derived SYNC MAC Mx route entry, as MAC only RT-2 advertisement is optional. Corresponding local MAC Mx (if present) Sequence number should be re-computed as follows:

- o If the current sequence number is less than the received SYNC MAC sequence number, it MUST be increased to be equal to received SYNC MAC sequence number.

- o If a LOCAL MAC sequence number is updated as a result of the above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated sequence number.

7.6 Inter-op

In general, if all GW nodes in the overlay network follow the above sequence number assignment procedure, and the GW is advertising both MAC+IP and MAC routes, sequence number advertised with the MAC and MAC+IP routes with the same MAC would always be the same. However, an inter-op scenario with a different implementation could arise, where a GW implementation non-compliant with this document or with RFC 7432 assigns and advertises independent sequence numbers to MAC and MAC+IP routes. To handle this case, if different sequence numbers are received for remote MAC+IP and corresponding remote MAC routes from a remote GW, sequence number associated with the remote MAC route should be computed as:

- o Highest of the all received sequence numbers with remote MAC+IP and MAC routes with the same MAC.

- o MAC sequence number would be re-computed on a MAC or MAC+IP route withdraw as per above.

A MAC and / or IP move to the local GW would now result in the MAC (and hence all MAC-IP) sequence numbers incremented from the above computed remote MAC sequence number.

8. Routed Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

- A host MAC is never advertised in EVPN overlay control plane

- Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI

- An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged

- Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC 7814] for more details.

Host mobility within the stretched subnet would still need to be supported for this use. In the absence of any host MAC routes, sequence number mobility EXT-COMM specified in [RFC7432], section 7.7 may be associated with a /32 OR /128 host IP prefix advertised via EVPN route type 5. MAC mobility procedures defined in RFC 7432 can now be applied as is to host IP prefixes:

- On LOCAL learning of a host IP, on a new ESI, host IP MUST be advertised with a sequence number attribute that is higher than what is currently advertised with the old ESI

- on receiving a host IP route advertisement with a higher sequence number, a PE MUST trigger ARP/ND probe and deletion procedure on any LOCAL route for that IP with a lower sequence number. A PE would essentially move the forwarding entry to point to the remote route with a higher sequence number and send an ARP/ND PROBE for the local IP route. If the IP has indeed moved, PROBE would timeout and the local IP host route would be deleted.

Note that there is still only one sequence number associated with a host route at any time. For earlier use cases where a host MAC is advertised along with the host IP, a sequence number is only associated with a MAC. Only if the MAC is not advertised at all, as in this use case, is a sequence number associated with a host IP.

Note that this mobility procedure would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

9. Duplicate Host Detection

Duplicate host detection scenarios across EVPN IRB can be classified as follows:

- o Scenario A: where two hosts have the same MAC (host IPs may or may not be duplicate)

- o Scenario B: where two hosts have the same IP but different MACs

- o Scenario C: where two hosts have the same IP and host MAC is not advertised at all

Duplicate detection procedures for scenario B and C would not apply to "anycast IPv6" hosts advertised via NA messages with 0-bit=0. Please refer to [EVPN-PROXY-ARP].

9.1 Scenario A

For all use cases where duplicate hosts have the same MAC, MAC is detected as duplicate via duplicate MAC detection procedure described in RFC 7432. Corresponding MAC-IP routes with the same MAC do not require duplicate detection and MUST simply inherit the DUPLICATE property from the corresponding MAC route. In other words, if a MAC route is in DUPLICATE state, all corresponding MAC-IP routes MUST also be treated as DUPLICATE. Duplicate detection procedure need only be applied to MAC routes.

9.2 Scenario B

Due to misconfiguration, a situation may arise where hosts with different MACs are configured with the same IP. This scenario would not be detected by existing duplicate MAC detection procedure and would result in incorrect forwarding of routed traffic destined to this IP.

Such a situation, on LOCAL MAC-IP learning, would be detected as a move scenario via the following local MAC sequence number computation procedure described earlier in section 5.1:

- o If the IP is also associated with a different remote MAC "Mz", MUST be higher than "Mz" sequence number

Such a move that results in sequence number increment on local MAC because of a remote MAC-IP route associated with a different MAC MUST be counted as an "IP move" against the "IP" independent of MAC. Duplicate detection procedure described in RFC 7432 can now be applied to an "IP" entity independent of MAC. Once an IP is detected as DUPLICATE, corresponding MAC-IP route should be treated as DUPLICATE. Associated MAC routes and any other MAC-IP routes associated with this MAC should not be affected.

9.2.1 Duplicate IP Detection Procedure for Scenario B

Duplicate IP detection procedure for such a scenario is specified in [EVPN-PROXY-ARP]. What counts as an "IP move" in this scenario is further clarified as follows:

- o On learning a LOCAL MAC-IP route Mx-IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different MAC association, say, Mz-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC

- o On learning a REMOTE MAC-IP route Mz-IPx, check if there is an existing LOCAL route for IPx with a different MAC association, say, Mx-IPx. If so, count this as an "IP move" count for IPx, independent of the MAC

A MAC-IP route SHOULD be treated as DUPLICATE if either of the following two conditions are met:

- o Corresponding MAC route is marked as DUPLICATE via existing duplicate detection procedure

- o Corresponding IP is marked as DUPLICATE via extended procedure described above

9.3 Scenario C

For a purely routed overlay scenario described in section 8, where only a host IP is advertised via EVPN RT-5, together with a sequence number mobility attribute, duplicate MAC detection procedures specified in RFC 7432 can be intuitively applied to IP only host routes for the purpose of duplicate IP detection.

- o On learning a LOCAL host IP route IPx, check if there is an existing REMOTE OR LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx.

- o On learning a REMOTE host IP route IPx, check if there is an existing LOCAL route for IPx with a different ESI association. If so, count this as an "IP move" count for IPx

- o With configurable parameters "N" and "M", If "N" IP moves are detected within "M" seconds for IPx, treat IPx as DUPLICATE

9.4 Duplicate Host Recovery

Once a MAC or IP is marked as DUPLICATE and FROZEN, corrective action must be taken to un-provision one of the duplicate MAC or IP. Un-provisioning a duplicate MAC or IP in this context refers to a corrective action taken on the host side. Once one of the duplicate MAC or IP is un-provisioned, normal operation would not resume until the duplicate MAC or IP ages out, following this correction, unless additional action is taken to speed up recovery.

This section lists possible additional corrective actions that could be taken to achieve faster recovery to normal operation.

9.4.1 Route Un-freezing Configuration

Unfreezing the DUPLICATE OR FROZEN MAC or IP via a CLI can be leveraged to recover from DUPLICATE and FROZEN state following corrective un-provisioning of the duplicate MAC or IP.

Unfreezing the frozen MAC or IP via a CLI at a GW should result in that MAC OR IP being advertised with a sequence number that is higher than the sequence number advertised from the other location of that MAC or IP.

Two possible corrective un-provisioning scenarios exist:

- Scenario A: A duplicate MAC or IP may have been un-provisioned at the location where it was NOT marked as DUPLICATE and FROZEN

- Scenario B: A duplicate MAC or IP may have been un-provisioned at the location where it was marked as DUPLICATE and FROZEN

Unfreezing the DUPLICATE and FROZEN MAC or IP, following the above corrective un-provisioning scenarios would result in recovery to steady state as follows:

- Scenario A: If the duplicate MAC or IP was un-provisioned at the location where it was NOT marked as DUPLICATE, unfreezing the route at the FROZEN location will result in the route being advertised with a higher sequence number. This would in-turn result in automatic clearing of local route at the GW location, where the host was un-provisioned via ARP/ND PROBE and DELETE procedure specified earlier in section 8 and in [RFC 7432].

- Scenario B: If the duplicate host is un-provisioned at the location where it was marked as DUPLICATE, unfreezing the route will trigger an advertisement with a higher sequence number to the other location. This would in-turn trigger re-learning of local route at the remote location, resulting in another advertisement with a higher sequence number from the remote location. Route at the local location would now be cleared on receiving this remote route advertisement, following the ARP/ND PROBE.

9.4.2 Route Clearing Configuration

In addition to the above, route clearing CLIs may also be leveraged to clear the local MAC or IP route, to be executed AFTER the duplicate host is un-provisioned:

- o clear mac CLI: A clear MAC CLI can be leveraged to clear a DUPLICATE MAC route, to recover from a duplicate MAC scenario

- o clear ARP/ND: A clear ARP/ND CLI may be leveraged to clear a DUPLICATE IP route to recover from a duplicate IP scenario

Note that the route unfreeze CLI may still need to be run if the route was un-provisioned and cleared from the NON-DUPLICATE / NON-FROZEN location. Given that unfreezing of the route via the un-freeze CLI would any ways result in auto-clearing of the route from the "un-provisioned" location, as explained in the prior section, need for a route clearing CLI for recovery from DUPLICATE / FROZEN state is truly optional.

10. Security Considerations

11. IANA Considerations

12. References

12.1 Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[EVPN-PROXY-ARP] Rabadan et al., "Operational Aspects of Proxy-ARP/ND in EVPN Networks", draft-ietf-bess-evpn-proxy-arp-nd-02, work in progress, April 2017, <https://tools.ietf.org/html/draft-ietf-bess-evpn-proxy-arp-nd-02>.

[EVPN-INTER-SUBNET] Sajassi et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-03, work in progress, Feb 2017, <https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-03>.

[RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., Fee, B., "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension Solution", RFC 7814, March 2016, <https://tools.ietf.org/html/rfc7814>.

12.2 Informative References

13. Acknowledgements

Authors would like to thank Vibov Bhan and Patrice Brisset for feedback and comments through the process.

Authors' Addresses

Neeraj Malhotra (Editor)
Arrcus
EMail: neeraj.ietf@gmail.com

Ali Sajassi
Cisco
EMail: sajassi@cisco.com

Aparna Pattekar
Cisco
Email: apjoshi@cisco.com

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

Avinash Lingala
AT&T
Email: ar977m@att.com

John Drake
Juniper Networks
EMail: jdrake@juniper.net

Appendix A

An alternative approach considered was to associate two independent sequence number attributes with MAC and IP components of a MAC-IP route. However, the approach of enabling IRB mobility procedures using a single sequence number associated with a MAC, as specified in this document was preferred for the following reasons:

- o Procedural overhead and complexity associated with maintaining two separate sequence numbers all the time, only to address scenarios with changing MAC-IP bindings is a big overhead for topologies where MAC-IP bindings never change.

- o Using a single sequence number associated with MAC is much simpler and adds no overhead for topologies where MAC-IP bindings never change.

- o Using a single sequence number associated with MAC is aligned with existing MAC mobility implementations. On other words, it is an easier implementation extension to existing MAC mobility procedure.

INTERNET-DRAFT

Intended Status: Proposed Standard

Expires: May 3, 2018

October 30, 2017

N. Malhotra, Ed.
S. Thoria
A. Sajassi
(Cisco)
A. Lingala
(AT&T)

Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing
draft-malhotra-bess-evpn-unequal-lb-00

Abstract

In an EVPN-IRB based network overlay, EVPN LAG enables all-active multi-homing for a host or CE device connected to two or more PEs via a LAG bundle, such that bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs. This document defines extensions to EVPN procedures to optimally handle unequal access bandwidth distribution across a set of multi-homing PEs in order to:

- o provide greater flexibility, with respect to adding or removing individual PE-CE links within the access LAG

- o handle PE-CE LAG member link failures that can result in unequal PE-CE access bandwidth across a set of multi-homing PEs

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

In an EVPN-IRB based network overlay, with access an access CE multi-homed via a LAG interface, bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs:

- o ECMP Load-balancing for bridged unicast traffic is enabled via aliasing and mass-withdraw procedures detailed in RFC 7432.

- o ECMP Load-balancing for routed unicast traffic is enabled via existing L3 ECMP mechanisms.

- o Load-sharing of bridged BUM traffic on local ports is enabled via EVPN DF election procedure detailed in RFC 7432

All of the above load-balancing and DF election procedures implicitly assume equal bandwidth distribution between the CE and the set of multi-homing PEs. Essentially, with this assumption of equal "access" bandwidth distribution across all PEs, ALL remote traffic is equally load balanced across the multi-homing PEs. This assumption of equal access bandwidth distribution can be restrictive with respect to adding / removing links in a multi-homed LAG interface and may also be easily broken on individual link failures. A solution to handle unequal access bandwidth distribution across a set of multi-homing EVPN PEs is proposed in this document. Primary motivation behind this proposal is to enable greater flexibility with respect to adding / removing member PE-CE links, as needed and optimally handle PE-CE link failures.

1.1 PE CE Link Provisioning

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+

                          +-----+   +-----+
                          | PE1 |   | PE2 |
                          +-----+   +-----+
                             \         /
                              \ ESI-1 /
                               \     /
                               +\---/+
                               | \ / |
                               +--+--+
                                  |
                                 CE1
```

Figure 1

Consider a CE1 that is dual-homed to PE1 and PE2 via EVPN-LAG with single member links of equal bandwidth to each PE (aka, equal access band-width distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it MUST add a link to both PE1 and PE2 in order to maintain equal access bandwidth distribution and inter-work with EVPN ECMP load-balancing. In other words, for a dual-homed CE, total number of CE links must be provisioned in multiples of 2 (2, 4, 6, and so on). For a triple-homed CE, number of CE links must be provisioned in multiples of three (3, 6, 9, and so on). To generalize, for a CE that is multi-homed to "n" PEs, number of PE-CE physical links provisioned must be an integral multiple of "n". This is restrictive in case of dual-homing and very quickly becomes prohibitive in case of multi-homing.

Instead, a provider may wish to increase PE-CE bandwidth OR number of links in ANY link increments. As an example, for CE1 dual-homed to PE1 and PE2 in all-active mode, provider may wish to add a third link to ONLY PE1 to increase total band-width for this CE by 50%, rather than being required to increase access bandwidth by 100% by adding a link to each of the two PEs. While existing EVPN based all-active load-balancing procedures do not necessarily preclude such asymmetric access bandwidth distribution among the PEs providing redundancy, it may result in unexpected traffic loss due to congestion in the access interface towards CE. This traffic loss is due to the fact that PE1 and PE2 will continue to attract equal amount of CE1 destined traffic from remote PEs, even when PE2 only has half the bandwidth to CE1 as PE1. This may lead to congestion and traffic loss on the PE2-CE1 link. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner.

1.2 PE CE Link Failures

More importantly, unequal PE-CE bandwidth distribution described above may occur during regular operation following a link failure, even when PE-CE links were provisioned to provide equal bandwidth distribution across multi-homing PEs.

```
                      +-----------------------+
                      | Underlay Network Fabric|
                      +-----------------------+

                          +-----+   +-----+
                          | PE1 |   | PE2 |
                          +-----+   +-----+
                            \\         //
                             \\ ESI-1 //
                              \\     /X
                              +\\---//+
                              | \\ // |
                              +---+---+
                                  |
                                 CE1
```

Consider a CE1 that is multi-homed to PE1 and PE2 via a link bundle with two member links to each PE. On a PE2-CE1 physical link failure, link bundle represented by ESI-1 on PE2 stays up, however, it’s bandwidth is cut in half. With the existing ECMP procedures, both PE1 and PE2 will continue to attract equal amount of traffic from remote PEs, even when PE1 has double the bandwidth to CE1. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner to avoid unexpected congestion and traffic loss on PE2-CE1 links within the LAG.

1.3 Design Requirement

```
                           +-----------------------+
                           |Underlay Network Fabric|
                           +-----------------------+

                 +-----+   +-----+           +-----+   +-----+
                 | PE1 |   | PE2 |   .....   | PEx |   | PEn |
                 +-----+   +-----+           +-----+   +-----+
                    \       \                 //        //
                     \ L1    \ L2            // Lx     // Ln
                      \       \             //        //
                     +-\-------\-----------//--------//-+
                     |  \       \  ESI-1  //        //  |
                     +----------------------------------+
                                      |
                                      CE
```

To generalize, if total link band-width to a CE is distributed across "n" multi-homing PEs, with Lx being the number of links / bandwidth to PEx, traffic from remote PEs to this CE MUST be load-balanced unequally across [PE1, PE2, ..., PEn] such that, the proportion of unicast and BUM flows destined for CE that are serviced by PEx is:

  Lx / [L1+L2+.....+Ln]

Solution proposed below includes extensions to EVPN procedures to achieve the above.

1.1 Terminology

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

"LOCAL PE" in the context of an ESI refers to a provider edge switch OR router that physically hosts the ESI.

"REMOTE PE" in the context of an ESI refers to a provider edge switch OR router in an EVPN overlay, who’s overlay reachability to the ESI is via the LOCAL PE.

2. Solution Overview

In order to achieve weighted load balancing for overlay unicast traffic, EVPN per-ESI EAD (Route Type 1) is leveraged to signal the ESI bandwidth to remote PEs. Using per-ESI EAD route to signal the ESI bandwidth provides a mechanism to be able to react to changes in access bandwidth in a service and host independent manner. Remote PEs computing the MAC path-lists based on global and aliasing EAD routes now have the ability to computed weighted load-balancing based on the ESI access bandwidth received from each PE that the ESI is multi-homed to. If per-ESI EAD route is also leveraged for IP path-list computation, as per [EVPN-IP-ALIASING], it would also provide a method to do weighted load-balancing for IP routed traffic.

In order to achieve weighted load-balancing of overlay BUM traffic, EVPN ES route (Route Type 4) is leveraged to signal the ESI bandwidth to PEs within an ESI’s redundancy group to influence per-service DF election. PEs in an ESI redundancy group now have the ability to do per-service DF election in a manner that is proportionate to their relative ESI bandwidth.

Procedures to accomplish this are described in greater detail next.

3. Weighted Unicast Traffic Load-balancing

3.1 LOCAL PE Behavior

A PE that is part of an ESI’s redundancy group would advertise a additional "link bandwidth" EXT-COMM attribute with per-ESI EAD route (EVPN Route Type 1), that represents total band-width of PE’s physical links in an ESI. BGP link bandwidth EXT-COMM defined in [BGP-LINK-BW] would be re-used for this purpose.

3.2 REMOTE PE Behavior

A receiving PE should use per-ESI link band-width attribute received from each PE to compute a relative weight for each remote PE, per-ESI, as shown below.

if,

```
    L(x,y) : link band-width advertised by PE-x for ESI-y

    W(x,y) : normalized weight assigned to PE-x for ESI-y

    H(y)   : Highest Common Factor (HCF) of [L(1,y), L(2,y), .....,
             L(n,y)]
```

then, the normalized weight assigned to PE-x for ESI-y may be computed as follows:

$W(x,y) = L(x,y) / H(y)$

For a MAC+IP route (EVPN Route Type 2) received with ESI-y, receiving PE MUST compute MAC and IP forwarding path-list weighted by the above normalized weights.

As an example, for a CE dual-homed to PE-1, PE-2, PE-3 via 2, 1, and 1 GE physical links respectively, as part of a link bundle represented by ESI-10:

L(1, 10) = 2000 Mbps

L(2, 10) = 1000 Mbps

L(3, 10) = 1000 Mbps

H(10) = 1000

Normalized weights assigned to each PE for ESI-10 are as follows:

W(1, 10) = 2000 / 1000 = 2.

W(2, 10) = 1000 / 1000 = 1.

W(3, 10) = 1000 / 1000 = 1.

For a remote MAC+IP host route received with ESI-10, forwarding load-balancing path-list must now be computed as: [PE-1, PE-1, PE-2, PE-3] instead of [PE-1, PE-2, PE-3]. This now results in load-balancing of all traffic destined for ESI-10 across the three multi-homing PEs in proportion to ESI-10 band-width at each PE.

Above weighted path-list computation MUST only be done for an ESI, IF a link bandwidth attribute is received from ALL of the PE’s advertising reachability to that ESI via per-ESI EAD Route Type 1. In the event that link bandwidth attribute is not received from one or more PEs, forwarding path-list would be computed using regular ECMP semantics.

4. Weighted BUM Traffic Load-Sharing

Load sharing of per-service DF role, weighted by link-bandwidth is currently under discussion and needs to be reconciled with [EVPN-PREF-DF]. This will closed in the next revision of this draft.

5. Routed EVPN Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

- o A host MAC is never advertised in EVPN overlay control plane
- o Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI
- o An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged
- o Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC 7814] for more details.

Weighted multi-path procedure described in this document may be used together with procedures described in [EVPN-IP-ALIASING] for this use case. per-ES EAD route advertised with Layer 3 VRF RTs would be used to signal ES link bandwidth attribute instead of the per-ES EAD route with Layer 2 VRF RTs. All other procedures described earlier in this document would as is.

6. EVPN-IRB Multi-homing with non-EVPN routing

EVPN-LAG based multi-homing on an IRB gateway may also be deployed together with non-EVPN routing, such as global routing or an L3VPN routing control plane. Key property that differentiates this set of use cases from EVPN IRB use cases discussed earlier is that EVPN control plane is used only to enable LAG interface based multi-homing and NOT as an overlay VPN control plane. EVPN control plane in this case enables:

- o DF election via EVPN RT-4 based procedures described in [RFC7432]
- o LOCAL MAC sync across multi-homing PEs via EVPN RT-2
- o LOCAL ARP and ND sync across multi-homing PEs via EVPN RT-2

Applicability of weighted ECMP procedures proposed in this document to these set of use cases are still under discussion and will be addressed in subsequent revisions.

7. References

7.1 Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[BGP-LINK-BW] Mohapatra, P., Fernando, R., "BGP Link Bandwidth Extended Community", January 2013, <https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-06>.

[EVPN-IP-ALIASING] Sajassi, A., Badoni, G., "L3 Aliasing and Mass Withdrawal Support for EVPN", July 2017, <https://tools.ietf.org/html/draft-sajassi-bess-evpn-ip-aliasing-00>.

[EVPN-PREF-DF-ELECT] Rabadan, J., et al., "Preference-based EVPN DF Election", June 2017, <https://www.ietf.org/id/draft-ietf-bess-evpn-pref-df-00.txt>.

7.2 Informative References

8. Acknowledgements

Authors’ Addresses

Neeraj Malhotra
Cisco
Email: nmalhotr@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Avinash Lingala
AT&T
Email: ar977m@att.com

INTERNET-DRAFT

Intended Status: Proposed Standard

N. Malhotra, Ed.
Arrcus
A. Sajassi
Cisco
J. Rabadan
Nokia
J. Drake
Juniper
A. Lingala
AT&T
S. Thoria
Cisco

Expires: Jan 17, 2019 July 16, 2018

Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing
draft-malhotra-bess-evpn-unequal-lb-04

Abstract

In an EVPN-IRB based network overlay, EVPN LAG enables all-active multi-homing for a host or CE device connected to two or more PEs via a LAG bundle, such that bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs. This document defines extensions to EVPN procedures to optimally handle unequal access bandwidth distribution across a set of multi-homing PEs in order to:

- o provide greater flexibility, with respect to adding or removing individual PE-CE links within the access LAG

- o handle PE-CE LAG member link failures that can result in unequal PE-CE access bandwidth across a set of multi-homing PEs

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

In an EVPN-IRB based network overlay, with an access CE multi-homed via a LAG interface, bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs:

- ECMP Load-balancing for bridged unicast traffic is enabled via aliasing and mass-withdraw procedures detailed in RFC 7432.

- ECMP Load-balancing for routed unicast traffic is enabled via existing L3 ECMP mechanisms.

- Load-sharing of bridged BUM traffic on local ports is enabled via EVPN DF election procedure detailed in RFC 7432

All of the above load-balancing and DF election procedures implicitly assume equal bandwidth distribution between the CE and the set of multi-homing PEs. Essentially, with this assumption of equal "access" bandwidth distribution across all PEs, ALL remote traffic is equally load balanced across the multi-homing PEs. This assumption of equal access bandwidth distribution can be restrictive with respect to adding / removing links in a multi-homed LAG interface and may also be easily broken on individual link failures. A solution to handle unequal access bandwidth distribution across a set of multi-homing EVPN PEs is proposed in this document. Primary motivation behind this proposal is to enable greater flexibility with respect to adding / removing member PE-CE links, as needed and to optimally handle PE-CE link failures.

1.1 PE CE Link Provisioning

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+

                          +-----+   +-----+
                          | PE1 |   | PE2 |
                          +-----+   +-----+
                             \         /
                              \ ESI-1 /
                               \     /
                               +\---/+
                               | \ / |
                               +--+--+
                                  |
                                 CE1
```

Figure 1

Consider a CE1 that is dual-homed to PE1 and PE2 via EVPN-LAG with single member links of equal bandwidth to each PE (aka, equal access bandwidth distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it MUST add a link to both PE1 and PE2 in order to maintain equal access bandwidth distribution and inter-work with EVPN ECMP load-balancing. In other words, for a dual-homed CE, total number of CE links must be provisioned in multiples of 2 (2, 4, 6, and so on). For a triple-homed CE, number of CE links must be provisioned in multiples of three (3, 6, 9, and so on). To generalize, for a CE that is multi-homed to "n" PEs, number of PE-CE physical links provisioned must be an integral multiple of "n". This is restrictive in case of dual-homing and very quickly becomes prohibitive in case of multi-homing.

Instead, a provider may wish to increase PE-CE bandwidth OR number of links in ANY link increments. As an example, for CE1 dual-homed to PE1 and PE2 in all-active mode, provider may wish to add a third link to ONLY PE1 to increase total bandwidth for this CE by 50%, rather than being required to increase access bandwidth by 100% by adding a link to each of the two PEs. While existing EVPN based all-active load-balancing procedures do not necessarily preclude such asymmetric access bandwidth distribution among the PEs providing redundancy, it may result in unexpected traffic loss due to congestion in the access interface towards CE. This traffic loss is due to the fact that PE1 and PE2 will continue to attract equal amount of CE1 destined traffic from remote PEs, even when PE2 only has half the bandwidth to CE1 as PE1. This may lead to congestion and traffic loss on the PE2-CE1 link. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner.

1.2 PE CE Link Failures

More importantly, unequal PE-CE bandwidth distribution described above may occur during regular operation following a link failure, even when PE-CE links were provisioned to provide equal bandwidth distribution across multi-homing PEs.

```
                      +------------------------+
                      | Underlay Network Fabric|
                      +------------------------+

                          +-----+   +-----+
                          | PE1 |   | PE2 |
                          +-----+   +-----+
                            \\         //
                             \\ ESI-1 //
                              \\     /X
                              +\\---//+
                              | \\ // |
                              +---+---+
                                  |
                                 CE1
```

Consider a CE1 that is multi-homed to PE1 and PE2 via a link bundle with two member links to each PE. On a PE2-CE1 physical link failure, link bundle represented by ESI-1 on PE2 stays up, however, it’s bandwidth is cut in half. With the existing ECMP procedures, both PE1 and PE2 will continue to attract equal amount of traffic from remote PEs, even when PE1 has double the bandwidth to CE1. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner to avoid unexpected congestion and traffic loss on PE2-CE1 links within the LAG.

1.3 Design Requirement

```
                           +-----------------------+
                           |Underlay Network Fabric|
                           +-----------------------+

                 +-----+   +-----+           +-----+   +-----+
                 | PE1 |   | PE2 |   .....   | PEx |   | PEn |
                 +-----+   +-----+           +-----+   +-----+
                    \       \                 //        //
                     \ L1    \ L2            // Lx     // Ln
                      \       \             //        //
                     +-\-------\-----------//--------//-+
                     |  \       \  ESI-1  //        //  |
                     +----------------------------------+
                                      |
                                      CE
```

To generalize, if total link bandwidth to a CE is distributed across "n" multi-homing PEs, with Lx being the number of links / bandwidth to PEx, traffic from remote PEs to this CE MUST be load-balanced unequally across [PE1, PE2, ..., PEn] such that, fraction of total unicast and BUM flows destined for CE that are serviced by PEx is:

Lx / [L1+L2+.....+Ln]

Solution proposed below includes extensions to EVPN procedures to achieve the above.

1.4 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

"LOCAL PE" in the context of an ESI refers to a provider edge switch OR router that physically hosts the ESI.

"REMOTE PE" in the context of an ESI refers to a provider edge switch OR router in an EVPN overlay, who’s overlay reachability to the ESI is via the LOCAL PE.

2. Solution Overview

In order to achieve weighted load balancing for overlay unicast traffic, Ethernet A-D per-ES route (EVPN Route Type 1) is leveraged to signal the ESI bandwidth to remote PEs. Using Ethernet A-D per-ES route to signal the ESI bandwidth provides a mechanism to be able to react to changes in access bandwidth in a service and host independent manner. Remote PEs computing the MAC path-lists based on global and aliasing Ethernet A-D routes now have the ability to setup weighted load-balancing path-lists based on the ESI access bandwidth received from each PE that the ESI is multi-homed to. If Ethernet A-D per-ES route is also leveraged for IP path-list computation, as per [EVPN-IP-ALIASING], it also provides a method to do weighted load-balancing for IP routed traffic.

In order to achieve weighted load-balancing of overlay BUM traffic, EVPN ES route (Route Type 4) is leveraged to signal the ESI bandwidth to PEs within an ESI’s redundancy group to influence per-service DF election. PEs in an ESI redundancy group now have the ability to do service carving in proportion to each PE’s relative ESI bandwidth.

Procedures to accomplish this are described in greater detail next.

3. Weighted Unicast Traffic Load-balancing

3.1 LOCAL PE Behavior

A PE that is part of an ESI’s redundancy group would advertise a additional "link bandwidth" EXT-COMM attribute with Ethernet A-D per-ES route (EVPN Route Type 1), that represents total bandwidth of PE’s physical links in an ESI. BGP link bandwidth EXT-COMM defined in [BGP-LINK-BW] is re-used for this purpose.

3.1 Link Bandwidth Extended Community

Link bandwidth extended community described in [BGP-LINK-BW] for layer 3 VPNs is re-used here to signal local ES link bandwidth to remote PEs. link-bandwidth extended community is however defined in [BGP-LINK-BW] as optional non-transitive. In inter-AS scenarios, link-bandwidth may need to be signaled to an eBGP neighbor along with next-hop unchanged. It is work in progress with authors of [BGP-LINK-BW] to allow for this attribute to be used as transitive in inter-AS scenarios.

3.2 REMOTE PE Behavior

A receiving PE should use per-ES link bandwidth attribute received from each PE to compute a relative weight for each remote PE, per-ES, as shown below.

if,

```
    L(x,y) : link bandwidth advertised by PE-x for ESI-y

    W(x,y) : normalized weight assigned to PE-x for ESI-y

    H(y)   : Highest Common Factor (HCF) of [L(1,y), L(2,y), .....,
             L(n,y)]
```

then, the normalized weight assigned to PE-x for ESI-y may be computed as follows:

```
    W(x,y) = L(x,y) / H(y)
```

For a MAC+IP route (EVPN Route Type 2) received with ESI-y, receiving PE MUST compute MAC and IP forwarding path-list weighted by the above normalized weights.

As an example, for a CE dual-homed to PE-1, PE-2, PE-3 via 2, 1, and 1 GE physical links respectively, as part of a link bundle represented by ESI-10:

```
    L(1, 10) = 2000 Mbps

    L(2, 10) = 1000 Mbps

    L(3, 10) = 1000 Mbps

    H(10) = 1000
```

Normalized weights assigned to each PE for ESI-10 are as follows:

```
    W(1, 10) = 2000 / 1000 = 2.

    W(2, 10) = 1000 / 1000 = 1.

    W(3, 10) = 1000 / 1000 = 1.
```

For a remote MAC+IP host route received with ESI-10, forwarding load-balancing path-list must now be computed as: [PE-1, PE-1, PE-2, PE-3] instead of [PE-1, PE-2, PE-3]. This now results in load-balancing of all traffic destined for ESI-10 across the three multi-homing PEs in proportion to ESI-10 bandwidth at each PE.

Above weighted path-list computation MUST only be done for an ESI, IF a link bandwidth attribute is received from ALL of the PE’s advertising reachability to that ESI via Ethernet A-D per-ES Route Type 1. In the event that link bandwidth attribute is not received from one or more PEs, forwarding path-list would be computed using regular ECMP semantics.

4. Weighted BUM Traffic Load-Sharing

Optionally, load sharing of per-service DF role, weighted by individual PE’s link-bandwidth share within a multi-homed ES may also be achieved.

In order to do that, a new DF Election Capability [EVPN-DF-ELECT-FRAMEWORK] called "BW" (Bandwidth Weighted DF Election) is defined. BW may be used along with some DF Election Types, as described in the following sections.

4.1 The BW Capability in the DF Election Extended Community

[EVPN-DF-ELECT-FRAMEWORK] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests a bit in the DF Election extended community Bitmap:

Bit 28: BW (Bandwidth Weighted DF Election)

ES routes advertised with the BW bit set will indicate the desire of the advertising PE to consider the link-bandwidth in the DF Election algorithm defined by the value in the "DF Type".

As per [EVPN-DF-ELECT-FRAMEWORK], all the PEs in the ES MUST advertise the same Capabilities and DF Type, otherwise the PEs will fall back to Default [RFC7432] DF Election procedure.

The BW Capability MAY be advertised with the following DF Types:

- o Type 0: Default DF Election algorithm, as in [RFC7432]
- o Type 1: HRW algorithm, as in [EVPN-DF-ELECT-FRAMEWORK]
- o Type 2: Preference algorithm, as in [EVPN-DF-PREF]
- o Type 4: HRW per-multicast flow DF Election, as in [XXX]

The following sections describe how the DF Election procedures are modified for the above DF Types when the BW Capability is used.

4.2 BW Capability and Default DF Election algorithm

When all the PEs in the ES agree to use the BW Capability with DF Type 0, the Default DF Election procedure is modified as follows:

- o Each PE advertises a "Link Bandwidth" EXT-COMM attribute along with the ES route to signal the PE-CE link bandwidth (LBW) for the ES.
- o A receiving PE MUST use the ES link bandwidth attribute received from each PE to compute a relative weight for each remote PE.
- o The DF Election procedure MUST now use this weighted list of PEs to compute the per-VLAN Designated Forwarder, such that the DF role is distributed in proportion to this normalized weight.

Considering the same example as in Section 3, the candidate PE list for DF election is:

[PE-1, PE-1, PE-2, PE-3].

The DF for a given VLAN-a on ES-10 is now computed as (VLAN-a % 4). This would result in the DF role being distributed across PE1, PE2, and PE3 in portion to each PE’s normalized weight for ES-10.

4.3 BW Capability and HRW DF Election algorithm (Type 1 and 4)

[EVPN-DF-ELECT-FRAMEWORK] introduces Highest Random Weight (HRW) algorithm (DF Type 1) for DF election in order to solve potential DF election skew depending on Ethernet tag space distribution. [EVPN-PER-MCAST-FLOW-DF] further extends HRW algorithm for per-multicast flow based hash computations (DF Type 4). This section describes extensions to HRW Algorithm for EVPN DF Election specified in [EVPN-DF-ELECT-FRAMEWORK] and in [EVPN-PER-MCAST-FLOW-DF] in order to achieve DF election distribution that is weighted by link bandwidth.

4.3.1 BW Increment

A new variable called "bandwidth increment" is computed for each [PE, ES] advertising the ES link bandwidth attribute as follows:

In the context of an ES,

L(i) = Link bandwidth advertised by PE(i) for this ES

L(min) = lowest link bandwidth advertised across all PEs for this ES

Bandwidth increment, "b(i)" for a given PE(i) advertising a link bandwidth of L(i) is defined as an integer value computed as:

b(i) = L(i) / L(min)

As an example,

with PE(1) = 10, PE(2) = 10, PE(3) = 20

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 2

with PE(1) = 10, PE(2) = 10, PE(3) = 10

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 1

Note that the bandwidth increment must always be an integer, including, in an unlikely scenario of a PE’s link bandwidth not being an exact multiple of L(min). If it computes to a non-integer value (including as a result of link failure), it MUST be rounded down to an integer.

4.3.2 HRW Hash Computations with BW Increment

HRW algorithm as described in [EVPN-DF-ELECT-FRAMEWORK] and in [EVPN-PER-MCAST-FLOW-DF] compute a random hash value (referred to as affinity here) for each PE(i), where, (0 < i <= N), PE(i) is the PE at ordinal i, and Address(i) is the IP address of PE at ordinal i.

For ’N’ PEs sharing an Ethernet segment, this results in ’N’ candidate hash computations. PE that has the highest hash value is selected as the DF.

Affinity computation for each PE(i) is extended to be computed one per-bandwidth increment associated with PE(i) instead of a single affinity computation per PE(i).

PE(i) with b(i) = j, results in j affinity computations:

affinity(i, x), where 1 < x <= j

This essentially results in number of candidate HRW hash computations for each PE that is directly proportional to that PE’s relative bandwidth within an ES and hence gives PE(i) a probability of being DF in proportion to it’s relative bandwidth within an ES.

As an example, consider an ES that is multi-homed to two PEs, PE1 and PE2, with equal bandwidth distribution across PE1 and PE2. This would result in a total of two candidate hash computations:

affinity(PE1, 1)

affinity(PE2, 1)

Now, consider a scenario with PE1’s link bandwidth as 2x that of PE2. This would result in a total of three candidate hash computations to be used for DF election:

affinity(PE1, 1)

affinity(PE1, 2)

affinity(PE2, 1)

which would give PE1 2/3 probability of getting elected as a DF, in proportion to its relative bandwidth in the ES.

Depending on the chosen HRW hash function, affinity function MUST be extended to include bandwidth increment in the computation.

For e.g.,

affinity function specified in [EVPN-PER-MCAST-FLOW-DF] MAY be extended as follows to incorporate bandwidth increment j:

affinity(S,G,V, ESI, Address(i,j)) =
(1103515245.((1103515245.Address(i).j + 12345) XOR
D(S,G,V,ESI))+12345) (mod 2^31)

affinity or random function specified in [EVPN-DF-ELECT-FRAMEWORK] MAY be extended as follows to incorporate bandwidth increment j:

affinity(v, Es, Address(i,j)) = (1103515245((1103515245.Address(i).j
+ 12345) XOR D(v,Es))+12345)(mod 2^31)

4.3.3 Cost-Benefit Tradeoff on Link Failures

While incorporating link bandwidth into the DF election process provides optimal BUM traffic distribution across the ES links, it also implies that affinity values for a given PE are re-computed, and DF elections are re-adjusted on changes to that PE’s bandwidth increment that might result from link failures or link additions. If the operator does not wish to have this level of churn in their DF election, then they should not advertise the BW capability. Not advertising BW capability may result in less than optimal BUM traffic distribution while still retaining the ability to allow a remote ingress PE to do weighted ECMP for its unicast traffic to a set of multi-homed PEs, as described in section 3.2.

Same also applies to use of BW capability with service carving (DF Type 0), as specified in section 4.2.

4.4 BW Capability and Preference DF Election algorithm

This section applies to ES’es where all the PEs in the ES agree use the BW Capability with DF Type 2. The BW Capability modifies the Preference DF Election procedure [EVPN-DF-PREF], by adding the LBW value as a tie-breaker as follows:

- Section 4.1, bullet (f) in [EVPN-DF-PREF] now considers the LBW value:

  f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP bit, the LBW value and the lowest IP PE in that order. For instance:

  - If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=1, LBW=2000] in PE2, PE2 would be elected due to the DP bit.
  - If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=0, LBW=2000] in PE2, PE2 would be elected due to a higher LBW, even if PE1’s IP address is lower.
  - The LBW exchanged value has no impact on the Non-Revertive option described in [EVPN-DF-PREF].

5. Real-time Available Bandwidth

PE-CE link bandwidth availability may sometimes vary in real-time disproportionately across PE_CE links within a multi-homed ESI due to various factors such as flow based hashing combined with fat flows and unbalanced hashing. Reacting to real-time available bandwidth is at this time outside the scope of this document. Procedures described in this document are strictly based on static link bandwidth parameter.

6. Routed EVPN Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

o A host MAC is never advertised in EVPN overlay control plane o Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI

o An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged

o Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC 7814] for more details.

Weighted multi-path procedure described in this document may be used together with procedures described in [EVPN-IP-ALIASING] for this use case. Ethernet A-D per-ES route advertised with Layer 3 VRF RTs would be used to signal ES link bandwidth attribute instead of the Ethernet A-D per-ES route with Layer 2 VRF RTs. All other procedures described earlier in this document would apply as is.

If [EVPN-IP-ALIASING] is not used for routed fast convergence, link bandwidth attribute may still be advertised with IP routes (RT-5) to achieve PE-CE link bandwidth based load-balancing as described in this document. In the absence of [EVPN-IP-ALIASING], re-balancing of traffic following changes in PE-CE link bandwidth will require all IP routes from that CE to be re-advertised in a prefix dependent manner.

7. EVPN-IRB Multi-homing with non-EVPN routing

EVPN-LAG based multi-homing on an IRB gateway may also be deployed together with non-EVPN routing, such as global routing or an L3VPN routing control plane. Key property that differentiates this set of use cases from EVPN IRB use cases discussed earlier is that EVPN control plane is used only to enable LAG interface based multi-homing and NOT as an overlay VPN control plane. EVPN control plane in this case enables:

- DF election via EVPN RT-4 based procedures described in [RFC7432]
- LOCAL MAC sync across multi-homing PEs via EVPN RT-2
- LOCAL ARP and ND sync across multi-homing PEs via EVPN RT-2

Applicability of weighted ECMP procedures proposed in this document to these set of use cases will be addressed in subsequent revisions.

7. References

7.1 Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[BGP-LINK-BW] Mohapatra, P., Fernando, R., "BGP Link Bandwidth Extended Community", January 2013, <https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-06>.

[EVPN-IP-ALIASING] Sajassi, A., Badoni, G., "L3 Aliasing and Mass Withdrawal Support for EVPN", July 2017, <https://tools.ietf.org/html/draft-sajassi-bess-evpn-ip-aliasing-00>.

[EVPN-DF-PREF] 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-01.txt, April 2018.

[EVPN-PER-MCAST-FLOW-DF] Sajassi, et al., "Per multicast flow Designated Forwarder Election for EVPN", March 2018, <https://tools.ietf.org/html/draft-sajassi-bess-evpn-per-mcast-flow-df-election-00>.

[EVPN-DF-ELECT-FRAMEWORK] Rabadan, Mohanty, et al., "Framework for EVPN Designated Forwarder Election Extensibility", March 2018, <https://tools.ietf.org/html/draft-ietf-bess-evpn-df-election-framework-03>.

[RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", March 1997, <https://tools.ietf.org/html/rfc2119>.

[RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", May 2017, <https://tools.ietf.org/html/rfc8174>.

7.2 Informative References

8. Acknowledgements

Authors would like to thank Satya Mohanty for valuable review and inputs with respect to HRW algorithm refinements proposed in this document.

Authors’ Addresses

Neeraj Malhotra, Ed.
Arrcus
Email: neeraj.ietf@gmail.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

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

John Drake
Juniper
EMail: jdrake@juniper.net

Avinash Lingala
AT&T
Email: ar977m@att.com

Samir Thoria
Cisco
Email: sthoria@cisco.com

BESS Workgroup   J. Rabadan, Ed.
Internet Draft   Nokia
Intended status: Standards Track   A. Sajassi, Ed.
                                                                      Cisco

                                                                   E. Rosen
                                                                   J. Drake
                                                                     W. Lin
                                                                    Juniper

                                                                  J. Uttaro
                                                                       AT&T

                                                                 A. Simpson
                                                                      Nokia

Expires: April 28, 2018   October 25, 2017

EVPN Interworking with IPVPN
draft-rabadan-sajassi-bess-evpn-ipvpn-interworking-00

Abstract

EVPN is used as a unified control plane for tenant intra and inter-subnet-forwarding. When tenant connectivity spans not only EVPN domains but also domains where IPVPN provides inter-subnet-forwarding, there is a need to specify the interworking aspects between both EVPN and IPVPN domains, so that the end to end tenant connectivity can be accomplished. This document specifies how EVPN should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families for inter-subnet-forwarding.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction And Problem Statement

EVPN is used as a unified control plane for tenant intra and inter-subnet-forwarding. When tenant connectivity spans not only EVPN domains but also domains where IPVPN provides inter-subnet-forwarding, there is a need to specify the interworking aspects between both EVPN and IPVPN domains, so that the end to end tenant connectivity can be accomplished. This document specifies how EVPN should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families for inter-subnet-forwarding.

EVPN supports the advertisement of ipv4 or ipv6 prefixes in two different route types:

- Route Type 2 - MAC/IP route (only for /32 and /128 host routes), as described by [INTER-SUBNET].

- Route Type 5 - IP Prefix route, as described by [IP-PREFIX].

When interworking with other BGP address families (AFIs/SAFIs) for inter-subnet-forwarding, the IP- Prefixes in those two EVPN route types must be propagated to other domains using different SAFIs. Some aspects of that propagation must be clarified. Examples of these aspects or procedures across BGP families are: route selection, loop prevention or BGP Path attribute propagation. The Interworking PE concepts are defined in section 2, and the rest of the document describes the interaction between Interworking PEs and other PEs for end-to-end inter-subnet-forwarding.

2. Terminology and Interworking PE components

This section summarizes the terminology related to the "Interworking PE" concept that will be used throughout the rest of the document.

```
       +------------------------------------------------------------+
       |                                                            |
       |              +-----------------+           Interworking PE |
       | Attachment   | +-----------------+                         |
       | Circuit(AC1) | |  +----------+   |                MPLS/NVO tnl
     ----------------------*Bridge    |   |                    +------
       |              | |  |Table(BT1)|   |    +-----------+  / \     \
    MPLS/NVO tnl +-------->|          *--------*           |<--> | Eth |
      -------+   |    | |  |Eth-Tag x +   |IRB1|           |  \ /     /
     / Eth  / \<-+    | |  +----------+   |    |           |   +------
    |      |   |    | |     ...         |    |  IP-VRF1  |        |
     \      \ /<-+    | |  +----------+   |    |  RD2/RT2  |MPLS/NVO tnl
      -------+   |    | |  |Bridge    |   |    |           |   +------
       |         +-------->|Table(BT2)|   |IRB2|           |  / \     \
       |              | |  |          *--------*           |<--> | IP  |
     ----------------------*Eth-Tag y |   |    +-----*-----+  \ /     /
       |  AC2         | |  +----------+   |      AC3|         +------
       |              | |    MAC-VRF1     |          |              |
       |              +-+    RD1/RT1      |          |              |
       |                +-----------------+          |  SAFIs       |
       |                                             |  1     +---+ |
     ------------------------------------------------+  128   |BGP| |
       |                                                EVPN  +---+ |
       |                                                            |
       +------------------------------------------------------------+
```

Figure 1 EVPN-IPVPN Interworking PE

- o ISF SAFI: Inter-Subnet-Forwarding (ISF) SAFI is a MP-BGP Sub-Address Family that advertises reachability for IP-Prefixes and can be used for inter-subnet-forwarding within a given tenant Domain. The ISF SAFIs are 1 (including IPv4 and IPv6 AFIs), 128 (including IPv4 and IPv6 AFIs) and 70 (EVPN, including only AFI 25).

- o IP-VRF: an IP Virtual Routing and Forwarding table, as defined in [RFC4364]. It is the instantiation of an IPVPN in a PE. Route-distinghisher and route-target(s) are required properties of an IP-VRF.

- o MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [RFC7432]. The instantiation of an EVI (EVPN Instance) in a PE. Route-distinghisher and route-target(s) are required properties and they are normally different than the ones defined in the associated IP-VRF.

- o BT: a Bridge Table, as defined in [RFC7432]. A BT is the instantiation of a Broadcast Domain in a PE. When there is a single

Broadcast Domain in a given EVI, the MAC-VRF in each PE will contain a single BT. When there are multiple BTs within the same MAC-VRF, each BT is associated to a different Ethernet Tag. The EVPN routes specific to a BT, will indicate which Ethernet Tag the route corresponds to.

Example: In Figure 1, MAC-VRF1 has two BTs: BT1 and BT2. Ethernet Tag x is defined in BT1 and Ethernet Tag y in BT2.

o AC: Attachment Circuit or logical interface associated to a given BT or IP-VRF. To determine the AC on which a packet arrived, the PE will examine the combination of a physical port and VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).

Example: In Figure 1, AC1 is associated to BT1, AC2 to BT2 and AC3 to IP-VRF1.

o IRB: Integrated Routing and Bridging interface. It refers to the logical interface that connects a BT to an IP-VRF and allows to forward packets with destination in a different subnet.

o MPLS/NVO tnl: It refers to a tunnel that can be MPLS or NVO-based (Network Virtualization Overlays) and it is used by MAC-VRFs and IP-VRFs. Irrespective of the type, the tunnel may carry an Ethernet or an IP payload. MAC-VRFs can only use tunnels with Ethernet payloads (setup by EVPN), whereas IP-VRFs can use tunnels with Ethernet (setup by EVPN) or IP payloads (setup by EVPN or IPVPN). IPVPN-only PEs have IP-VRFs but they cannot send or receive traffic on tunnels with Ethernet payloads.

Example: Figure 1 shows an MPLS/NVO tunnel that is used to transport Ethernet frames to/from MAC-VRF1. The PE determines the MAC-VRF and BT the packets belong to based on the EVPN label (MPLS or VNI). Figure 1 also shows two MPLS/NVO tunnels being used by IP-VRF1, one carrying Ethernet frames and the other one carrying IP packets.

o RT-2: Route Type 2 or MAC/IP route, as per [RFC7432].

o RT-5: Route Type 5 or IP-Prefix route, as per [IP-PREFIX].

o Interworking PE: a PE that may advertise a given prefix in both an EVPN route (RT-2 or RT-5) and in a route of another ISF SAFI. An Interworking PE has one IP-VRF per tenant, and one or multiple MAC-VRFs per tenant. Each MAC-VRF may contain one or more BTs, where each BT may be attached to that IP-VRF via IRB. There are two types of Interworking PEs: Composite PEs and Gateway PEs. Both PE functions can be independently implemented per tenant and they may both be implemented for the same tenant.

Example: Figure 1 shows an Interworking PE, where ISF SAFIs 1, 128 and 70 are enabled. IP-VRF1 and MAC-VRF1 are instantiated on the PE, and together provide inter-subnet-forwarding for the tenant.

- Composite PE: an Interworking PE that advertises a given IP Prefix multiple times to the same BGP peer, but using a different ISF SAFI each time, and being EVPN one of them.

  Example: Figure 2 shows an example where PE1 is a Composite PE since PE1 has EVPN and another ISF SAFI enabled to the same route-reflector, and PE1 advertises a given IP Prefix IPn/x twice, one using EVPN and another one using ISF SAFI 128. PE2 and PE3 are not Composite PEs.

```
                                   +---+
                                   |PE2|
                                   +---+
                                    ^
                                    |EVPN
                       IW    EVPN   v
                      +---+  IPVPN ++-+       +---+
                      |PE1| <----> |RR| <---> |PE3|
                      +---+        +--+ IPVPN +---+
                    Composite
```

Figure 2 Interworking Composite PE example

- Gateway PE: an Interworking PE that advertises IP Prefixes to different BGP peers, using EVPN to one BGP peer and another ISF SAFI to another BGP peer.

  Example: Figure 3 illustrates an example where PE1 is a Gateway PE since the EVPN and IPVPN SAFIs are enabled on different BGP peers, and a given local IP Prefix IPn/x is sent to both BGP peers for the same tenant.

```
                                     IW
                       +---+ EVPN   +---+ IPVPN  +---+
                       |PE2| <----> |PE1| <----> |PE3|
                       +---+        +---+        +---+
                                   Gateway
```

Figure 3 Interworking Gateway PE example

o Domain: a set of IP-VRFs for the same tenant that have been configured with the same Domain-ID. Two PEs are in the same Domain if they are attached to the same tenant and the packets between them do not require a data path IP lookup (in the tenant space) in any intermediate router. A Gateway PE is always configured with multiple Domain-IDs.

Example 1: Figure 4 depicts an example where TS1 and TS2 belong to the same tenant, and they are located in different Data Centers that are connected by Gateway PEs. These Gateway PEs speak IPVPN in the WAN. When TS1 sends traffic to TS2, the intermediate routers between PE1 and PE2 require a tenant IP lookup in their IP-VRFs so that the packets can be forwarded. In this example there are three different Domains. The Gateway PEs connect the EVPN Domain to the IPVPN Domain.

```
                           GW1------------GW3
                         +------+       +------+
           +-------------|IP-VRF|       |IP-VRF|-------------+
          PE1            +------+       +------+            PE2
        +------+   DC1      |     WAN      |     DC2     +------+
    TS1-|IP-VRF|   EVPN     |    IPVPN     |     EVPN    |IP-VRF|-TS2
        +------+           GW2            GW4            +---+--+
           |             +------+       +------+             |
           +-------------|IP-VRF|       |IP-VRF|-------------+
                         +------+       +------+
                            +--------------+
               DOMAIN 1         DOMAIN 2       DOMAIN 3
           <---------------> <------------> <---------------->
```

Figure 4 Multiple Domain DCI example

Example 2: Figure 5 illustrates a similar example, but PE1 and PE2 are now connected by a BGP-LU (BGP Labeled Unicast) tunnel, and they have a BGP peer relationship for EVPN. Contrary to Example 1, there is no need for tenant IP lookups on the intermediate routers in order to forward packets between PE1 and PE2. Therefore, there is only one Domain in the network and PE1/PE2 belong to it.

```
                                EVPN
           <-------------------------------------------------->
                                BGP-LU
           <-------------------------------------------------->

                          ASBR------------ASBR
                         +------+       +------+
           +-------------|      |       |      |-------------+
          PE1            +------+       +--+---+           PE2
        +------+   DC1      |     WAN      |     DC2     +------+
    TS1-|IP-VRF|   EVPN     |              |     EVPN    |IP-VRF|-TS2
        +------+          ASBR            ASBR           +---+--+
           |             +------+       +------+             |
           +-------------|      |       |      |-------------+
                         +------+       +------+
                            +--------------+

           <--------------------DOMAIN-1---------------------->
```

Figure 5 Single Domain DCI example

The Domain-ID is encoded in the Domain Path Attribute (D-PATH), and advertised along with EVPN and other ISF SAFI routes. Section 3 describes the D-PATH attribute.

3. Domain Path Attribute (D-PATH)

The BGP Domain Path (D-PATH) attribute is an optional and transitive BGP path attribute.

Similar to AS_PATH, D-PATH is composed of a length field followed by a sequence of Domain segments, where each Domain segment is represented by <DOMAIN-ID:ISF_SAFI_TYPE>.

- The length field is a 1-octet field, containing the number of Domain segments. Each Domain segment value field contains one or more Domain segments, each encoded as a 7-octet length field.

- DOMAIN-ID is a 6-octet field that represents a Domain. It is composed of a 4-octet Global Administrator sub-field and a 2-octet Local Administrator sub-field. The Global Administrator sub-field

MAY be filled with an Autonomous System Number (ASN), an IPv4 address, or any value that guarantees the uniqueness of the DOMAIN-ID when the tenant expands across multiple Operators.

o ISF_SAFI_TYPE is a 1-octet field that indicates the Inter-Subnet-Forwarding SAFI type of the DOMAIN. The following types are valid in this document:

Value	Type
1	SAFI 1
70	EVPN
128	SAFI 128

About the BGP D-PATH attribute:

a) Identifies the sequence of <DOMAIN-ID:ISF_SAFI_TYPE> segments through which the update message has passed.

- This attribute list may contain zero, one or more entries.

- The leftmost entry in the list is the <DOMAIN-ID:ISF_SAFI_TYPE> that the Gateway PE added when sending the prefix into the local DOMAIN. The rightmost entry in the list is the originating <DOMAIN-ID:ISF_SAFI_TYPE> for the prefix, that was added by the first Gateway PE propagating the update between Domains. Intermediate entries are transit <DOMAIN-ID:ISF_SAFI_TYPE> that the update has passed through on its way.

- As an example, an EVPN Prefix route received with D-PATH {<6500:2:IPVPN>,<6500:1:EVPN>} indicates that the Prefix was originally advertised in EVPN within Domain 6500:1, re-advertised by a first Gateway PE using an IPVPN route in Domain 6500:2 and re-advertised by a second Gateway PE into the local Domain.

b) It is added/modified by a Gateway PE when propagating an update to a different Domain:

- A Gateway PE’s IP-VRF, that connects two Domains, belongs to two DOMAIN-IDs, e.g. 6500:1 for EVPN and 6500:2 for IPVPN.

- Whenever a Prefix arrives at a Gateway PE in a particular ISF SAFI route, if the Gateway PE needs to export that Prefix to a BGP peer using a different ISF SAFI, the Gateway PE will prepend a <DOMAIN-ID:ISF_SAFI_TYPE> segment to the list of segments in the received D-PATH.

- For instance, in an IP-VRF configured with DOMAIN-IDs 6500:1 for EVPN and 6500:2 for IPVPN, if an EVPN route for Prefix P is received and P installed in the IP-VRF, the IPVPN route for P that is exported to an IPVPN peer will prepend the segment <6500:1:EVPN> to the previously received D-PATH attribute. Likewise, IP-VRF prefixes that are received from IP-VPN, will be exported to EVPN peers with the additional segment <6500:2:IPVPN>.

- In the above example, if the EVPN route is received without D-PATH, the Gateway PE will add the D-PATH attribute with segment <6500:1:EVPN> when re-advertising to Domain 6500:2.

- Within the originating Domain, the update does not contain a D-PATH attribute because the update has not passed through a Gateway PE yet.

c) The Gateway PE MUST NOT add the D-PATH attribute to routes generated for IP-VRF Prefixes that are not learned via any ISF SAFI, for instance, local prefixes.

d) A route received with a D-PATH that contains at least one of the locally configured Domains for the IP-VRF MUST be flagged as a looped update.

e) The number of <DOMAIN-ID:ISF_SAFI_TYPE> segments in the D-PATH list reflects the number of Gateway PEs the update has gone through, irrespective of the actual number of BGP speakers along the way.

4. Route selection process between EVPN and other ISF SAFIs

A PE may receive an IP Prefix simultaneously from EVPN and another ISF SAFI, e.g. SAFI 128 or 1, from the same or different BGP neighbor. In addition, a router may receive the same IP Prefix (host route) simultaneously from an EVPN RT-5 and a RT-2. A route selection algorithm across EVPN and other ISF SAFIs is needed so that:

o Different Gateway and Composite PEs have a consistent and deterministic view on how to reach a given Prefix.

o Prefixes advertised in EVPN and other ISF SAFIs can be compared based on path attributes commonly used by operators across networks.

- o Equal Cost Multi-Path (ECMP) is allowed across EVPN and other ISF SAFI routes.

For a given prefix advertised in one or more non-EVPN ISF SAFIs, the BGP best path selection procedure will produce a set of "non-EVPN best paths". For a given prefix advertised in the EVPN ISF SAFI, the BGP best path selection procedure will produce a set of "EVPN best paths". To support IP/EVPN interworking, it is then necessary to run a tie-breaking selection algorithm on the union of these two sets. This tie-breaking algorithm begins by considering all EVPN and other ISF SAFI routes, equally preferable routes to the same destination, and then selects routes to be removed from consideration. The process terminates as soon as only one route remains in consideration.

The route selection algorithm must remove from consideration the routes following the rules and the order defined in [RFC4271], with the following exceptions and in the following order:

1- Immediately after removing from consideration all routes that are not tied for having the highest Local Preference, any routes that do not have the shortest D-PATH are also removed from consideration. Routes with no D-PATH are considered to have a zero-length D-PATH.

2- Then regular [RFC4271] selection criteria is followed.

3- At the end of the selection algorithm, if at least one route still under consideration is an RT-2 route, remove from consideration any RT-5 routes.

4- Steps 1-3 could possibly leave Equal Cost Multi-Path (ECMP) between IP and EVPN paths. By default, the EVPN path is considered (and the IP path removed from consideration). However, if ECMP across ISF SAFIs is enabled by policy, and an "IP path" and an "EVPN path" remain at the end of step 3, both path types will be used.

Example 1 - PE1 receives the following routes for IP1/32, that are candidate to be imported in IP-VRF-1:

```
{SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=(100,200)}
{SAFI=EVPN, RT-5, Local-Pref=100, AS-Path=(100,200)}
{SAFI=128, Local-Pref=100, AS-Path=(100,200)}

Selected route: {SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=100,200]
(due to step 3, and no ECMP)
```

Example 2 - PE1 receives the following routes for IP2/24, that are candidate to be imported in IP-VRF-1:

{SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-Path=(100,200), MED=10}
{SAFI=128, D-PATH=(6500:1:EVPN,6500:2:IPVPN), AS-Path=(200), MED=200}

Selected route: {SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-Path=(100,200), MED=10} (due to step 1)

5. Loop Prevention

The D-PATH attribute (see section 3) is used to prevent loops in interworking PE networks. For instance, in the example of Figure 4, Gateway GW1 receives TS1 Prefix in two different updates:

o In an EVPN RT-5 with next-hop PE1 and no D-PATH attribute.

o In a SAFI 128 route with next-hop GW2 and D-PATH = (6500:1:EVPN), assuming that DOMAIN-ID for Domain 1 is 6500:1.

Gateway GW1 flags the SAFI 128 route as a loop, and does not re-advertise it to the EVPN neighbors since the route includes the GW1’s local Domain.

In general, any Interworking PE that imports a Prefix route MUST flag the route as "looped" if its D-PATH contains a <DOMAIN-ID:ISF_SAFI_TYPE> segment, where DOMAIN-ID matches a local DOMAIN-ID in the tenant IP-VRF.

6. BGP Path Attribute Propagation Across ISF SAFIs

The following modes of operation are defined on the Gateway PEs.

6.1. No-Propagation-Mode

This is the default mode of operation. In this mode, the Gateway PE will simply re-initialize the Path Attributes when re-advertising a route to a different SAFI, as though it would for direct or local IP-Prefixes. This model may be enough in those use-cases where the EVPN Domain is considered an "abstracted" CE and remote IPVPN/IP PEs don’t need to consider the original EVPN Attributes for path calculations.

However, since this mode of operation does not propagate the D-PATH attribute either, redundant Gateway PEs are exposed to routing loops. Those loops may be resolved by policies and the use of other attributes, such as the Route Origin extended community [RFC4360], however not all the loop situations may be solved.

6.2. Uniform-Propagation-Mode

In this mode, the Gateway PE simply keeps accumulating or mapping certain key commonly used Path Attributes when re-advertising routes to a different ISF SAFI. This mode is typically used in networks where EVPN and IPVPN SAFIs are used seamlessly to distribute IP Prefixes.

The following rules MUST be observed by the Gateway PE when propagating Path Attributes:

- The Gateway PE imports the routes in the IP-VRF and stores the original Path Attributes. The following set of Path Attributes SHOULD be propagated by the Gateway PE to other ISF SAFIs (other Path Attributes SHOULD NOT be propagated):
    - AS_PATH
    - D-PATH
    - IBGP-only Path Attributes: LOCAL_PREF, ORIGINATOR_ID, CLUSTER_ID
    - MED
    - AIGP
    - Communities, (non-EVPN) Extended Communities and Large Communities

- When re-advertising a route to a different ISF SAFI and IBGP peer, the Gateway PE SHOULD copy the AS_PATH of the originating family and add it to the destination family without any modification. When re-advertising to a different ISF SAFI and EBGP peer, the Gateway PE SHOULD copy the AS_PATH of the originating family and prepend the IP-VRF’s AS before sending the route.

- When re-advertising a route to IBGP peers, the Gateway PE SHOULD copy the IBGP-only Path Attributes from the originating SAFI to the re-advertised route.

- Communities, non-EVPN Extended Communities and Large Communities SHOULD be copied by the Gateway PE from the originating SAFI route.

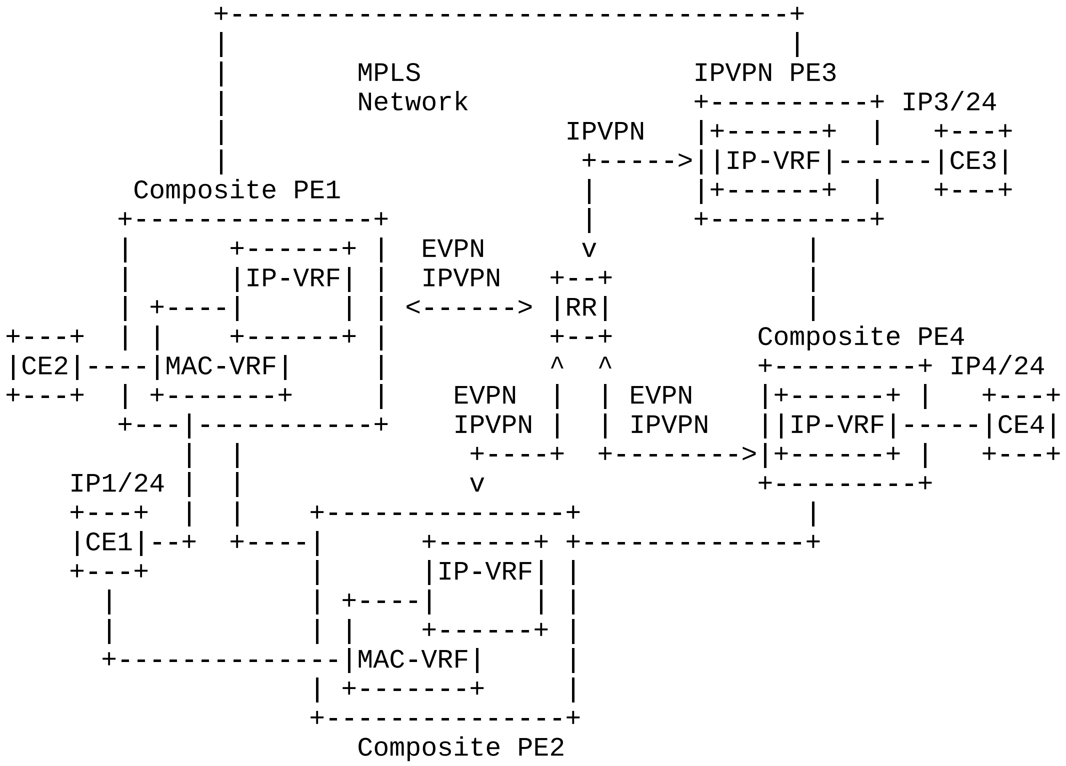
6.3. Aggregation of Host Routes and Path Attribute Propagation

This section will be addressed in future revisions of the document.

7. Composite PE Procedures

As described in Section 2, a Composite PE is defined as an Interworking PE where the same IP Prefix is advertised multiple times to the same BGP peer, but using EVPN and another ISF SAFI. Composite PEs are typically used in tenant networks where EVPN and IPVPN are both used to provide inter-subnet-forwarding within the same tenant Domain.

Figure 6 depicts an example of a tenant Domain. As defined in section 2, two PEs are in the same Domain if they are attached to the same tenant and the traffic forwarding between them does not require any data path IP lookup (in the tenant space) in any intermediate router. PE1/PE2/PE4 are Composite PEs (they support EVPN and IPVPN ISFs on their peering to the Route Reflector), and PE3 is a regular IPVPN PE. The four PEs belong to the same Domain.



```
                +-----------------------------------+
                |                                   |
                |        MPLS                 IPVPN PE3
                |        Network              +----------+ IP3/24
                |                     IPVPN   |+------+  |   +---+
                |                      +----->||IP-VRF|------|CE3|
           Composite PE1               |      |+------+  |   +---+
          +---------------+            |      +----------+
          |      +------+ |  EVPN      v             |
          |      |IP-VRF| |  IPVPN   +--+            |
          | +----|      | | <------> |RR|            |
   +---+  | |    +------+ |          +--+         Composite PE4
   |CE2|----|MAC-VRF|     |          ^  ^         +---------+ IP4/24
   +---+  | +-------+     |    EVPN  |  | EVPN    |+------+ |   +---+
          +---|-----------+    IPVPN |  | IPVPN   ||IP-VRF|-----|CE4|
              |  |              +----+  +-------->|+------+ |   +---+
       IP1/24 |  |              v                 +---------+
       +---+  |  |    +---------------+              |
       |CE1|--+  +----|      +------+ +--------------+
       +---+          |      |IP-VRF| |
         |            | +----|      | |
         |            | |    +------+ |
         +--------------|MAC-VRF|     |
                      | +-------+     |
                      +---------------+
                         Composite PE2
```

Figure 6 Composite PE example

In a Domain with Composite and regular PEs:

- The Composite PEs advertise the same IP Prefixes in each ISF SAFI to the RR. For example, the prefix IP1/24 is advertised by PE1 and PE2 to the RR in two separate NLRIs, one for AFI/SAFI 1/128 and another one for EVPN.

- The RR does not forward EVPN routes to PE3 (since the RR does not have the EVPN SAFI enabled on its BGP session to PE3), whereas the IPVPN routes are forwarded to all the PEs.

- PE3 receives only the IPVPN route for IP1/24 and resolves the BGP next-hop to an MPLS tunnel (with IP payload) to PE1 and/or PE2.

- Composite PE4 receives IP1/24 encoded in EVPN and another ISF SAFI route (EVPN RT-5 and IPVPN). The route selection follows the procedures in section 4. Assuming an EVPN route is selected, PE4 resolves the BGP next-hop to an MPLS tunnel (with Ethernet or IP payload) to PE1 and/or PE2. As described in section 2, two EVPN PEs may use tunnels with Ethernet or IP payloads to connect their IP-VRFs, depending on the [IP-PREFIX] model implemented. If some attributes are modified so that the route selection process (section 4) results in PE4 selecting the IPVPN path instead of the EVPN path, the operator should be aware that the EVPN advanced forwarding features, e.g. recursive resolution to overlay indexes, will be lost for PE4.

- The other Composite PEs (PE1 and PE2) receive also the same IP Prefix via EVPN and IPVPN SAFIs and they also follow the route selection in section 4.

- When a given route has been selected as the route for a particular packet, the transmission of the packet is done according to the rules for that route’s AFI/SAFI.

- It is important to note that in mixed networks, such as the one in Figure 6, the EVPN advanced forwarding features will only be available to Composite and EVPN PEs (assuming they select an RT-5 to forward packets for a given IP Prefix), and not to IPVPN PEs. For example, assuming PE1 sends IP1/24 in an EVPN and an IPVPN route and the EVPN route is the best one in the selection, the recursive resolution of the EVPN RT-5s can only be used in PE2 and PE4 (Composite PEs), and not in PE3 (IPVPN PE). As a consequence of this, the indirection provided by the RT5’s recursive resolution and its benefits in a scaled network, will not be available in all the PEs in the network.

8. Gateway PE Procedures

Section 2 defines a Gateway PE as an Interworking PE that advertises IP Prefixes to different BGP peers, using EVPN to one BGP peer and another ISF SAFI to another BGP peer. Examples of Gateway PEs are Data Center Gateways connecting Domains that make use of EVPN and other ISF SAFIs for a given tenant. Figure 7 illustrates this use-case, in which PE1 and PE2 (and PE3/PE4) are Gateway PEs interconnecting Domains for the same tenant.

```
    <----EVPN---->    <----------IPVPN--------->   <----EVPN---->
      6500:1:EVPN             6500:2:IPVPN           6500:3:EVPN
 <DOMAIN-ID:ISF_SAFI_TYPE>
                       +-----------------------+
                Gateway PE1              Gateway PE3
                +----------+             +----------+
    +-----------|+------+  |  MPLS tnls  |+------+  |-------------+
    |           ||IP-VRF|  |             ||IP-VRF|  |             |
  PE5           |+------+  |             |+------+  |           PE6
 +------+       +----------+             +----------+          +------+
 |IP-VRF| NVO tnls |   |                       |  |  NVO tnls  |IP-VRF|
 |      |          |   |                       |  |            |      |
 +------+       +----------+             +----------+          +------+
 IP1/24-->      |+------+  |             |+------+  |             |
    |           ||IP-VRF|  |             ||IP-VRF|  |             |
    +-----------|+------+  |             |+------+  |-------------+
                +----------+             +----------+
                Gateway PE2   +------+   Gateway PE4
                      +-------|IP-VRF|---------+
                              |      |
                              +------+
                                PE7
```

Figure 7 Gateway PE example

The Gateway PE procedures are described as follows:

- A Gateway PE that imports an ISF SAFI-x route to prefix P in an IP-VRF, MUST export P in ISF SAFI-y if:

  1. P is installed in the IP-VRF (hence the SAFI-x route is the best one for P) and

  2. PE has a BGP peer for SAFI-y (enabled for the same IP-VRF) and

  3. Either x or y is EVPN.

In the example of Figure 7, Gateway PE1 and PE2 receive an EVPN RT-5 with IP1/24, install the prefix in the IP-VRF and re-advertise it using SAFI 128.

- ISF SAFI routes advertised by a Gateway PE MUST include a D-PATH attribute, so that loops can be detected in remote Gateway PEs. When a Gateway PE re-advertises an IP Prefix between EVPN and another ISF SAFI, it MUST prepend a <DOMAIN-ID:ISF_SAFI_TYPE> to the received D-PATH attribute. The DOMAIN-ID and ISF_SAFI_TYPE fields refer to the Domain over which the Gateway PE received the IP Prefix. If the received IP Prefix route did not include any D-PATH attribute, the Gateway IP MUST add the D-PATH when re-advertising. The D-PATH in this case will have only one segment on the list, the <DOMAIN-ID:ISF_SAFI_TYPE> of the received route.

  In the example of Figure 7, Gateway PE1/PE2 receive the EVPN RT-5 with no D-PATH attribute since the route is originated at PE5. Therefore PE1 and PE2 will add the D-PATH attribute including <DOMAIN-ID:ISF_SAFI_TYPE> = <6500:1:EVPN>. Gateways PE3/PE4 will re-advertise the route again, now prepending their <DOMAIN-ID:ISF_SAFI_TYPE> = <6500:2:IPVPN>. PE6 receives the EVPN RT-5 routes with D-PATH = {<6500:2:IPVPN>,<6500:1:EVPN>} and can use that information to make BGP path decisions.

- The Gateway PE MAY use the route-distinguisher of the IP-VRF to re-advertise IP Prefixes in EVPN or the other ISF SAFI.

- The label allocation used by each Gateway PE is a local implementation matter. The IP-VRF advertising IP Prefixes for EVPN and another ISF SAFI may use a label per-VRF, per-prefix, etc.

- The Gateway PE MUST be able to use the same or different set of route-targets per ISF SAFI on the same IP-VRF. In particular, if different Domains use different set of route-targets for the same tenant, the Gateway PE MUST be able to import and export routes with the different sets.

- Even though Figure 7 only shows two Domains per Gateway PE, the Gateway PEs may be connected to more than two Domains.

- There is no limitation of Gateway PEs that a given IP Prefix can pass through until it reaches a given PE.

- It is worth noting that an IP-Prefix that was originated in an EVPN Domain but traversed a different ISF SAFI Domain, will lose EVPN-specific attributes that are used in advanced EVPN procedures. For example, even if PE1 advertises IP1/24 along with a given non-zero ESI (for recursive resolution to that ESI), when PE6 receives the

IP-Prefix in an EVPN route, the ESI value will be zero. This is because the route traverses an ISF SAFI Domain that is different than EVPN.

9. Interworking Use-Cases

While Interworking PE networks may well be similar to the examples described in sections 7 and 8, in some cases a combination of both functions may be required. Figure 8 illustrates an example where the Gateway PEs are also Composite PEs, since not only they need to re-advertise IP Prefixes from EVPN routes to another ISF SAFI routes, but they also need to interwork with IPVPN-only PEs in a Domain with a mix of Composite and IPVPN-only PEs.

```
                       +-----------------------------------+
                       |                                   |
                       |        MPLS                 IPVPN PE3
                       |        Network              +---------+
                       |                     IPVPN   |+------+ |
                       |                      +----->||IP-VRF|---TS3
                (GW+Composite) PE1            |      |+------+ |
                 +---------------+            |      +---------+
                 |      +------+ |  EVPN      v            |
                 |      |IP+VRF| |  IPVPN   +-++           |
                 | +----|      | | <------> |RR|           |
        +--------| |    +------+ |          +--+        Composite PE4
        |        | |MAC+VRF|     |          ^  ^         +---------+
        |        | +-------+     |    EVPN  |  | EVPN    |+------+ |
     +----+      +---------------+    IPVPN |  | IPVPN   ||IP-VRF|---TS4
 TS1-|NVE1|             |              +----+  +-------->|+------+ |
     +----+             |              v                 +---------+
        |    EVPN DC    |    +---------------+             |
        |    NVO tnls   +----|      +------+ |-------------+
        |                    |      |IP+VRF| |
        |                    | +----|      | |
        |                    | |    +------+ |
        |     +----+         | |MAC+VRF|     |
        +-----|NVE2|---------| +-------+     |
              +----+         +---------------+
                |           (GW+Composite) PE2
               TS2
```

Figure 8 Gateway and Composite combined functions - example

In the example above, PE1 and PE2 MUST follow the procedures described in sections 7 and 8. Compared to section 8, PE1 and PE2 now need to also re-advertise prefixes from EVPN to EVPN, in addition to re-advertising prefixes from EVPN to IPVPN.

10. Conclusion

This document describes the procedures required in PEs that use EVPN and another Inter-Subnet-Forwarding SAFI to import and export IP Prefixes for a given tenant. In particular, this document defines:

- A route selection algorithm so that a PE can determine what path to choose between EVPN paths and other ISF SAFI paths.

- A new BGP Path attribute called D-PATH that provides loop protection and visibility on the Domains a particular route has traversed.

- The way Path attributes should be propagated between EVPN and another ISF SAFI.

- The procedures that must be followed on Interworking PEs that behave as Composite PEs, Gateway PEs or a combination of both.

The above procedures provide an operator with the required tools to build large tenant networks that may span multiple domains, use different ISF SAFIs to handle IP Prefixes, in a deterministic way and with routing loop protection.

11. Conventions used in this document

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

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

12. Security Considerations

This section will be added in future versions.

13. IANA Considerations

This document defines a new BGP path attribute known as the BGP Domain Path (D-PATH) attribute and requests IANA to assign a new attribute code type from the BGP Path Attributes registry.

14. References

14.1. Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

14.2. Informative References

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[IP-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-04, February, 2017.

[INTER-SUBNET] Sajassi et al., "IP Inter-Subnet Forwarding in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-03.txt, work in progress, February, 2017

[ENCAP-ATT] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-03.txt, work in progress, November, 2016.

15. Acknowledgments

16. Contributors

17. Authors' Addresses

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

Ali Sajassi (editor)
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
EMail: sajassi@cisco.com

Eric C. Rosen
Juniper Networks, Inc.
EMail: erosen@juniper.net

John Drake
Juniper Networks, Inc.
EMail: jdrake@juniper.net

Wen Lin
Juniper Networks, Inc.
EMail: wlin@juniper.net

Jim Uttaro
AT&T
Email: ju1738@att.com

Adam Simpson
Nokia
Email: adam.1.simpson@nokia.com

BESS Workgroup   J. Rabadan, Ed.
Internet Draft   Nokia
Intended status: Standards Track   A. Sajassi, Ed.
                                                               Cisco

E. Rosen
J. Drake
W. Lin
Juniper

J. Uttaro
AT&T

A. Simpson
Nokia

Expires: July 11, 2019   January 7, 2019

EVPN Interworking with IPVPN
draft-rabadan-sajassi-bess-evpn-ipvpn-interworking-02

Abstract

EVPN is used as a unified control plane for tenant network intra and inter-subnet forwarding. When a tenant network spans not only EVPN domains but also domains where IPVPN provides inter-subnet forwarding, there is a need to specify the interworking aspects between both EVPN and IPVPN domains, so that the end to end tenant connectivity can be accomplished. This document specifies how EVPN should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families for inter-subnet forwarding.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction and Problem Statement

EVPN is used as a unified control plane for tenant network intra and inter-subnet forwarding. When a tenant network spans not only EVPN domains but also domains where IPVPN provides inter-subnet forwarding, there is a need to specify the interworking aspects between both EVPN and IPVPN domains, so that the end to end tenant connectivity can be accomplished. This document specifies how EVPN should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families for inter-subnet forwarding.

EVPN supports the advertisement of IPv4 or IPv6 prefixes in two different route types:

- Route Type 2 - MAC/IP route (only for /32 and /128 host routes), as described by [INTER-SUBNET].

- Route Type 5 - IP Prefix route, as described by [IP-PREFIX].

When interworking with other BGP address families (AFIs/SAFIs) for inter-subnet forwarding, the IP prefixes in those two EVPN route types must be propagated to other domains using different SAFIs. Some aspects of that propagation must be clarified. Examples of these aspects or procedures across BGP families are: route selection, loop prevention or BGP Path attribute propagation. The Interworking PE concepts are defined in section 2, and the rest of the document describes the interaction between Interworking PEs and other PEs for end-to-end inter-subnet forwarding.

2. Terminology and Interworking PE Components

This section summarizes the terminology related to the "Interworking PE" concept that will be used throughout the rest of the document.

```
      +-------------------------------------------------------------+
      |                                                             |
      |              +------------------+           Interworking PE |
      | Attachment   | +------------------+                         |
      | Circuit(AC1) | |  +----------+    |                MPLS/NVO tnl
   -----------------------*Bridge    |    |                    +------
      |              | |  |Table(BT1)|    |    +-----------+  / \     \
   MPLS/NVO tnl +-------->|          *---------*           |<--> | Eth |
     -------+   |    | |  |Eth-Tag x +    |IRB1|           |  \ /     /
    / Eth  / \<-+    | |  +----------+    |    |           |   +------
   |      |   |      | |     ...          |    |  IP-VRF1  |        |
    \      \ /<-+    | |  +----------+    |    |  RD2/RT2  |MPLS/NVO tnl
     -------+   |    | |  |Bridge    |    |    |           |   +------
      |         +-------->|Table(BT2)|    |IRB2|           |  / \     \
      |              | |  |          *---------*           |<--> | IP  |
   -----------------------*Eth-Tag y |    |    +-----*-----+  \ /     /
      |  AC2         | |  +----------+    |       AC3|         +------
      |              | |   MAC-VRF1       |          |              |
      |              +-+   RD1/RT1        |          |              |
      |                +------------------+          |  SAFIs       |
      |                                              |  1     +---+ |
   --------------------------------------------------+  128   |BGP| |
      |                                                 EVPN  +---+ |
      |                                                             |
      +-------------------------------------------------------------+
```

Figure 1 EVPN-IPVPN Interworking PE

- o ISF SAFI: Inter-Subnet Forwarding (ISF) SAFI is a MP-BGP Sub-Address Family that advertises reachability for IP prefixes and can be used for inter-subnet forwarding within a given tenant network. The ISF SAFIs are 1 (including IPv4 and IPv6 AFIs), 128 (including IPv4 and IPv6 AFIs) and 70 (EVPN, including only AFI 25).

- o ISF route: a route for a given prefix whose ISF SAFI may change as it transits different domains.

- o IP-VRF: an IP Virtual Routing and Forwarding table, as defined in [RFC4364]. It is also the instantiation of an IPVPN in a PE. Route Distinguisher and Route Target(s) are required properties of an IP-VRF.

- o MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [RFC7432]. It is also the instantiation of an EVI (EVPN Instance) in a PE. Route Distinguisher and Route Target(s) are required properties and they are normally different than the ones defined in the associated IP-VRF.

o BT: a Bridge Table, as defined in [RFC7432]. A BT is the instantiation of a Broadcast Domain in a PE. When there is a single Broadcast Domain in a given EVI, the MAC-VRF in each PE will contain a single BT. When there are multiple BTs within the same MAC-VRF, each BT is associated to a different Ethernet Tag. The EVPN routes specific to a BT, will indicate which Ethernet Tag the route corresponds to.

Example: In Figure 1, MAC-VRF1 has two BTs: BT1 and BT2. Ethernet Tag x is defined in BT1 and Ethernet Tag y in BT2.

o AC: Attachment Circuit or logical interface associated to a given BT or IP-VRF. To determine the AC on which a packet arrived, the PE will examine the combination of a physical port and VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).

Example: In Figure 1, AC1 is associated to BT1, AC2 to BT2 and AC3 to IP-VRF1.

o IRB: Integrated Routing and Bridging interface. It refers to the logical interface that connects a BT to an IP-VRF and allows to forward packets with destination in a different subnet.

o MPLS/NVO tnl: It refers to a tunnel that can be MPLS or NVO-based (Network Virtualization Overlays) and it is used by MAC-VRFs and IP-VRFs. Irrespective of the type, the tunnel may carry an Ethernet or an IP payload. MAC-VRFs can only use tunnels with Ethernet payloads (setup by EVPN), whereas IP-VRFs can use tunnels with Ethernet (setup by EVPN) or IP payloads (setup by EVPN or IPVPN). IPVPN-only PEs have IP-VRFs but they cannot send or receive traffic on tunnels with Ethernet payloads.

Example: Figure 1 shows an MPLS/NVO tunnel that is used to transport Ethernet frames to/from MAC-VRF1. The PE determines the MAC-VRF and BT the packets belong to based on the EVPN label (MPLS or VNI). Figure 1 also shows two MPLS/NVO tunnels being used by IP-VRF1, one carrying Ethernet frames and the other one carrying IP packets.

o RT-2: Route Type 2 or MAC/IP route, as per [RFC7432].

o RT-5: Route Type 5 or IP Prefix route, as per [IP-PREFIX].

o Domain: Two PEs are in the same domain if they are attached to the same tenant and the packets between them do not require a data path IP lookup (in the tenant space) in any intermediate router. A gateway PE is always configured with multiple Domain-IDs.

Example 1: Figure 4 depicts an example where TS1 and TS2 belong to the same tenant, and they are located in different Data Centers that are connected by gateway PEs (see the gateway PE definition later). These gateway PEs use IPVPN in the WAN. When TS1 sends traffic to TS2, the intermediate routers between PE1 and PE2 require a tenant IP lookup in their IP-VRFs so that the packets can be forwarded. In this example there are three different domains. The gateway PEs connect the EVPN domains to the IPVPN domain.

```
                           GW1------------GW3
                         +------+       +------+
           +-------------|IP-VRF|       |IP-VRF|-------------+
          PE1            +------+       +------+            PE2
        +------+   DC1      |     WAN      |     DC2     +------+
    TS1-|IP-VRF|   EVPN     |    IPVPN     |     EVPN    |IP-VRF|-TS2
        +------+           GW2            GW4            +---+--+
           |             +------+       +------+             |
           +-------------|IP-VRF|       |IP-VRF|-------------+
                         +------+       +------+
                            +--------------+
               DOMAIN 1         DOMAIN 2       DOMAIN 3
           <---------------> <------------> <---------------->
```

Figure 4 Multiple domain DCI example

Example 2: Figure 5 illustrates a similar example, but PE1 and PE2 are now connected by a BGP-LU (BGP Labeled Unicast) tunnel, and they have a BGP peer relationship for EVPN. Contrary to Example 1, there is no need for tenant IP lookups on the intermediate routers in order to forward packets between PE1 and PE2. Therefore, there is only one domain in the network and PE1/PE2 belong to it.

```
                                EVPN
           <------------------------------------------------->
                                BGP-LU
           <------------------------------------------------->

                          ASBR------------ASBR
                         +------+       +------+
           +-------------|      |       |      |-------------+
          PE1            +------+       +--+---+            PE2
        +------+   DC1      |     WAN      |     DC2     +------+
    TS1-|IP-VRF|   EVPN     |              |     EVPN    |IP-VRF|-TS2
        +------+          ASBR            ASBR           +---+--+
           |             +------+       +------+             |
           +-------------|      |       |      |-------------+
                         +------+       +------+
                            +--------------+

           <--------------------DOMAIN-1--------------------->
```

Figure 5 Single domain DCI example

- Regular Domain: a domain in which a single control plane, IPVPN or EVPN, is used and which is composed of regular PEs, see below. In Figures 4 and 5, above, all domains are regular domains.

- Composite Domain: a domain in which multiple control planes, IPVPN and EVPN, are used and which is composed of regular PEs, see below, and composite PEs, see below.

- Regular PE: a PE that is attached to a domain, either regular or composite, and which uses one of the control plane protocols (IPVPN or EVPN) operating in the domain.

- Interworking PE: a PE that may advertise a given prefix with an EVPN ISF route (RT-2 or RT-5) and/or an IPVPN ISF route. An interworking PE has one IP-VRF per tenant, and one or multiple MAC-VRFs per tenant. Each MAC-VRF may contain one or more BTs, where each BT may be attached to that IP-VRF via IRB. There are two types of Interworking PEs: composite PEs and gateway PEs. Both PE functions can be independently implemented per tenant and they may both be implemented for the same tenant.

  Example: Figure 1 shows an interworking PE of type gateway, where ISF SAFIs 1, 128 and 70 are enabled. IP-VRF1 and MAC-VRF1 are instantiated on the PE, and together provide inter-subnet forwarding for the tenant.

o Composite PE: an interworking PE that is attached to a composite domain and which advertises a given prefix to an IPVPN peer with an IPVPN ISF route, to an EVPN peer with an EVPN ISF route, and to a route reflector with both an IPVPN and EVPN ISF route. A composite PE performs the procedures of Sections 5 and 6.

Example: Figure 2 shows an example where PE1 is a composite PE since PE1 has EVPN and another ISF SAFI enabled to the same route-reflector, and PE1 advertises a given IP prefix IPn/x twice, one using EVPN and another one using ISF SAFI 128. PE2 and PE3 are not composite PEs.

```
                                   +---+
                                   |PE2|
                                   +---+
                                    ^
                                    |EVPN
                       IW    EVPN   v
                      +---+  IPVPN ++-+       +---+
                      |PE1| <----> |RR| <---> |PE3|
                      +---+        +--+ IPVPN +---+
                    Composite
```

Figure 2 Interworking composite PE example

o Gateway PE: an interworking PE that is attached to two domains, each either regular or composite, and which, based on configuration, does one of the following:

- Propagates the same control plane protocol, either IPVPN or EVPN, between the two domains.

- Propagates an ISF route with different ISF SAFIs between the two domains. E.g., propagate an EVPN ISF route in one domain as an IPVPN ISF route in the other domain and vice versa. A gateway PE performs the procedures of Sections 3, 4, 5 and 7.

  A gateway PE is always configured with multiple Domain-IDs. The Domain-ID is encoded in the Domain Path Attribute (D-PATH), and advertised along with EVPN and other ISF SAFI routes. Section 3 describes the D-PATH attribute.

  Example: Figure 3 illustrates an example where PE1 is a gateway PE since the EVPN and IPVPN SAFIs are enabled on different BGP peers, and a given local IP prefix IPn/x is sent to both BGP peers for the same tenant. PE2 and PE1 are in one domain and PE3 and PE1 are in another domain.

```
                       IW
         +---+ EVPN   +---+ IPVPN  +---+
         |PE2| <----> |PE1| <----> |PE3|
         +---+        +---+        +---+
                     Gateway
```

Figure 3 Interworking gateway PE example

- o Composite/Gateway PE: an interworking PE that is both a composite PE and a gateway PE that is attached to two domains, one regular and one composite, and which does the following:

  - Propagates an ISF route, either IPVPN or EVPN, from the regular domain into the composite domain. Within the composite domain it acts as a composite PE.

  - Propagates an ISF route, either IPVPN or EVPN, from the composite domain into the regular domain. Within the regular domain it is propagated as an ISF route using the ISF SAFI for that domain.

    This is particularly useful when a tenant network is attached to both IPVPN and EVPN domains, any-to-any connectivity is required, and end-to-end control plane consistency, when possible, is desired.

    It would be instantiated by attaching the disparate, regular IPVPN and EVPN domains via these PEs to a central composite domain.

3. Domain Path Attribute (D-PATH)

The BGP Domain Path (D-PATH) attribute is an optional and transitive BGP path attribute.

Similar to AS_PATH, D-PATH is composed of a length field followed by a sequence of Domain segments, where each domain segment is represented by <DOMAIN-ID:ISF_SAFI_TYPE>.

- o The length field is a 1-octet field, containing the number of domain segments.

- o DOMAIN-ID is a 6-octet field that represents a domain. It is composed of a 4-octet Global Administrator sub-field and a 2-octet Local Administrator sub-field. The Global Administrator sub-field MAY be filled with an Autonomous System Number (ASN), an IPv4 address, or any value that guarantees the uniqueness of the DOMAIN-ID when the tenant network is connected to multiple Operators.

- o ISF_SAFI_TYPE is a 1-octet field that indicates the Inter-Subnet Forwarding SAFI type in which a route was advertised in the DOMAIN. The following types are valid in this document:

Value	Type
1	SAFI 1
70	EVPN
128	SAFI 128

About the BGP D-PATH attribute:

a) Identifies the sequence of domains, each identified by a <DOMAIN-ID:ISF_SAFI_TYPE> through which a given ISF route has passed.

- This attribute list may contain zero, one or more entries.

- The first entry in the list (leftmost) is the <DOMAIN-ID:ISF_SAFI_TYPE> from which a gateway PE is propagating an ISF route. The last entry in the list (rightmost) is the <DOMAIN-ID:ISF_SAFI_TYPE> from which a gateway PE received an ISF route without a D-PATH attribute. Intermediate entries in the list are domains that the ISF route has transited.

- As an example, an ISF route received with a D-PATH attribute of {<6500:2:IPVPN>,<6500:1:EVPN>} indicates that the ISF route was originated in EVPN domain 6500:1, and propagated into IPVPN domain 6500:2.

b) It is added/modified by a gateway PE when propagating an update to a different domain:

- A gateway PE's IP-VRF, that connects two domains, belongs to two DOMAIN-IDs, e.g. 6500:1 for EVPN and 6500:2 for IPVPN.

- Whenever a prefix arrives at a gateway PE in a particular ISF SAFI route, if the gateway PE needs to export that prefix to a BGP peer, the gateway PE will prepend a <DOMAIN-ID:ISF_SAFI_TYPE> segment to the list of segments in the received D-PATH.

- For instance, in an IP-VRF configured with DOMAIN-IDs 6500:1 for EVPN and 6500:2 for IPVPN, if an EVPN route for prefix P is received and P installed in the IP-VRF, the IPVPN route for P that is exported to an IPVPN peer will prepend the segment <6500:1:EVPN> to the previously received D-PATH attribute. Likewise, IP-VRF prefixes that are received from IP-VPN, will be exported to EVPN peers with the additional segment <6500:2:IPVPN>.

- In the above example, if the EVPN route is received without D-PATH, the gateway PE will add the D-PATH attribute with segment <6500:1:EVPN> when re-advertising to domain 6500:2.

- Within the originating domain, the update does not contain a D-PATH attribute because the update has not passed through a gateway PE yet.

c) The gateway PE MUST NOT add the D-PATH attribute to ISF routes generated for IP-VRF prefixes that are not learned via any ISF SAFI, for instance, local prefixes.

d) An ISF route received by a gateway PE with a D-PATH attribute that contains one or more of its locally configured domains for the IP-VRF is considered to be a looped ISF route and MUST be dropped.

e) The number of domain segments in the D-PATH attribute indicates the number of gateway PEs that the ISF route update has transited.

3.1. D-PATH and Loop Prevention

The D-PATH attribute is used to prevent loops in interworking PE networks. For instance, in the example of Figure 4, gateway GW1 receives TS1 prefix in two different ISF routes:

o In an EVPN RT-5 with next-hop PE1 and no D-PATH attribute.

o In a SAFI 128 route with next-hop GW2 and D-PATH = (6500:1:EVPN), assuming that DOMAIN-ID for domain 1 is 6500:1.

Gateway GW1 flags the SAFI 128 route as a loop, and does not re-advertise it to the EVPN neighbors since the route includes the GW1’s local domain.

In general, any interworking PE that imports an ISF route MUST flag the route as "looped" if its D-PATH contains a <DOMAIN-

ID:ISF_SAFI_TYPE> segment, where DOMAIN-ID matches a local DOMAIN-ID in the tenant IP-VRF.

4. BGP Path Attribute Propagation across ISF SAFIs

Based on configurations a gateway PE is required to propagate an ISF route with different ISF SAFIs between two domains. This requires a definition of what a gateway PE is to do with Path attributes attached to the ISF route that it is propagating.

4.1. No-Propagation-Mode

This is the default mode of operation. In this mode, the gateway PE will simply re-initialize the Path Attributes when propagating an ISF route, as though it would for direct or local IP prefixes. This model may be enough in those use-cases where the EVPN domain is considered an "abstracted" CE and remote IPVPN/IP PEs don't need to consider the original EVPN Attributes for path calculations.

Since this mode of operation does not propagate the D-PATH attribute either, redundant gateway PEs are exposed to routing loops. Those loops may be resolved by policies and the use of other attributes, such as the Route Origin extended community [RFC4360], however not all the loop situations may be solved.

4.2. Uniform-Propagation-Mode

In this mode, the gateway PE simply keeps accumulating or mapping certain key commonly used Path Attributes when propagating an ISF route. This mode is typically used in networks where EVPN and IPVPN SAFIs are used seamlessly to distribute IP prefixes.

The following rules MUST be observed by the gateway PE when propagating Path Attributes:

o The gateway PE imports an ISF route in the IP-VRF and stores the original Path Attributes. The following set of Path Attributes SHOULD be propagated by the gateway PE to other ISF SAFIs (other Path Attributes SHOULD NOT be propagated):

  - AS_PATH
  - D-PATH
  - IBGP-only Path Attributes: LOCAL_PREF, ORIGINATOR_ID, CLUSTER_ID
  - MED
  - AIGP
  - Communities, (non-EVPN) Extended Communities and Large Communities

- o When propagating an ISF route to a different ISF SAFI and IBGP peer, the gateway PE SHOULD copy the AS_PATH of the originating family and add it to the destination family without any modification. When re-advertising to a different ISF SAFI and EBGP peer, the gateway PE SHOULD copy the AS_PATH of the originating family and prepend the IP-VRF's AS before sending the route.

- o When propagating an ISF route to IBGP peers, the gateway PE SHOULD copy the IBGP-only Path Attributes from the originating SAFI to the re-advertised route.

- o Communities, non-EVPN Extended Communities and Large Communities SHOULD be copied by the gateway PE from the originating SAFI route.

4.3. Aggregation of Routes and Path Attribute Propagation

Instead of propagating a high number of (host) ISF routes between ISF SAFIs, a gateway PE that receives multiple ISF routes of one ISF SAFI MAY choose to propagate a single ISF aggregate route with a different ISF SAFI. In this document, aggregation is used to combine the characteristics of multiple ISF routes of the same ISF SAFI in such way that a single aggregate ISF route of a different ISF SAFI can be propagated. Aggregation of multiple ISF routes of one ISF SAFI into an aggregate ISF route of a different ISF SAFI is only done by a gateway PE.

Aggregation on gateway PEs may use either the No-Propagation-Mode or the Uniform-Propagation-Mode explained in Sections 4.1. and 4.2, respectively.

When using Uniform-Propagation-Mode, Path Attributes of the same type code MAY be aggregated according to the following rules:

- o AS_PATH is aggregated based on the rules in [RFC4271]. The gateway PEs SHOULD NOT receive AS_PATH attributes with path segments of type AS_SET [RFC6472]. Routes received with AS_PATH attributes including AS_SET path segments MUST NOT be aggregated.

- o ISF routes that have different attributes of the following type codes MUST NOT be aggregated: D-PATH, LOCAL_PREF, ORIGINATOR_ID, CLUSTER_ID, MED or AIGP.

- o The Community, Extended Community and Large Community attributes of the aggregate ISF route MUST contain all the Communities/Extended Communities/Large Communities from all of the aggregated ISF routes.

Assuming the aggregation can be performed (the above rules are applied), the operator should consider aggregation to deal with scaled tenant networks where a significant number of host routes exists. For a example, large Data Centers.

5. Route Selection Process between EVPN and other ISF SAFIs

A PE may receive an IP prefix in ISF routes with different ISF SAFIs, from the same or different BGP peer. It may also receive the same IP prefix (host route) in an EVPN RT-2 and RT-5. A route selection algorithm across all ISF SAFIs is needed so that:

o Different gateway and composite PEs have a consistent and deterministic view on how to reach a given prefix.

o Prefixes advertised in EVPN and other ISF SAFIs can be compared based on path attributes commonly used by operators across networks.

o Equal Cost Multi-Path (ECMP) is allowed across EVPN and other ISF SAFI routes.

For a given prefix advertised in one or more non-EVPN ISF routes, the BGP best path selection procedure will produce a set of "non-EVPN best paths". For a given prefix advertised in one or more EVPN ISF routes, the BGP best path selection procedure will produce a set of "EVPN best paths". To support IP/EVPN interworking, it is then necessary to run a tie-breaking selection algorithm on the union of these two sets. This tie-breaking algorithm begins by considering all EVPN and other ISF SAFI routes, equally preferable routes to the same destination, and then selects routes to be removed from consideration. The process terminates as soon as only one route remains in consideration.

The route selection algorithm must remove from consideration the routes following the rules and the order defined in [RFC4271], with the following exceptions and in the following order:

1- Immediately after removing from consideration all routes that are not tied for having the highest Local Preference, any routes that do not have the shortest D-PATH are also removed from consideration. Routes with no D-PATH are considered to have a zero-length D-PATH.

2- Then regular [RFC4271] selection criteria is followed.

3- At the end of the selection algorithm, if at least one route still under consideration is an RT-2 route, remove from consideration any RT-5 routes.

4- Steps 1-3 could possibly leave Equal Cost Multi-Path (ECMP) between IP and EVPN paths. By default, the EVPN path is considered (and the IP path removed from consideration). However, if ECMP across ISF SAFIs is enabled by policy, and an "IP path" and an "EVPN path" remain at the end of step 3, both path types will be used.

Example 1 - PE1 receives the following routes for IP1/32, that are candidate to be imported in IP-VRF-1:

```
{SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=(100,200)}
{SAFI=EVPN, RT-5, Local-Pref=100, AS-Path=(100,200)}
{SAFI=128, Local-Pref=100, AS-Path=(100,200)}
```

Selected route: {SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=100,200] (due to step 3, and no ECMP)

Example 2 - PE1 receives the following routes for IP2/24, that are candidate to be imported in IP-VRF-1:

```
{SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-Path=(100,200),
MED=10}
{SAFI=128, D-PATH=(6500:1:EVPN,6500:2:IPVPN), AS-Path=(200),
MED=200}
```

Selected route: {SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-Path=(100,200), MED=10} (due to step 1)

6. Composite PE Procedures

As described in Section 2, composite PEs are typically used in tenant networks where EVPN and IPVPN are both used to provide inter-subnet forwarding within the same composite domain.

Figure 6 depicts an example of a composite domain, where PE1/PE2/PE4 are composite PEs (they support EVPN and IPVPN ISF SAFIs on their peering to the Route Reflector), and PE3 is a regular IPVPN PE.

```
                +-----------------------------------+
                |                                   |
                |        MPLS                 IPVPN PE3
                |        Network              +----------+ IP3/24
                |                     IPVPN   |+------+  |   +---+
                |                      +----->||IP-VRF|------|CE3|
           Composite PE1               |      |+------+  |   +---+
          +---------------+            |      +----------+
          |      +------+ |  EVPN      v             |
          |      |IP-VRF| |  IPVPN   +--+            |
          | +----|      | | <------> |RR|            |
   +---+  | |    +------+ |          +--+         Composite PE4
   |CE2|----|MAC-VRF|     |          ^  ^         +---------+ IP4/24
   +---+  | +-------+     |    EVPN  |  | EVPN    |+------+ |   +---+
          +---|-----------+    IPVPN |  | IPVPN   ||IP-VRF|-----|CE4|
              |  |              +----+  +-------->|+------+ |   +---+
       IP1/24 |  |                v               +---------+
       +---+  |  |    +---------------+              |
       |CE1|--+  +----|      +------+ +--------------+
       +---+          |      |IP-VRF| |
         |            | +----|      | |
         |            | |    +------+ |
         +--------------|MAC-VRF|     |
                      | +-------+     |
                      +---------------+
                         Composite PE2
```

Figure 6 Composite PE example

In a composite domain with composite and regular PEs:

o The composite PEs advertise the same IP prefixes in each ISF SAFI to the RR. For example, in Figure 6, the prefix IP1/24 is advertised by PE1 and PE2 to the RR in two separate NLRIs, one for AFI/SAFI 1/128 and another one for EVPN.

o The RR does not forward EVPN routes to PE3 (since the RR does not have the EVPN SAFI enabled on its BGP session to PE3), whereas the IPVPN routes are forwarded to all the PEs.

o PE3 receives only the IPVPN route for IP1/24 and resolves the BGP next-hop to an MPLS tunnel (with IP payload) to PE1 and/or PE2.

o Composite PE4 receives IP1/24 encoded in EVPN and another ISF SAFI route (EVPN RT-5 and IPVPN). The route selection follows the procedures in Section 5. Assuming an EVPN route is selected, PE4 resolves the BGP next-hop to an MPLS tunnel (with Ethernet or IP payload) to PE1 and/or PE2. As described in Section 2, two EVPN PEs may use tunnels with Ethernet or IP payloads to connect their IP-VRFs, depending on the [IP-PREFIX] model implemented. If some attributes are modified so that the route selection process (Section 5) results in PE4 selecting the IPVPN path instead of the EVPN path, the operator should be aware that the EVPN advanced forwarding features, e.g. recursive resolution to overlay indexes, will be lost for PE4.

- The other composite PEs (PE1 and PE2) receive also the same IP prefix via EVPN and IPVPN SAFIs and they also follow the route selection in Section 5.

- When a given route has been selected as the route for a particular packet, the transmission of the packet is done according to the rules for that route's AFI/SAFI.

- It is important to note that in composite domains, such as the one in Figure 6, the EVPN advanced forwarding features will only be available to composite and EVPN PEs (assuming they select an RT-5 to forward packets for a given IP prefix), and not to IPVPN PEs. For example, assuming PE1 sends IP1/24 in an EVPN and an IPVPN route and the EVPN route is the best one in the selection, the recursive resolution of the EVPN RT-5s can only be used in PE2 and PE4 (composite PEs), and not in PE3 (IPVPN PE). As a consequence of this, the indirection provided by the RT5's recursive resolution and its benefits in a scaled network, will not be available in all the PEs in the network.

7. Gateway PE Procedures

Section 2 defines a gateway PE as an Interworking PE that advertises IP prefixes to different BGP peers, using EVPN to one BGP peer and another ISF SAFI to another BGP peer. Examples of gateway PEs are Data Center gateways connecting domains that make use of EVPN and other ISF SAFIs for a given tenant. Figure 7 illustrates this use-case, in which PE1 and PE2 (and PE3/PE4) are gateway PEs interconnecting domains for the same tenant.

```
    <----EVPN---->    <----------IPVPN--------->   <----EVPN---->
      6500:1:EVPN             6500:2:IPVPN           6500:3:EVPN
 <DOMAIN-ID:ISF_SAFI_TYPE>
                       +-----------------------+
                Gateway PE1              Gateway PE3
                +----------+             +----------+
    +-----------|+------+  |  MPLS tnls  |+------+  |-------------+
    |           ||IP-VRF|  |             ||IP-VRF|  |             |
  PE5           |+------+  |             |+------+  |           PE6
 +------+       +----------+             +----------+          +------+
 |IP-VRF| NVO tnls |   |                       |  |  NVO tnls  |IP-VRF|
 |      |          |   |                       |  |            |      |
 +------+       +----------+             +----------+          +------+
 IP1/24-->      |+------+  |             |+------+  |             |
    |           ||IP-VRF|  |             ||IP-VRF|  |             |
    +-----------|+------+  |             |+------+  |-------------+
                +----------+             +----------+
                Gateway PE2   +------+   Gateway PE4
                      +-------|IP-VRF|---------+
                              |      |
                              +------+
                                PE7
```

Figure 7 Gateway PE example

The gateway PE procedures are described as follows:

- A gateway PE that imports an ISF SAFI-x route to prefix P in an IP-VRF, MUST export P in ISF SAFI-y if:

  1. P is installed in the IP-VRF (hence the SAFI-x route is the best one for P) and

  2. PE has a BGP peer for SAFI-y (enabled for the same IP-VRF) and

  3. Either x or y is EVPN.

     In the example of Figure 7, gateway PE1 and PE2 receive an EVPN RT-5 with IP1/24, install the prefix in the IP-VRF and re-advertise it using SAFI 128.

- ISF SAFI routes advertised by a gateway PE MUST include a D-PATH attribute, so that loops can be detected in remote gateway PEs. When a gateway PE propagates an IP prefix between EVPN and another ISF SAFI, it MUST prepend a <DOMAIN-ID:ISF_SAFI_TYPE> to the received D-PATH attribute. The DOMAIN-ID and ISF_SAFI_TYPE fields refer to the domain over which the gateway PE received the IP prefix and the ISF SAFI of the route, respectively. If the received

IP prefix route did not include any D-PATH attribute, the gateway IP MUST add the D-PATH when readvertising. The D-PATH in this case will have only one segment on the list, the <DOMAIN-ID:ISF_SAFI_TYPE> of the received route.

In the example of Figure 7, gateway PE1/PE2 receive the EVPN RT-5 with no D-PATH attribute since the route is originated at PE5. Therefore PE1 and PE2 will add the D-PATH attribute including <DOMAIN-ID:ISF_SAFI_TYPE> = <6500:1:EVPN>. Gateways PE3/PE4 will propagate the route again, now prepending their <DOMAIN-ID:ISF_SAFI_TYPE> = <6500:2:IPVPN>. PE6 receives the EVPN RT-5 routes with D-PATH = {<6500:2:IPVPN>,<6500:1:EVPN>} and can use that information to make BGP path decisions.

- The gateway PE MAY use the Route Distinguisher of the IP-VRF to readvertise IP prefixes in EVPN or the other ISF SAFI.

- The label allocation used by each gateway PE is a local implementation matter. The IP-VRF advertising IP prefixes for EVPN and another ISF SAFI may use a label per-VRF, per-prefix, etc.

- The gateway PE MUST be able to use the same or different set of Route Targets per ISF SAFI on the same IP-VRF. In particular, if different domains use different set of Route Targets for the same tenant, the gateway PE MUST be able to import and export routes with the different sets.

- Even though Figure 7 only shows two domains per gateway PE, the gateway PEs may be connected to more than two domains.

- There is no limitation of gateway PEs that a given IP prefix can pass through until it reaches a given PE.

- It is worth noting that an IP prefix that was originated in an EVPN domain but traversed a different ISF SAFI domain, will lose EVPN-specific attributes that are used in advanced EVPN procedures. For example, even if PE1 advertises IP1/24 along with a given non-zero ESI (for recursive resolution to that ESI), when PE6 receives the IP prefix in an EVPN route, the ESI value will be zero. This is because the route traverses an ISF SAFI domain that is different than EVPN.

8. Interworking Use-Cases

While Interworking PE networks may well be similar to the examples described in Sections 6 and 7, in some cases a combination of both functions may be required. Figure 8 illustrates an example where the gateway PEs are also composite PEs, since not only they need to re-advertise IP prefixes from EVPN routes to another ISF SAFI routes, but they also need to interwork with IPVPN-only PEs in a domain with a mix of composite and IPVPN-only PEs.

```
                       +-----------------------------------+
                                                           |
                       |        MPLS                IPVPN PE3
                       |        Network             +---------+
                       |                     IPVPN  |+------+ |
                       |                      +----->||IP-VRF|---TS3
                (GW+composite) PE1            |     |+------+ |
                 +---------------+            |     +---------+
                 |      +------+ |  EVPN      v            |
                 |      |IP+VRF| |  IPVPN   +-++           |
                 | +----|      | | <------> |RR|           |
        +--------| |    +------+ |          +--+        Composite PE4
        |        | |MAC+VRF|     |          ^  ^         +---------+
        |        | +-------+     |    EVPN  |  | EVPN    |+------+ |
     +----+      +---------------+    IPVPN |  | IPVPN   ||IP-VRF|---TS4
 TS1-|NVE1|             |              +----+  +-------->|+------+ |
     +----+             |              v                 +---------+
        |    EVPN DC    |    +---------------+             |
        |    NVO tnls   +----|      +------+ |-------------+
        |                    |      |IP+VRF| |
        |                    | +----|      | |
        |                    | |    +------+ |
        |     +----+         | |MAC+VRF|     |
        +-----|NVE2|---------| +-------+     |
              +----+         +---------------+
                |           (GW+composite) PE2
               TS2
```

Figure 8 Gateway and composite combined functions - example

In the example above, PE1 and PE2 MUST follow the procedures described in Sections 6 and 7. Compared to section 7, PE1 and PE2 now need to also propagate prefixes from EVPN to EVPN, in addition to propagating prefixes from EVPN to IPVPN.

It is worth noting that PE1 and PE2 will receive TS4’s IP prefix via IPVPN and RT-5 routes. When readvertising to NVE1 and NVE2, PE1 and PE2 will consider the D-PATH rules and attributes of the selected route for TS4 (Section 5 describes the Route Selection Process).

9. Conclusion

This document describes the procedures required in PEs that use EVPN and another Inter-Subnet Forwarding SAFI to import and export IP prefixes for a given tenant. In particular, this document defines:

- A route selection algorithm so that a PE can determine what path to choose between EVPN paths and other ISF SAFI paths.

- A new BGP Path attribute called D-PATH that provides loop protection and visibility on the domains a particular route has traversed.

- The way Path attributes should be propagated between EVPN and another ISF SAFI.

- The procedures that must be followed on Interworking PEs that behave as composite PEs, gateway PEs or a combination of both.

The above procedures provide an operator with the required tools to build large tenant networks that may span multiple domains, use different ISF SAFIs to handle IP prefixes, in a deterministic way and with routing loop protection.

10. Conventions used in this document

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

11. Security Considerations

This section will be added in future versions.

12. IANA Considerations

This document defines a new BGP path attribute known as the BGP Domain Path (D-PATH) attribute and requests IANA to assign a new attribute code type from the "BGP Path Attributes" subregistry under the "Border Gateway Protocol (BGP) Parameters" registry.

13. References

13.1. Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

13.2. Informative References

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[IP-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11, May, 2018.

[INTER-SUBNET] Sajassi et al., "IP Inter-Subnet Forwarding in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-05.txt, work in progress, July, 2018

[ENCAP-ATT] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10.txt, work in progress, August, 2018.

[RFC6472] Kumari, W. and K. Sriram, "Recommendation for Not Using AS_SET and AS_CONFED_SET in BGP", BCP 172, RFC 6472, DOI 10.17487/RFC6472, December 2011, <https://www.rfc-editor.org/info/rfc6472>.

14. Acknowledgments

15. Contributors

16. Authors' Addresses

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

Ali Sajassi (editor)
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
EMail: sajassi@cisco.com

Eric C. Rosen
Juniper Networks, Inc.
EMail: erosen@juniper.net

John Drake
Juniper Networks, Inc.
EMail: jdrake@juniper.net

Wen Lin
Juniper Networks, Inc.
EMail: wlin@juniper.net

Jim Uttaro
AT&T
Email: ju1738@att.com

Adam Simpson
Nokia
Email: adam.1.simpson@nokia.com

BESS Workgroup   J. Rabadan, Ed.
Internet Draft   J. Kotalwar
Intended status: Standards Track   S. Sathappan
Nokia

Z. Zhang
Juniper

A. Sajassi
Cisco

Expires: May 3, 2018   October 30, 2017

PIM Proxy in EVPN Networks
draft-skr-bess-evpn-pim-proxy-01

Abstract

Ethernet Virtual Private Networks [RFC7432] are becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. One of the goals that EVPN pursues is the reduction of flooding and the efficiency of CE-based control plane procedures in Broadcast Domains. Examples of this are Proxy ARP/ND and IGMP/MLD Proxy. This document complements the latter, describing the procedures required to minimize the flooding of PIM messages in EVPN Broadcast Domains, and optimize the IP Multicast delivery between PIM routers.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

Ethernet Virtual Private Networks [RFC7432] are becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. One of the goals that EVPN pursues is the reduction of flooding and the efficiency of CE-based control plane procedures in Broadcast Domains. Examples of this are [EVPN-PROXY-ARP-ND] for improving the efficiency of CE's ARP/ND protocols, and [EVPN-IGMP-MLD-PROXY] for IGMP/MLD protocols.

This document focuses on optimizing the behavior of PIM in EVPN Broadcast Domains and re-uses some procedures of [EVPN-IGMP-MLD-PROXY]. The reader is also advised to check out [RFC8220] to understand certain aspects of the procedures of PIM Join/Prune messages received on Attachment Circuits (ACs).

Section 2 describes the PIM Proxy procedures that the implementation should follow, including:

- o The use of EVPN to suppress the flooding of PIM Hello messages in shared Broadcast Domains. The benefit of this is twofold:
  - PIM Hello messages will ONLY be flooded to Attachment Circuits that are connected to PIM routers, as opposed to all the CEs and hosts in the Broadcast Domain.
  - Soft-state PIM Hello messages will be replaced by hard-state BGP messages that don't need to be refreshed periodically.

- o The use of EVPN to discover IGMP Queriers, while avoiding the flooding of IGMP Queries in the core.

- o The procedures to proxy PIM Join/Prune messages and replace them by hard-state EVPN routes that don't need to be refreshed periodically. By using BGP EVPN to propagate both, Hello and Join/Prune messages, we also avoid out-of-order delivery between both types of PIM messages.

- o This document also describes an EVPN based procedure so that the PIM routers connected to the shared Broadcast Domain don't need to run any PIM Assert procedure. PIM Assert procedures may be expensive for PIM routers in terms of resource consumption. With this procedure, there is no PIM Assert needed on PIM routers.

- The use of procedures similar to the ones defined in [EVPN-IGMP-MLD-PROXY] to synchronize multicast states among the PEs in the same Ethernet Segment.

Section 3 describes the interaction of PIM Proxy with IGMP Proxy PEs and Multicast Sources connected to the same EVPN Broadcast Domain.

Section 4 defines the BGP Information Model that this document requires to address the PIM Proxy procedures.

This document assumes the reader is familiar with PIM and IGMP protocols.

2. PIM Proxy Operation in EVPN Broadcast Domains

This section describes the operation of PIM Proxy in EVPN Broadcast Domains (BDs). Figure 1 depicts an EVPN Broadcast Domain defined in four PEs that are connected to PIM routers. This example will be used throughout this section and assumes both R4 and R5 are PIM Upstream Neighbors for PIM routers R1, R2 and R3 and multicast group G1. In this situation, the PIM multicast traffic flows from R4 or R5 to R1, R2 and R3. The PIM Join/Prune signaling will flow in the opposite direction. From a terminology perspective, we consider PE1 and PE2 as egress or downstream PEs, whereas PE3 and PE4 are ingress or upstream PEs.

```
          J(*,G1,IP5)
      +--+
      |R1+------>               XXXXXXXX
      +--+       +-----+     XXXX      XX  XXXXX  +-----+      +--+
                 | PE1 |XXXXX           XXXX    XX| PE3 +----> |R4|
      +--+       |     |                          |     |      +--+
      |R2+-----> +-----+                          +-----+ <----
      +--+          X                            XX         multicast
        J(*,G1,IP5) X                             XXX        (S1,G1)
                 XXX        EVPN Broadcast          XX
                 X             Domain                 X
   +--+     +-----+                                   X           RP
   |R3+---> | PE2 |                                 XX+-----+    +--+
   +--+     |     |                              XXXX | PE4 +--> |R5|
            +-----+XXXX                      XXXXX    |     |    +--+
     J(S1,G1,IP4)      X          X           X       +-----+
                       XX      XXX XX       XXX
                         XXXXXX     XXXXX XXX
```

Figure 1 - PIM Routers connected by an EVPN Broadcast Domain

It is important to note that any Router’s PIM message not explicitly specified in this document will be forwarded by the PEs normally, in the data path, as a unicast or multicast packet.

2.1. Multicast Router Discovery Procedures in EVPN

The procedures defined in this section make use of the Multicast Router Discovery (MRD) route described in section 4 and are OPTIONAL. An EVPN router not implementing this specification will transparently flood PIM Hello messages and IGMP Queries to remote PEs.

2.1.1. Discovering PIM Routers

As described in [RFC4601] for shared LANs, an EVPN Broadcast Domain may have multiple PIM routers connected to it and a single one of these routers, the DR, will act on behalf of directly connected hosts with respect to the PIM-SM protocol. The DR election, as well as discovery and negotiation of options in PIM, is performed using Hello messages. PIM Hello messages are periodically exchanged and flooded in EVPN Broadcast Domains that don’t follow this specification.

When PIM Proxy is enabled, an EVPN PE will snoop PIM Hello messages and forward them only to local ACs where PIM routers have been detected. This document assumes that all the procedures defined in

[RFC8220] to snoop PIM Hellos on local ACs and build the PIM Neighbor DB on the PEs are followed. PIM Hello messages MUST NOT be forwarded to remote EVPN PEs though.

Using Figure 1 as an example, the PIM Proxy operation for Hello messages is as follows:

1) The arrival of a new PIM Hello message at e.g. PE1 will trigger an MRD route advertisement including:
   - o The IP address and length of the multicast router that issued the Hello message. E.g. R1’s IP address and length.
   - o The DR Priority copied from the Hello DR Priority TLV.
   - o Q flag set (if the multicast router is a Querier).
   - o P flag set that indicates the router is PIM capable.

2) All other PEs import the MRD route and do the following:
   - o Add the multicast router address to the PIM Neighbor Database (PIM Nbr DB) associated to the Originator Router Address.
   - o Generate a PIM hello where the IP Source Address is the Multicast Router IP and the DR Priority is copied from the route. This PIM hello is sent to all the local ACs connected to a PIM router. For example, PE3 will send the generated hello message to R4.

3) Each PE will build its PIM Nbr DB out of the local PIM hello messages and/or remote MRD routes. The PIM hello timers and other hello parameters are not propagated in the MRD routes.

   - o The timers are handled locally by the PE and as per [RFC4601]. This is valid for the hold_time (when a PIM router or PE receives a hello message, resets the neighbor-expiry timer), and other timers.

   - o The Generation ID option is also processed locally on the PE, as well as the Generation ID changes for a given multicast router. It is not propagated in the MRD route.

   - o Procedures described in [RFC4601] are used to remove a local AC PIM router from the PIM Nbr DB. When a local router is removed from the DB, the MRD route is withdrawn. If the local router is still sending Queries, the route is updated with flags P=0 and Q=1. Upon receiving the update, the other PEs will remove the router from the PIM Nbr DB but not from the list of queriers.

4) Based on regular PIM DR election procedures (highest DR Priority or highest IP), each PE is aware of who the DR is for the BD. For more information, refer to section "3. Interaction with IGMP-snooping and Sources".

2.1.2. Discovering IGMP Queriers

In (EVPN) Broadcast Domains that are shared among not only PIM routers but also IGMP hosts, one or more PIM routers will also be configured as IGMP Queriers. The proxy Querier mechanism described in [EVPN-IGMP-MLD-PROXY] suppresses the flooding of queries on the Broadcast Domain, by using PE generated Queries from an anycast IP address.

While the proxy Querier mechanism works in most of the use-cases, sometimes it is desired to have a more transparent behavior and propagate existing multicast router IGMP Queries as opposed to "blindly" querying all the hosts from the PEs. The MRD route defined in section 4 can be used for that purpose.

When the discovered local PIM router is also sending IGMP Queries, the PE will issue an MRD route for the multicast router with both Q (IGMP Querier) and P (PIM router) flags set. Note that the PE may set both flags or only one of them, depending on the capabilities of the local router.

A PE receiving an MRD route with Q=1 will generate IGMP Query messages, using the multicast router IP address encoded in the received MRD route. If more than one IGMP Queriers exist in the EVI, the PE receiving the MRD routes with Q=1 will select the lower IP address, as per [RFC2236]. Note that, upon receiving the MRD routes with Q=1, the PE must generate IGMP Queries and forward them to all the local ACs. Other Queriers listening to these received Query messages will stop sending Queries if they are no longer the selected Querier, as per [RFC2236].

This procedure allows the EVPN PEs to act as proxy Queriers, but using the IP address of the best existing IGMP Querier in the EVPN Broadcast Domain. This can help IGMP hosts troubleshoot any issues on the IGMP routers and check their connectivity to them.

2.2. PIM Join/Prune Proxy Procedures

This section describes the procedures associated to the PIM Proxy function for Join and Prune messages. This document assumes that all the procedures defined in [RFC8220] to build multicast states on the PEs’ local ACs are followed. Figure 2 illustrates an scenario where PIM Proxy is enabled on the EVPN PEs.

```
          J(*,G1,IP5)
      +--+                                               J(*,G1,IP5)
      |R1+------>               XXXXXXXX               P(S1,G1,IP5,rpt)
      +--+       +-----+     XXXX      XX  XXXXX  +-----+      +--+
                 | PE1 |XXXXX           XXXX    XX| PE3 +----> |R4|
      +--+       |     |   SMET                   |     |      +--+
      |R2+-----> +-----+   (*,G1,IP5)             +-----+
      +--+          X        +--------->         XX
        J(*,G1,IP5) X                             XXX
                  XX                                XX
                 X                                    X  J(*,G1,IP5)
   +--+     +-----+     SMET                          X P(S1,G1,IP5,rpt)
   |R3+---> | PE2 |     (S1,G1,IP5,rpt)             XX+-----+    +--+
   +--+     |     |        +-------->            XXXX | PE4 +--> |R5|
            +-----+XXXX                      XXXXX    |     |    +--+
     P(S1,G1,IP5,rpt)  X          X           X       +-----+     RP
                       XX      XXX XX       XXX
                         XXXXXX     XXXXX XXX
```

Figure 2 - Proxy PIM Join/Prune in EVPN

PIM J/P messages are sent by the routers towards upstream sources and RPs:

- o (*,G) is used in Join/Prune messages that are sent towards the RP for the specified group.
- o (S,G) used in Join/Prune messages sent towards the specified source.
- o (S,G,rpt) is used in Join/Prune messages sent towards the RP. We refer to this as RPT message and the Prune message always precedes the Join message. The typical sequence of PIM messages (for a group) seen in a BD connecting PIM routers is the following:

  a) (*,G) Join issued by a downstream router to the RP (to join the RP Tree).
  b) (S,G) Join issued by a downstream router switching to the SPT.
  c) (S,G,rpt) Prune issued by a downstream router to the RP to prune a specific source from the RPT.
  d) (S,G) Prune issued by a downstream router no longer interested in the SPT.
  e) (S,G,rpt) Join issued by a downstream router interested (again) in the RPT for (S,G).

The Proxy PIM procedures for Join/Prune messages are summarized as follows:

1) Downstream PE procedures:

- A downstream PE will snoop PIM Join/Prune messages and won’t forward them to remote PEs.

- Triggered by the reception of the PIM Join message, a downstream PE will advertise an SMET route, including the source, group and Upstream Neighbor as received from the PIM Join message. A single SMET route is advertised per source, group, with the P flag set. As an example, in Figure 2, PE1 receives two PIM Join messages for the same source, group and Upstream Neighbor, however PE1 advertises a single SMET route.

- When the last connected router sends a PIM Prune message for a given source, group and Upstream Neighbor and the state is removed, the PE will withdraw the SMET route (note that the state is removed once the prune-pend timer expires).

- SMET routes must always be generated upon receiving a PIM Join message, irrespective of the location of the Upstream Neighbor and even if the Upstream Neighbor is local to the PE.

- A downstream PE receiving a PIM Prune (S,G,rpt) message will trigger an RPT-Prune route for the source and group. Subsequently, if the downstream PE receives a PIM Join (S,G,rpt) to cancel the previous Prune (S,G,rpt) and keep pulling the multicast traffic from the RPT, the downstream PE will withdraw the RPT-Prune route.

- PIM Timers are handled locally. If the holdtime expires for a local Join the PE withdraws the SMET route.

3) Upstream PE procedures:

- A received SMET route with P=1 will add state for the source and group and will generate a PIM Join message for the source, group that will be forwarded to all the local AC PIM routers.

- A received SMET route withdrawal will remove the state and generate a PIM Prune message for the source, group and upstream neighbor that will be forwarded to all the local AC PIM routers.

- A received RPT-Prune route for (S,G) will generate a PIM Prune (S,G,rpt) message that will be forwarded to all the local AC PIM routers.

- A received RPT-Prune withdrawal for (S,G) will generate a PIM Join (S,G,rpt) message that will be forwarded to all the local AC PIM routers.

It is important to note that, compared to a solution that does not snoop PIM messages and does not use BGP to propagate states in the core, this EVPN PIM Proxy solution will add some latency derived from the procedures described in this document.

2.3. PIM Assert Optimization

The PIM Assert process described in [RFC4601] is intense in terms of resource consumption in the PIM routers, however it is needed in case PIM routers share a multi-access transit LAN. The use of PIM Proxy for EVPN BDs can minimize and even suppress the need for PIM Assert as described in this section.

As a refresher, the PIM Assert procedures are needed to prevent two or more Upstream PIM routers from forwarding the same multicast content to the group of Downstream PIM routers sharing the same (EVPN) Broadcast Domain. This multicast packet duplication may happen in any of the following cases:

- Two or more Downstream PIM routers on the BD may issue (*,G) Joins to different upstream routers on the BD because they have inconsistent MRIB entries regarding how to reach the RP. Both paths on the RP tree will be set up, causing two copies of all the shared tree traffic to appear on the EVPN Broadcast Domain.

- Two or more routers on the BD may issue (S,G) Joins to different upstream routers on the BD because they have inconsistent MRIB entries regarding how to reach source S. Both paths on the source-specific tree will be set up, causing two copies of all the traffic from S to appear on the BD.

- A router on the BD may issue a (*,G) Join to one upstream router on the BD, and another router on the BD may issue an (S,G) Join to a different upstream router on the same BD. Traffic from S may reach the BD over both the RPT and the SPT. If the receiver behind the downstream (*,G) router doesn’t issue an (S,G,rpt) prune, then this condition would persist.

PIM does not prevent such duplicate joins from occurring; instead, when duplicate data packets appear on the same BD from different routers, these routers notice this and then elect a single forwarder. This election is performed using the PIM Assert procedure.

The issue is minimized or suppressed in this document by making sure all the Upstream PEs select the same Upstream Neighbor for a given (*,G) or (S,G) in any of the three above situations. If there is only one upstream PIM router selected and the same multicast content is not allowed to be flooded from more than one Upstream Neighbor, there will not be multicast duplication or need for Assert procedures in the EVPN Broadcast Domain.

Figure 3 illustrates an example of the PIM Assert Optimization in EVPN.

```
          J(*,G1,IP5)
      +--+                                               J(*,G1,IP5)
      |R1+------>               XXXXXXXX                 J(S1,G1,IP4)
      +--+       +-----+     XXXX      XX  XXXXX  +-----+      +--+
                 | PE1 |XXXXX           XXXX    XX| PE3 +----> |R4|
      +--+       |     |   SMET                   |     |      +--+
      |R2+-----> +-----+   (*,G1,IP5)             +-----+
      +--+          X        +--------->         XX
        J(*,G1,IP4) X                             XXX
                  XX                                XX
                 X                                    X  J(*,G1,IP5)
   +--+     +-----+     SMET                          X  J(S1,G1,IP4)
   |R3+---> | PE2 |     (S1,G1,IP4)                 XX+-----+    +--+
   +--+     |     |        +-------->            XXXX | PE4 +--> |R5|
            +-----+XXXX                      XXXXX    |     |    +--+
      J(S1,G1,IP4)     X          X           X       +-----+     RP
                       XX      XXX XX       XXX  P(S1,G1,IP5,rpt)-->
                         XXXXXX     XXXXX XXX
```

Figure 3 - Proxy PIM Assert Optimization in EVPN

2.3.1 Assert Optimization Procedures in Downstream PEs

The Downstream PEs will trigger SMET routes based on the received PIM Join messages. This is their behavior when any of the three situations described in section 2.3 occurs:

- o If the Downstream PE receives two local (*,G) Joins to different Upstream Neighbors, the PE will generate a single SMET route, selecting the highest IP address. In Figure 3, if we assume R1 issues J(*,G1,IP5) and R2 J(*,G1,IP4), PE1 will advertise an SMET route for (*,G,IP5). If PE1 had already advertised (*,G1,IP4), it would have sent an update with (*,G1,IP5). Note that the Upstream Router IP address is not part of the SMET route key, hence there is no need to withdraw the previous (*,G1,IP4).

- o In the same way, if the Downstream PE receives two local (S,G) Joins to different Upstream Neighbors, the PE will generate a single SMET route, selecting the highest IP address.

- If the Downstream PE receives a local (S,G) and a local (*,G) Joins for the same group but to different Upstream Neighbors, the PE will generate two different SMET routes (since *,G and S,G make two different route keys), keeping the original Upstream Neighbors in the SMET routes.

2.3.2 Assert Optimization Procedures in Upstream PEs

Upon receiving two or more SMET routes for the same group but different Upstream Neighbors, the Upstream PEs will follow this procedure:

1) The Upstream PE will select a unique Upstream Neighbor based on the following rules:

   a) The Upstream Neighbor encoded in a (S,G) SMET route has precedence over the Upstream Neighbor on the (*,G) SMET route for the same group. This is consistent with the Assert winner election in [RFC4601]. In the example of Figure 3, PE3 and PE4 will select IP4 as the Upstream Neighbor for (S1,G1) and (*,G1).

   b) In case the SMET routes have the same source (* or S), the higher Upstream Neighbor IP Address wins.

2) After selecting the Unique Upstream Neighbor, the PE will instruct the data path to discard any ingress multicast stream that is coming from an interface different than the selected Upstream Neighbor for the multicast group. In the example in Figure 3, PE4 will not accept G1 multicast traffic from R5.

   NOTE: when the procedure selects an Upstream Neighbor between the (S,G) and (*,G) routes, we assume that the PE’s interface that is connected to the non-selected Upstream Neighbor, is not shared with another Source for the same Group. In the example of Figure 3, this means that PE4’s AC cannot be shared by R5 and S2 for the same group G. If PE4’s AC is connected to a switch where R5 (RP) and S2 are connected, multicast traffic (S2,G) will be dropped by PE4, as per (2).

3) Then the PE will generate the corresponding local PIM messages as usual. In the example, PE3 and PE4 generate PIM Join messages for (S1,G1,IP4) and (*,G1,IP5).

4) The PE connected to the non-selected Upstream Neighbor will issue a PIM (S,G)/(*,G) Prune or a PIM (S,G,rpt) Prune to make sure the non-selected Upstream Router does not forward traffic for the group anymore. In the example, PE4 will issue a local (S1,G1,rpt) Prune message to R5, so that R5 does not forward G1 traffic.

In case of any change that impacts on the Upstream Neighbor selection for a given group G1, the upstream PEs will simply update the Upstream Neighbor selection and follow the above procedure. This mechanism prevents the multicast duplication in the EVPN Broadcast Domain and avoids PIM Assert procedures among PIM routers in the BD.

2.4. EVPN Multi-Homing and State Synchronization

PIM Join/Prune States will be synchronized across all the PEs in an Ethernet Segment by using the procedures described in [EVPN-IGMP-MLD-PROXY] and the IGMP/PIM Join Synch Route with the corresponding Flag P set. This document does not require the use of IGMP Leave Synch Routes.

In the same way, RPT-Prune States can be synchronized by using the PIM RPT-Prune Synch route. The generation and process for this route follows similar procedures as for the IGMP/PIM Join Synch Route.

In order to synchronize the PIM Neighbors discovered on an Ethernet Segment, the MRD route and its ESI value will be used. Upon receiving a Hello message on a link that is part of a multi-homed Ethernet Segment, the PE will issue an MRD route that encodes the ESI value of the AC over which the Hello was received. Upon receiving the non-zero ESI MRD route, the PEs in the same ES will add the router to their PIM Neighbor DB, using their AC on the same ES as the PIM Neighbor port. This will allow the DF on the ES to generate Hello messages for the local PIM router.

A PE that is not part of the ESI would normally receive a single non-zero ESI MRD route per multicast router. In certain transient situations the PE may receive more than one non-zero ESI MRD route for the same multicast router. The PE should recognize this and not generate additional PIM Hello messages for the local ACs.

3. Interaction with IGMP-snooping and Sources

Figure 4 illustrates an example with a multicast source, an IGMP host and a PIM router in the same EVPN BD.

```
                               XXXXX       J(*,G1)
                        XXXXXXX     +-----+      +--+
                    XXXX            | PE3 |  <---+H3|
                   X                |     |      +--+
   +------+        X    +-------->  +-----+ +--->
   |Source|     +-----+ |   S1,G1      X      S1,G1 mcast
   | S1   +---> | PE1 | +   mcast     XX
   +------+     |     |              XX      Hello
           G1   +-----+ +   S1,G1     X    <---+
                   XX   |   mcast   +-----+    +--+
                   X    +---------> | PE4 +--> |R4|
                   X                |     |    +--+
                    XX   XXX        +-----+     DR
                      XXX  XXX     XXX
                             XXXXXXX        S1,G1, mcast
```

Figure 4 - Proxy PIM interaction with local sources and hosts

When PIM routers, multicast sources and IGMP hosts coexist in the same EVPN Broadcast domain, the PEs supporting both IGMP and PIM proxy will provide the following optimizations in the EVPN BD:

- If an IGMP host and a PIM router are connected to the same BD on a PE, the PE will advertise a single SMET route per (S,G) or (*,G) irrespective of the received IGMP or PIM message. The IGMP flags can be simultaneously set along with the P flag.

- In the same way, if IGMP hosts and PIM routers are connected to the same BD and Ethernet Segment, the IGMP/PIM Join Synch route can be shared by a host and a router requesting the same multicast source and group.

- A PE connected to a Source and using Ingress Replication will forward a multicast stream (S1,G1) to all the egress PEs that advertised an SMET route for (S1,G1) and all the egress PEs that advertised an MRD route for the EVPN BD.

4. BGP Information Model

This document defines the following additional routes and requests IANA to allocate a type value in the EVPN route type registry:

+ Type TBD - Multicast Router Discovery (MRD) Route
+ Type TBD - PIM RPT-Prune Route

+ Type TBD - PIM RPT-Prune Join Synch Route

In addition, the following routes defined in [EVPN-IGMP-MLD-PROXY] are re-used and extended in this document’s procedures:

+ Type 6 - Selective Multicast Ethernet Tag Route
+ Type 7 - IGMP Join Synch Route

Where Type 7 is requested to be re-named as IGMP/PIM Join Synch Route.

4.1 Multicast Router Discovery (MRD) Route

Figure 5 shows the content of the MRD route:

```
+--------------------------------------------------+
|  RD (8 octets)                                   |
+--------------------------------------------------+
|  Ethernet Segment ID (10 octets)                 |
+--------------------------------------------------+
|  Ethernet Tag ID (4 octets)                      |
+--------------------------------------------------+
|  Originator Router Length (1 octet)              |
+--------------------------------------------------+
|  Originator Router Address (Variable)            |
+--------------------------------------------------+
|  Mcast Router Length (1 octet)                   |
+--------------------------------------------------+
|  Mcast Router Address 1 (variable)               |
+--------------------------------------------------+
|  Secondary Address List Length (1 octet)         |
+--------------------------------------------------+
|  Secondary Mcast Router Address 1 (variable)     |
+--------------------------------------------------+
|              .                                   |
|              .                                   |
|  Secondary Mcast Router Address n (variable)     |
+--------------------------------------------------+
|  DR Priority    (4 octets)                       |
+--------------------------------------------------+
|  Flags (1 octet)                                 |
+--------------------------------------------------+
```

Figure 5 Multicast Router Discovery Route

The support for this new route type is OPTIONAL. Since this new route type is OPTIONAL, an implementation not supporting it MUST ignore the route, based on the unknown route type value, as specified by Section 5.4 in [RFC7606].

The encoding of this route is defined as follows:

o RD, ESI and Ethernet Tag ID are defined as per [RFC7432] for MAC/IP routes.

o The Originator Router Length and Address encode and IPv4 or IPv6 address that belongs to the advertising PE.

o The Multicast Router Length and Address field encode the Primary IP address of the PIM neighbor added to the PE’s DB.

o The Secondary Address List Length encodes the number of Secondary IP addresses advertised by the PIM router in the PIM Hello message. If this field is zero, the NLRI will not include any Secondary Multicast Router Address. All the IP addresses will have the same Length, that is, they will all be either IPv4 or IPv6, but not a mix of both.

o DR Priority is copied from the same field in Hello packets, as per [RFC4601].

o Flags:
  - Q: Querier flag. Least significant bit. It indicates the encoded multicast router is an IGMP Querier.
  - P: PIM router flag. Second low order bit in the Flags octet. It indicates that the multicast router is a PIM router.
  - Q and P may be set simultaneously.

For BGP processing purposes, only the RD, Ethernet Tag ID, Originator Router Length and Address, and Multicast Router Length and Address are considered part of the route key. The Secondary Multicast Router Addresses and the rest of the fields are not part of the route key.

4.2 Selective Multicast Ethernet Tag Route for PIM Proxy

This document extends the SMET route defined in [EVPN-IGMP-MLD-PROXY] as shown in Figure 6.

```
             +---------------------------------------+
             |  RD (8 octets)                        |
             +---------------------------------------+
             |  Ethernet Tag ID (4 octets)           |
             +---------------------------------------+
             |  Multicast Source Length (1 octet)    |
             +---------------------------------------+
             |  Multicast Source Address (variable)  |
             +---------------------------------------+
             |  Multicast Group Length (1 octet)     |
             +---------------------------------------+
             |  Multicast Group Address (Variable)   |
             +---------------------------------------+
             |  Originator Router Length (1 octet)   |
             +---------------------------------------+
             |  Originator Router Address (variable) |
             +---------------------------------------+
             |  Flags (1 octets) (optional)          |
             +---------------------------------------+
             |  Upstream Router Length (1B)(optional)|
             +---------------------------------------+
             |  Upstream Router Addr (variable)(opt) |
             +---------------------------------------+


             Flags:

             0  1  2  3  4  5  6  7
             +--+--+--+--+--+--+--+--+
             |     |  | P|IE|v3|v2|v1|
             +--+--+--+--+--+--+--+--+
```

Figure 6 Selective Multicast Ethernet Tag Route and Flags

As in the case of the MRD route, this route type is OPTIONAL.

This route will be used as per [EVPN-IGMP-MLD-PROXY], with the following extra and optional fields:

- Upstream Router Length and Address will contain the same information as received in a PIM Join/Prune message on a local AC. There is only one Upstream Router Address per route.

- Flags: This field encodes Flags that are now relevant to IGMP and PIM. The following new Flag is defined:

  - Flag P: Indicates the SMET route is generated by a received PIM

Join on a local AC. When P=1, the Upstream Router Length and Address fields are present in the route. Otherwise the two fields will not be present.

Compared to [EVPN-IGMP-MLD-PROXY] there is no change in terms of fields considered part of the route key for BGP processing. The Upstream Router Length and Address are not considered part of the route key.

4.3 PIM RPT-Prune Route

The RPT-Prune route is analogous to the SMET route but for PIM RPT-Prune messages. The SMET routes cannot be used to convey RPT-Prune messages because they are always triggered by IGMP or PIM Join messages. A PIM RPT-Prune message is used to Prune a specific (S,G) from the RP Tree by downstream routers. An RPT-Prune message is typically seen prior to an RPT-Join message for the (S,G), hence it requires its own BGP route type (since the SMET route is always advertised based on the received Join messages).

```
+---------------------------------------+
|  RD (8 octets)                        |
+---------------------------------------+
|  Ethernet Tag ID (4 octets)           |
+---------------------------------------+
|  Multicast Source Length (1 octet)    |
+---------------------------------------+
|  Multicast Source Address (variable)  |
+---------------------------------------+
|  Multicast Group Length (1 octet)     |
+---------------------------------------+
|  Multicast Group Address (Variable)   |
+---------------------------------------+
|  Originator Router Length (1 octet)   |
+---------------------------------------+
|  Originator Router Address (variable) |
+---------------------------------------+
|  Upstream Router Length (1B)          |
+---------------------------------------+
|  Upstream Router Addr (variable)      |
+---------------------------------------+
```

Figure 7 PIM RPT-Prune Route

Fields are defined in the same way as for the SMET route.

4.4 IGMP/PIM Join Synch Route for PIM Proxy

This document renames the IGMP Join Synch Route defined in [EVPN-IGMP-MLD-PROXY] as IGMP/PIM Join Synch Route and extends it with new fields and Flags as shown in Figure 8:

```
             +-----------------------------------------------+
             |  RD (8 octets)                                |
             +-----------------------------------------------+
             | Ethernet Segment Identifier (10 octets)       |
             +-----------------------------------------------+
             |  Ethernet Tag ID  (4 octets)                  |
             +-----------------------------------------------+
             |  Multicast Source Length (1 octet)            |
             +-----------------------------------------------+
             |  Multicast Source Address (variable)          |
             +-----------------------------------------------+
             |  Multicast Group Length (1 octet)             |
             +-----------------------------------------------+
             |  Multicast Group Address (Variable)           |
             +-----------------------------------------------+
             |  Originator Router Length (1 octet)           |
             +-----------------------------------------------+
             |  Originator Router Address (variable)         |
             +-----------------------------------------------+
             |  Flags (1 octet)                              |
             +-----------------------------------------------+
             |  Upstream Router Length (1B)(optional)        |
             +-----------------------------------------------+
             |  Upstream Router Addr (variable)(opt)         |
             +-----------------------------------------------+

             Flags:

             0  1  2  3  4  5  6  7
             +--+--+--+--+--+--+--+--+
             |  |  |  | P|IE|v3|v2|v1|
             +--+--+--+--+--+--+--+--+
```

Figure 8 IGMP/PIM Join Synch Route and Flags

This route will be used as per [EVPN-IGMP-MLD-PROXY], with the following extra and optional fields:

- o Upstream Router Length and Address will contain the same information as received in a PIM Join/Prune message on a local AC. There is only one Upstream Router Address per route.

- o Flags: This field encodes Flags that are now relevant to IGMP and PIM. The following new Flag is defined:

  - Flag P: Indicates the Join Synch route is generated by a received PIM Join on a local AC. When P=1, the Upstream Router Length and Address fields are present in the route. Otherwise the two fields will not be present.

Compared to [EVPN-IGMP-MLD-PROXY] there is no change in terms of fields considered part of the route key for BGP processing. The Upstream Router Length and Address are not considered part of the route key.

4.5 IGMP/PIM RPT-Prune Synch Route for PIM Proxy

This new route is used to Synch RPT-Prune states among the PEs in the Ethernet Segment.

```
+-----------------------------------------------+
|  RD (8 octets)                                |
+-----------------------------------------------+
| Ethernet Segment Identifier (10 octets)       |
+-----------------------------------------------+
|  Ethernet Tag ID  (4 octets)                  |
+-----------------------------------------------+
|  Multicast Source Length (1 octet)            |
+-----------------------------------------------+
|  Multicast Source Address (variable)          |
+-----------------------------------------------+
|  Multicast Group Length (1 octet)             |
+-----------------------------------------------+
|  Multicast Group Address (Variable)           |
+-----------------------------------------------+
|  Originator Router Length (1 octet)           |
+-----------------------------------------------+
|  Originator Router Address (variable)         |
+-----------------------------------------------+
|  Upstream Router Length (1B)(optional)        |
+-----------------------------------------------+
|  Upstream Router Addr (variable)(opt)         |
+-----------------------------------------------+
```

Figure 9 IGMP/PIM RPT-Prune Synch Route

The RD, Ethernet Segment Identifier and other fields are defined as for the IGMP/PIM Join Synch Route. In addition, the Upstream Router Length and Address will contain the same information as received in a PIM RPT-Prune message on a local AC. The Upstream Router points at the RP for the source and group and there is only one Upstream Router Address per route.

The route key for BGP processing is defined as per the IGMP/PIM Join Synch route.

5. Conclusions

This document extends the IGMP Proxy concept of [EVPN-IGMP-MLD-PROXY] to PIM, so that EVPN can also be used to minimize the flooding of PIM control messages and optimize the delivery of IP multicast traffic in EVPN Broadcast Domains that connect PIM routers.

This specification describes procedures to Discover new PIM routers in the BD, as well as propagate PIM Join/Prune messages using EVPN SMET routes and other optimizations.

6. Conventions used in this document

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

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

7. Security Considerations

This section will be added in future versions.

8. IANA Considerations

This document requests IANA to allocate a new EVPN route type in the corresponding registry:

+ Type TBD - Multicast Router Discovery (MRD) Route

+ Type TBD - PIM RPT-Prune Route
+ Type TBD - PIM RPT-Prune Join Synch Route

In addition, the following route defined in [EVPN-IGMP-MLD-PROXY] should be renamed as follows:

+ Type 7 - IGMP/PIM Join Synch Route

9. Terminology

- o EVI: EVPN Instance.

- o EVPN Broadcast Domain: it refers to an EVI in case of VLAN-based and VLAN-bundle interfaces. It refers to a Bridge Domain identified by an Ethernet-Tag (in the control plane) in case of VLAN-Aware Bundle interfaces.

- o AC: Attachment Circuit.

- o PIM-DM: Protocol Independent Multicast - Dense Mode.

- o PIM-SM: Protocol Independent Multicast - Sparse Mode.

- o PIM-SSM: Protocol Independent Multicast - Source Specific Mode.

- o S: IP address of the multicast source.

- o G: IP address of the multicast group.

- o N: Upstream neighbor field in a Join/Prune/Graft message.

- o PIM J/P: PIM Join/Prune messages.

- o RP: PIM Rendezvous Point.

- o MRD route: Multicast Router Discovery.

- o PIM Nbr: PIM Neighbor.

10. References

10.1 Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <http://www.rfc-editor.org/info/rfc4601>.

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

[RFC8220] Dornon, O. et al, "Protocol Independent Multicast (PIM) over Virtual Private LAN Service (VPLS)", RFC 8220, DOI 10.17487/RFC8220, September 2017, <http://www.rfc-editor.org/info/rfc8220>.

[EVPN-IGMP-MLD-PROXY] Sajassi, A. et al, "IGMP and MLD Proxy for EVPN", March 2017, work-in-progress, draft-ietf-bess-evpn-igmp-mld-proxy-00.

10.2 Informative References

[EVPN-PROXY-ARP-ND] Rabadan, J. et al, "Operational Aspects of Proxy-ARP/ND in EVPN Networks", October 2017, work-in-progress, draft-ietf-bess-evpn-proxy-arp-nd-03.

11. Acknowledgments

12. Contributors

13. Authors’ Addresses

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

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA
Email: senthil.sathappan@nokia.com

Jayant Kotalwar
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jayant.kotalwar@nokia.com

Zhaohui Zhang
Juniper Networks
EMail: zzhang@juniper.net

Ali Sajassi
Cisco
Email: sajassi@cisco.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: March 25, 2018

Z. Zhang
Juniper Networks
R. Raszuk
Bloomberg LP
D. Pacella
Verizon
A. Gulko
Thomson Reuters

September 21, 2017

Controller Based BGP Multicast Signaling
draft-zzhang-bess-bgp-multicast-controller-00

Abstract

This document specifies a way that one or more centralized controllers can use BGP to set up a multicast distribution tree in a network. In the case of labeled tree, the labels are assigned by the controllers either from the controllers’ local label spaces, or from a common Segment Routing Global Block (SRGB), or from each routers Segment Routing Local Block (SRLB) that the controllers learn. In case of labeled unidirectional tree and label allocation from the common SRGB or from the controllers’ local spaces, a single common label can be used for all routers on the tree to send and receive traffic with. Since the controllers caculate the trees, they can use sophisticated algorithms and constraints to achieve traffic engineering.

Requirements Language

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

Status of This Memo

Copyright Notice

Table of Contents

1. Overview

1.1. Introduction

[I-D.zzhang-bess-bgp-multicast] describes a way to use BGP as a replacement signaling for PIM [RFC7761] or mLDP [RFC6388]. The BGP-based multicast signaling described there provides a mechanism for setting up both (s,g)/(*,g) multicast trees (as PIM does, but optionally with labels) and labeled (MPLS) multicast tunnels (as mLDP does). Each router on a tree performs essentially the same procedures as it would perform if using PIM or mLDP, but all the inter-router signaling is done using BGP.

These procedures allow the routers to set up a separate tree for each individual multicast (x,g) flow where the 'x' could be either 's' or '*', but they also allow the routers to set up trees that are used for more than one flow. In the latter case, the trees are often referred to as "multicast tunnels" or "multipoint tunnels", and specifically in this document they are mLDP tunnels (except that they are set up with BGP signaling). While it actually does not have to be restricted to mLDP tunnels, mLDP FEC is conveniently borrowed to identify the tunnel. In the rest of the document, the term tree and tunnel are used interchangeably.

The trees/tunnels are set up using the "receiver-initiated join" technique of PIM/mLDP, hop by hop from downstream routers towards the root. The BGP messages are either sent hop by hop between downstream routers and their upstream neighbors, or can be reflected by Route Reflectors (RRs).

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the controller, using the same BGP messages as defined in [I-D.zzhang-bess-bgp-multicast]. For that, some additional procedures and optimizations are specified in this document.

While it is outside the scope of this document, signaling from the controllers could be done via other means as well, like Netconf or any other SDN methods.

1.2. Resilience

Each router could establish direct BGP sessions with one or more controllers, or it could establish BGP sessions with RRs who in turn peer with controllers. For the same tree/tunnel, each controller may independentantly calculate the tree/tunnel and signal the routers on the tree/tunnel using CMCAST S-PMSI/Leaf A-D routes [I-D.zzhang-bess-bgp-multicast]. How the tree/tunnel roots/leaves are discovered and how the calculation is done are outside the scope of this document.

On each router, BGP route selection rules will lead to one controller’s route for the tree/tunnel being selected as the active route and used for setting up forwarding state. As long as all the routers on a tree/tunnel consistently pick the same controller’s routes for the tree/tunnel, the setup should be consistent. If the tree/tunnel is labeled, different labels will be used from different controllers so there is no traffic loop issue even if the routers do not consistently select the same controlle’s routes. In the unlabeled case, to ensure the consistency the selection SHOULD be solely based on the identifier of the controller, which could be carried in an Address Specific Extended Community (EC).

Another consistency issue is when a bidirectional tree/tunnel needs to be re-routed. Because this is no longer triggered hop-by-hop from downstream to upstream, it is possible that the upstream change happens before the downstream, causing traffic loop. In the unlabeled case, there is no good solution (other than that the controller issues upstream change only after it gets acknowledgement from downstream). In the labeled case, as long as a new label is used there should be no problem.

Besides the traffic loop issue, there could be transient traffic loss before both the upstream and downstream’s forwarding state are updated. This could be mitigated if the upstream keep sending traffic on the old path (in addition to the new path) and the downstream keep accepting traffic on the old path (but not on the new path) for some time. It is a local matter when for the downstream to switch to the new path - it could be data driven (e.g., after traffic arrives on the new path) or timer driven.

For each tree, multiple disjoint instances could be calculated and signaled for live-live protection. Different labels are used for different instances, so that the leaves can differentiate incoming traffic on different instances. As far as tranist routers are concerned, the insances are just independent. Note that the two instances are not expected to share common transit routers (it is otherwise outside the scope of this document/revision).

1.3. Signaling

Each router only receives S-PMSI/Leaf A-D routes from the controllers but does not originate or re-advertise those routes. The re-advertisement of a received route can be blocked based on the fact that a configured import RT matches the RT of the route, which indicates that this router is the target and consumer of the route hence it should not be re-advertised further. The routes includes the outgoing forwarding information in the form of Tunnel Encapsulation Attributes (TEA), with optional enhancements specified in this document. The router infers the incoming forwarding information from the Upstream Router’s IP Address field in the NLRI in case of an unlabeled tree.

Suppose that for a particular tree, there are two downstream routers D1 and D2 for a particular upstream router U. A controller C may send two Leaf A-D routes to U, as if the two routes were originated by D1 and D2 but reflected by the controller. As an alternative in case of a labeled tree, C could just send one route to U, with a Composite Tunnel in TEA (in this case, the Originating Router’s Address field of the Leaf A-D route is set to the controller’s address) and the Composite Tunnel specifies both downstreams. The tunnel in a TEA or Composite Tunnel is of type "MPLS Encapsulation" with a Label Stack Sub-TLV to encode label information.

For comparison, the existing TEA as specified in [I-D.ietf-idr-tunnel-encaps] can include multiple tunnels, but only one of those is used, while with a Composite Tunnel, traffic is sent out of all the enclosed tunnels to reach multiple endpoints.

Note that, in case of labeled trees, the (x,g) or mLDP FEC signaling is actually not needed to transit routers but only needed on tunnel root/leaves. However, for consistency, the same signaling is used to all routers.

1.4. Label Allocation

In the case of labeled multicast signaled hop by hop towards the root, whether it’s (x,g) multicast or "mLDP" tunnel, labels are assigned by a downstream router and advertised to its upstream router (from traffic direction point of view). In the case of controller based signaling, routers do not originate tree join (S-PMSI/Leaf A-D) routes anymore, so the controllers have to assign labels on behalf of routers, and there are three options for label assignment:

- From each router’s SRLB that the controller learns
- From the common SRGB that the controller learns
- From the controller’s local label space

Assignment from each router’s SRLB is no different from each router assigning labels from its own local label space in the hop-by-hop signaling case. The assignments for a router is independent of assignments for another router, even for the same tree.

Assignment from the controller’s local label space is upstream-assigned [RFC5331]. It is used if the controller does not learn the common SRGB or each router’s SRLB. Assignment from the SRGB [I-D.ietf-spring-segment-routing] is only meaningful if all SRGBs are the same and a single common label is used for all the routers on a tree in case of unidirectional tree/tunnel (Section 1.4.1). Otherwise, assignment from SRLB is preferred.

The choice of which of the options to use depends on many factors. An operator may want to use a single common label per tree for ease of monitoring and debugging, but that requires explicit RPF checking and either SRGB or upstream assigned labels, which may not be supported due to either the software or hardware limitations (e.g. label imposition/disposition limits). In an SR network, assignment from the common SRGB if it’s required to use a single common label per unidirectional tree, or otherwise assignment from SRLB is a good choice because it does not require support for context label spaces.

1.4.1. Using a Common per-tree Label for All Routers

MPLS labels only have local significance. For an LSP that goes through a series of routers, each router allocates a label independently and it swaps the incoming label (that it advertised to its upstream) to an outgoing label (that it received from its downstream) when it forwards a labeled packet. Even if the incoming and outgoing labels happen to be the same on a particular router, that is just incidental.

With Segment Routing, it is becoming a common practice that all routers use the same SRGB so that a SID maps to the same label on all routers. This makes it easier for operators to monitor and debug their network. The same concept applies to multicast trees as well - a common per-tree label is used for a router to receive traffic from its upstream neighbor and replicate traffic to all its downstream neighbor.

However, a common per-tree label can only be used for unidirectional trees. Additionally, it requires each router to do explicit RPF check, so that only packets from its expected upstream neighbor are accepted. Otherwise, traffic loop may form during topology changes, because the forwarding state update is no longer ordered.

Traditionally, p2mp mpls forwarding does not require explicit RPF check as a downstream router advertises a label only to its upstream router and all traffic with that incoming label is presumed to be from the upstream router and accepted. When a downtream router switches to a different upstream router a different label will be advertised, so it can determine if traffic is from its expected upstream neighbor purely based on the label. Now with a single common label used for all routers on a tree to send and receive traffic with, a router can no longer determine if the traffic is from its expected neighbor just based on that common tree label. Therefore, explicit RPF check is needed. Instead of interface based RPF checking as in PIM case, neighbor based RPF checking is used - a label identifying the upstream neighbor preceeds the tree label and the receiving router checks if that preceeding neighbor label matches its expected upstream neighbor. Notice that this is similar to what's described in Section "9.1.1 Discarding Packets from Wrong PE" of RFC 6513 (an egress PE discards traffic sent from a wrong ingress PE). The only difference is one is used for label based forwarding and the other is used for (s,g) based forwarding. [note: for bidirectional trees, we may be able to use two labels per tree - one for upstream traffic and one for downstream traffic. This needs further verification].

Both the common per-tree label and the neighbor label are allocated either from the common SRGB or from the controller's local label space. In the latter case, an additional label identifying the controller's label space is needed, as descrbibed in the following section.

1.4.2. Upstream-assignment from Controller's Local Label Space

In this case in the multicast packet's label stack the tree label and upstream neighbor label (if used in case of single common-label per tree) are preceded by a downstream-assigned "context label". The context label identifies a context-specific label space (the controller's local label space), and the upstream-assigned label that follows it is looked up in that space.

This specification requires that, in case of upstream-assignment from a controller's local label space, each router D to assign, corresponding to each controller C, a context label that identifies the upstream-assigned label space used by that controller. This label, call it Lc-D, is communicated by D to C.

Suppose a controller is setting up unidirectional tree T. It assigns that tree the label Lt, and assigns label Lu to identify router U which is the upstream of router D on tree T. C needs to tell U: "to send a packet on the given tree/tunnel, one of the things you have to do is push Lt onto the packet's label stack, then push Lu, then push Lc-D onto the packet's label stack, then unicast the packet to D. Controller C also needs to inform router D of the correspondence between <Lc-D, Lu, Lt> and tree T.

To achieve that, when C sends an S-PMSI/Leaf A-D route, for each tunnel in the TEA or in the Composite Tunnel TLV, it includes a label stack Sub-TLV [I-D.ietf-idr-tunnel-encaps], with the outer label being the context label Lc-D (received by the controller from the corresponding downstream), the next label being the upstream neighbor label Lu, and the inner label being the label Lt assigned by the controller for the tree. The router receiving the route will use the label stacks to send traffic to its downstreams.

For C to sginal the expected label stack for D to receive traffic with, we overload a tunnel TLV in either the TEA or the Composite Tunnel in the Leaf A-D route sent to D - if the remote endpoint of that tunnel TLV matches the Upstream Router field in the Leaf A-D route, then it indicates that this is actually for receiving traffic from the upstream. If a common tree label is used, then the TLV contains a variant of the Label Stack Sub-TLV because the D needs to treat the second inner most label as the upstream neighbor label and set up forwarding state accordingly for explicit RPF check. This variant is referred to as RPF Label Stack Sub-TLV (Section 2.2).

Note that the use of TEA to specify downstream and upstream forwarding information also apply to label assignment from the common SRGB or each router’s SRLB, with the differences that the context label is not needed in the SRGB/SRLB case, and that in SRLB case only a Label Stack Sub-TLV with a single SRLB label is used for upstream and downstream forwarding information (no RPF Label Stack Sub-TLV is needed) in the SRLB case.

2. Specification

2.1. Additional Tunnel Type for TEA

This document specifies a Composite Tunnel TLV and a TEA Tunnel TLV. The type codes will be assigned by IANA.

A Tunnel Encapsulation Attribute includes Tunnel TLVs and a router receiving the TEA (associated with a route) selects one of the Tunnel TLVs to set up forwarding state - a packet is sent out of only one of the tunnels. To specify that traffic needs to be sent out of multiple tunnels, a Composite Tunnel TLV is used. The value part of the TLV includes a list of sub-TLVs, each being a Tunnel TLV. Obviously, a Composite Tunnel TLV MUST not be a sub-TLV of a Composite Tunnel TLV.

Consider that a Composite Tunnel TLV that includes a bunch of sub-TLVs specifying a bunch of tunnels used to send traffic to a bunch of endpoints. For a particular endpoint, there are multiple ways to reach it - any one but only one should be used. For that purpose, a

TEA Tunnel TLV (for lack of a better name) is usded for that endpoint. The TEA Tunnel TLV includes a bunch of sub-TLVs, each being a Tunnel TLV that specifies one way to reach the same endpoint. This is similar to a Tunnel Encapsulation Attribute, hence the name TEA Tunnel TLV.

2.2. RPF Label Stack Sub-TLV

This is almost identifcal to Label Stack Sub-TLV. The only difference is that the second inner most label in the stack identifies the expected upstream neighbor and explicit RPF checking needs to be set up for the tree label accordingly.

2.3. Context Label Wide Community

For a router to signal the context label that it assigns for a controller (or any label allocator that assigns labels that will be seen by this router), it attaches a Context Label Wide Community [I-D.ietf-idr-wide-bgp-communities] to the host route for its own address used in its BGP session towards the controllers (directly or via RRs). This is a new wide community that specifies the (Label Allocator, Context Label) tuple, and the exactly format will be specified in a future revision.

2.4. Procedures

Details to be added. The general idea is described in the introduction section.

3. Security Considerations

This document does not introduce new security risks?

4. IANA Considerations

To be added.

5. Acknowledgements

The authors Eric Rosen for his questions, suggestions, and help finding solutions to some issues like the neighbor based explicit RPF checcking. The authors also thank Lenny Giuliano and IJsbrand Wijnands for their review and comments.

6. References

6.1. Normative References

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-07 (work in progress), July 2017.

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", draft-ietf-idr-wide-bgp-communities-04 (work in progress), March 2017.

[I-D.zzhang-bess-bgp-multicast]
Zhang, Z., Patel, K., Wijnands, I., and a. arkadiy.gulko@thomsonreuters.com, "BGP Based Multicast", draft-zzhang-bess-bgp-multicast-01 (work in progress), 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>.

6.2. Informative References

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-12 (work in progress), June 2017.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[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>.

[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>.

Authors’ Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Robert Raszuk
Bloomberg LP

EMail: robert@raszuk.net

Dante Pacella
Verizon

EMail: dante.j.pacella@verizon.com

Arkadiy Gulko
Thomson Reuters

EMail: arkadiy.gulko@thomsonreuters.com

BESS   Z. Zhang
Internet-Draft   Juniper Networks
Intended status: Standards Track   R. Raszuk
Expires: August 10, 2019   Bloomberg LP
                                                              D. Pacella
                                                                 Verizon
                                                                A. Gulko
                                                         Thomson Reuters
                                                        February 6, 2019

Controller Based BGP Multicast Signaling
draft-zzhang-bess-bgp-multicast-controller-01

Abstract

This document specifies a way that one or more centralized controllers can use BGP to set up a multicast distribution tree in a network. In the case of labeled tree, the labels are assigned by the controllers either from the controllers' local label spaces, or from a common Segment Routing Global Block (SRGB), or from each routers Segment Routing Local Block (SRLB) that the controllers learn. In case of labeled unidirectional tree and label allocation from the common SRGB or from the controllers' local spaces, a single common label can be used for all routers on the tree to send and receive traffic with. Since the controllers calculate the trees, they can use sophisticated algorithms and constraints to achieve traffic engineering.

Requirements Language

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

Status of This Memo

Copyright Notice

Table of Contents

1. Overview

1.1. Introduction

[I-D.zzhang-bess-bgp-multicast] describes a way to use BGP as a replacement signaling for PIM [RFC7761] or mLDP [RFC6388]. The BGP-based multicast signaling described there provides a mechanism for setting up both (s,g)/(*,g) multicast trees (as PIM does, but optionally with labels) and labeled (MPLS) multicast tunnels (as mLDP does). Each router on a tree performs essentially the same procedures as it would perform if using PIM or mLDP, but all the inter-router signaling is done using BGP.

These procedures allow the routers to set up a separate tree for each individual multicast (x,g) flow where the 'x' could be either 's' or '*', but they also allow the routers to set up trees that are used for more than one flow. In the latter case, the trees are often referred to as "multicast tunnels" or "multipoint tunnels", and specifically in this document they are mLDP tunnels (except that they are set up with BGP signaling). While it actually does not have to be restricted to mLDP tunnels, mLDP FEC is conveniently borrowed to identify the tunnel. In the rest of the document, the term tree and tunnel are used interchangeably.

The trees/tunnels are set up using the "receiver-initiated join" technique of PIM/mLDP, hop by hop from downstream routers towards the root. The BGP messages are either sent hop by hop between downstream routers and their upstream neighbors, or can be reflected by Route Reflectors (RRs).

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the controller, using the same BGP messages as defined in [I-D.zzhang-bess-bgp-multicast]. For that, some additional procedures and optimizations are specified in this document.

While it is outside the scope of this document, signaling from the controllers could be done via other means as well, like Netconf or any other SDN methods.

1.2. Resilience

Each router could establish direct BGP sessions with one or more controllers, or it could establish BGP sessions with RRs who in turn peer with controllers. For the same tree/tunnel, each controller may independently calculate the tree/tunnel and signal the routers on the tree/tunnel using MCAST-TREE S-PMSI/Leaf A-D routes [I-D.zzhang-bess-bgp-multicast]. How the tree/tunnel roots/leaves are discovered and how the calculation is done are outside the scope of this document.

On each router, BGP route selection rules will lead to one controller's route for the tree/tunnel being selected as the active route and used for setting up forwarding state. As long as all the routers on a tree/tunnel consistently pick the same controller's routes for the tree/tunnel, the setup should be consistent. If the tree/tunnel is labeled, different labels will be used from different controllers so there is no traffic loop issue even if the routers do not consistently select the same controlle's routes. In the unlabeled case, to ensure the consistency the selection SHOULD be solely based on the identifier of the controller, which could be carried in an Address Specific Extended Community (EC).

Another consistency issue is when a bidirectional tree/tunnel needs to be re-routed. Because this is no longer triggered hop-by-hop from downstream to upstream, it is possible that the upstream change happens before the downstream, causing traffic loop. In the unlabeled case, there is no good solution (other than that the controller issues upstream change only after it gets acknowledgement from downstream). In the labeled case, as long as a new label is used there should be no problem.

Besides the traffic loop issue, there could be transient traffic loss before both the upstream and downstream's forwarding state are updated. This could be mitigated if the upstream keep sending traffic on the old path (in addition to the new path) and the downstream keep accepting traffic on the old path (but not on the new path) for some time. It is a local matter when for the downstream to switch to the new path - it could be data driven (e.g., after traffic arrives on the new path) or timer driven.

For each tree, multiple disjoint instances could be calculated and signaled for live-live protection. Different labels are used for different instances, so that the leaves can differentiate incoming traffic on different instances. As far as transit routers are concerned, the instances are just independent. Note that the two instances are not expected to share common transit routers (it is otherwise outside the scope of this document/revision).

1.3. Signaling

Each router only receives S-PMSI/Leaf A-D routes from the controllers but does not originate or re-advertise those routes. The re-advertisement of a received route can be blocked based on the fact that a configured import RT matches the RT of the route, which indicates that this router is the target and consumer of the route hence it should not be re-advertised further. The routes includes the outgoing forwarding information in the form of Tunnel Encapsulation Attributes (TEA), with optional enhancements specified in this document. The router infers the incoming forwarding information from the Upstream Router's IP Address field in the NLRI in case of an unlabeled tree.

Suppose that for a particular tree, there are two downstream routers D1 and D2 for a particular upstream router U. A controller C may send two Leaf A-D routes to U, as if the two routes were originated by D1 and D2 but reflected by the controller. As an alternative in case of a labeled tree, C could just send one route to U, with a Composite Tunnel in TEA (in this case, the Originating Router's Address field of the Leaf A-D route is set to the controller's address) and the Composite Tunnel specifies both downstreams. The tunnel in a TEA or Composite Tunnel is of type "MPLS Encapsulation" with a Label Stack Sub-TLV to encode label information.

For comparison, the existing TEA as specified in [I-D.ietf-idr-tunnel-encaps] can include multiple tunnels, but only one of those is used, while with a Composite Tunnel, traffic is sent out of all the enclosed tunnels to reach multiple endpoints.

Note that, in case of labeled trees, the (x,g) or mLDP FEC signaling is actually not needed to transit routers but only needed on tunnel root/leaves. However, for consistency, the same signaling is used to all routers.

1.4. Label Allocation

In the case of labeled multicast signaled hop by hop towards the root, whether it's (x,g) multicast or "mLDP" tunnel, labels are assigned by a downstream router and advertised to its upstream router (from traffic direction point of view). In the case of controller based signaling, routers do not originate tree join (S-PMSI/Leaf A-D) routes anymore, so the controllers have to assign labels on behalf of routers, and there are three options for label assignment:

- From each router's SRLB that the controller learns

- From the common SRGB that the controller learns

- From the controller's local label space

Assignment from each router's SRLB is no different from each router assigning labels from its own local label space in the hop-by-hop signaling case. The assignments for a router is independent of assignments for another router, even for the same tree.

Assignment from the controller's local label space is upstream-assigned [RFC5331]. It is used if the controller does not learn the common SRGB or each router's SRLB. Assignment from the SRGB [RFC8402] is only meaningful if all SRGBs are the same and a single common label is used for all the routers on a tree in case of unidirectional tree/tunnel (Section 1.4.1). Otherwise, assignment from SRLB is preferred.

The choice of which of the options to use depends on many factors. An operator may want to use a single common label per tree for ease of monitoring and debugging, but that requires explicit RPF checking and either SRGB or upstream assigned labels, which may not be supported due to either the software or hardware limitations (e.g. label imposition/disposition limits). In an SR network, assignment from the common SRGB if it's required to use a single common label per unidirectional tree, or otherwise assignment from SRLB is a good choice because it does not require support for context label spaces.

1.4.1. Using a Common per-tree Label for All Routers

MPLS labels only have local significance. For an LSP that goes through a series of routers, each router allocates a label independently and it swaps the incoming label (that it advertised to its upstream) to an outgoing label (that it received from its downstream) when it forwards a labeled packet. Even if the incoming and outgoing labels happen to be the same on a particular router, that is just incidental.

With Segment Routing, it is becoming a common practice that all routers use the same SRGB so that a SID maps to the same label on all routers. This makes it easier for operators to monitor and debug their network. The same concept applies to multicast trees as well - a common per-tree label is used for a router to receive traffic from its upstream neighbor and replicate traffic to all its downstream neighbor.

However, a common per-tree label can only be used for unidirectional trees. Additionally, it requires each router to do explicit RPF check, so that only packets from its expected upstream neighbor are accepted. Otherwise, traffic loop may form during topology changes, because the forwarding state update is no longer ordered.

Traditionally, p2mp mpls forwarding does not require explicit RPF check as a downstream router advertises a label only to its upstream router and all traffic with that incoming label is presumed to be from the upstream router and accepted. When a downstream router switches to a different upstream router a different label will be advertised, so it can determine if traffic is from its expected upstream neighbor purely based on the label. Now with a single common label used for all routers on a tree to send and receive traffic with, a router can no longer determine if the traffic is from its expected neighbor just based on that common tree label. Therefore, explicit RPF check is needed. Instead of interface based RPF checking as in PIM case, neighbor based RPF checking is used - a label identifying the upstream neighbor precedes the tree label and the receiving router checks if that preceding neighbor label matches its expected upstream neighbor. Notice that this is similar to what's described in Section "9.1.1 Discarding Packets from Wrong PE" of RFC 6513 (an egress PE discards traffic sent from a wrong ingress PE). The only difference is one is used for label based forwarding and the other is used for (s,g) based forwarding. [note: for bidirectional trees, we may be able to use two labels per tree - one for upstream traffic and one for downstream traffic. This needs further verification].

Both the common per-tree label and the neighbor label are allocated either from the common SRGB or from the controller's local label space. In the latter case, an additional label identifying the controller's label space is needed, as described in the following section.

1.4.2. Upstream-assignment from Controller's Local Label Space

In this case in the multicast packet's label stack the tree label and upstream neighbor label (if used in case of single common-label per tree) are preceded by a downstream-assigned "context label". The context label identifies a context-specific label space (the controller's local label space), and the upstream-assigned label that follows it is looked up in that space.

This specification requires that, in case of upstream-assignment from a controller's local label space, each router D to assign, corresponding to each controller C, a context label that identifies the upstream-assigned label space used by that controller. This label, call it Lc-D, is communicated by D to C.

Suppose a controller is setting up unidirectional tree T. It assigns that tree the label Lt, and assigns label Lu to identify router U which is the upstream of router D on tree T. C needs to tell U: "to send a packet on the given tree/tunnel, one of the things you have to do is push Lt onto the packet's label stack, then push Lu, then push Lc-D onto the packet's label stack, then unicast the packet to D".

Controller C also needs to inform router D of the correspondence between <Lc-D, Lu, Lt> and tree T.

To achieve that, when C sends an S-PMSI/Leaf A-D route, for each tunnel in the TEA or in the Composite Tunnel TLV, it includes a label stack Sub-TLV [I-D.ietf-idr-tunnel-encaps], with the outer label being the context label Lc-D (received by the controller from the corresponding downstream), the next label being the upstream neighbor label Lu, and the inner label being the label Lt assigned by the controller for the tree. The router receiving the route will use the label stacks to send traffic to its downstreams.

For C to signal the expected label stack for D to receive traffic with, we overload a tunnel TLV in either the TEA or the Composite Tunnel in the Leaf A-D route sent to D - if the remote endpoint of that tunnel TLV matches the Upstream Router field in the Leaf A-D route, then it indicates that this is actually for receiving traffic from the upstream. If a common tree label is used, then the TLV contains a variant of the Label Stack Sub-TLV because the D needs to treat the second inner most label as the upstream neighbor label and set up forwarding state accordingly for explicit RPF check. This variant is referred to as RPF Label Stack Sub-TLV (Section 2.2).

Note that the use of TEA to specify downstream and upstream forwarding information also apply to label assignment from the common SRGB or each router's SRLB, with the differences that the context label is not needed in the SRGB/SRLB case, and that in SRLB case only a Label Stack Sub-TLV with a single SRLB label is used for upstream and downstream forwarding information (no RPF Label Stack Sub-TLV is needed) in the SRLB case.

2. Specification

2.1. Additional Tunnel Type for TEA

This document specifies a Composite Tunnel TLV and a TEA Tunnel TLV. The type codes will be assigned by IANA.

A Tunnel Encapsulation Attribute includes Tunnel TLVs and a router receiving the TEA (associated with a route) selects one of the Tunnel TLVs to set up forwarding state - a packet is sent out of only one of the tunnels. To specify that traffic needs to be sent out of multiple tunnels, a Composite Tunnel TLV is used. The value part of the TLV includes a list of sub-TLVs, each being a Tunnel TLV. Obviously, a Composite Tunnel TLV MUST not be a sub-TLV of a Composite Tunnel TLV.

Consider that a Composite Tunnel TLV that includes a bunch of sub-TLVs specifying a bunch of tunnels used to send traffic to a bunch of endpoints. For a particular endpoint, there are multiple ways to reach it - any one but only one should be used. For that purpose, a TEA Tunnel TLV (for lack of a better name) is used for that endpoint. The TEA Tunnel TLV includes a bunch of sub-TLVs, each being a Tunnel TLV that specifies one way to reach the same endpoint. This is similar to a Tunnel Encapsulation Attribute, hence the name TEA Tunnel TLV.

2.2. RPF Label Stack Sub-TLV

This is almost identical to Label Stack Sub-TLV. The only difference is that the second inner most label in the stack identifies the expected upstream neighbor and explicit RPF checking needs to be set up for the tree label accordingly.

2.3. Context Label Wide Community

For a router to signal the context label that it assigns for a controller (or any label allocator that assigns labels that will be seen by this router), it attaches a Context Label Wide Community [I-D.ietf-idr-wide-bgp-communities] to the host route for its own address used in its BGP session towards the controllers (directly or via RRs). This is a new wide community that specifies the (Label Allocator, Context Label) tuple, and the exactly format will be specified in a future revision.

2.4. Procedures

Details to be added. The general idea is described in the introduction section.

3. Security Considerations

This document does not introduce new security risks?

4. IANA Considerations

To be added.

5. Acknowledgements

The authors Eric Rosen for his questions, suggestions, and help finding solutions to some issues like the neighbor based explicit RPF checking. The authors also thank Lenny Giuliano and IJsbrand Wijnands for their review and comments.

6. References

6.1. Normative References

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10 (work in progress), August 2018.

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", draft-ietf-idr-wide-bgp-communities-05 (work in progress), July 2018.

[I-D.zzhang-bess-bgp-multicast]
Zhang, Z., Giuliano, L., Patel, K., Wijnands, I., mishra, m., and A. Gulko, "BGP Based Multicast", draft-zzhang-bess-bgp-multicast-02 (work in progress), December 2018.

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

[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>.

6.2. Informative References

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[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>.

[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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

Authors’ Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Robert Raszuk
Bloomberg LP

EMail: robert@raszuk.net

Dante Pacella
Verizon

EMail: dante.j.pacella@verizon.com

Arkadiy Gulko
Thomson Reuters

EMail: arkadiy.gulko@thomsonreuters.com

BESS   Z. Zhang
Internet-Draft   L. Giuliano
Intended status: Standards Track   Juniper Networks
Expires: March 31, 2018   September 27, 2017

MVPN and MSDP SA Interoperation
draft-zzhang-bess-mvpn-msdp-sa-interoperation-00

Abstract

This document specifies the procedures for interoperation between MVPN Source Active routes and customer MSDP Source Active routes, which is useful for MVPN provider networks offering services to customers with an existing MSDP infrastructure. Without the procedures described in this document, VPN-specific MSDP sessions are required among the PEs that are customer MSDP peers.

Requirements Language

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

Status of This Memo

This Internet-Draft will expire on March 31, 2018.

Copyright Notice

Table of Contents

1. Terminologies

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

- ASM: Any source multicast.

- SPT: Source-specific Shortest-path Tree.

- C-S: A multicast source address, identifying a multicast source located at a VPN customer site.

- C-G: A multicast group address used by a VPN customer.

- C-RP: A multicast Rendezvous Point for a VPN customer.

- EC: Extended Community.

2. Introduction

Section "14. Supporting PIM-SM without Inter-Site Shared C-Trees" of [RFC6514] specifies the procedures for MVPN PEs to discover (C-S,C-G) via MVPN Source Active A-D routes and then send (C-S,C-G) C-multicast routes towards the ingress PEs, to establish SPTs for customer ASM flows for which they have downstream receivers. (C-*,C-G) C-multicast routes are not sent among the PEs so inter-site shared C-Trees are not used and the method is generally referred to as "spt-only" mode.

With this mode, the MVPN Source Active routes are functionally similar MSDP Source-Active messages [RFC3618]. One or more of the PEs, say PE1, either act as a C-RP and learn of (C-S,C-G) via PIM Register messages, or have MSDP sessions with some MSDP peers and learn (C-S,C-G) via MSDP SA messages. In either case, PE1 will then originate MVPN SA routes for other PEs to learn the (C-S,C-G).

[RFC6514] only specifies that a PE receiving the MVPN SA routes, say PE2, will advertise (C-S,C-G) C-multicast routes if it has corresponding (C-*,C-G) state learnt from its CE. PE2 may also have MSDP sessions with other C-RPs at its site, but [RFC6514] does not specify that it advertise MSDP SA messages to those MSDP peers for the (C-S,C-G) that it learns via MVPN SA routes. PE2 would need to have an MSDP session with PE1 (that advertised the MVPN SA messages) to learn the sources via MSDP SA messages, for it to advertise the MSDP SA to its local peers. To make things worse, unless blocked by policy control, PE2 would in turn advertise MVPN SA routes because of those MSDP SA messages that it receives from PE1, which are redundant and unnecessary. Also notice that the PE1-PE2 MSDP session is VPN-specific, while the BGP sessions over which the MVPN routes are advertised are not.

If a PE does advertise MSDP SA messages based on received MVPN SA routes, the VPN-specific MSDP sessions are no longer needed. Additionally, this MVPN/MSDP SA interoperation has the following inherent benefits for a BGP based solution.

- MSDP SA refreshes are replaced with BGP hard state.

- Route Reflectors can be used instead of having peer-to-peer sessions.

- BGP route propagation/selection rules remove the need for RPF checking required by MSDP.

- VPN extranet mechanisms can be used to propagate (C-S,C-G) information across VPNs with flexible policy control.

While MSDP Source Active routes contain the source, group and RP address of a given multicast flow, MVPN Source Active routes only contain the source and group. MSDP requires the RP address information in order to perform peer-RPF. Therefore, this document describes how to convey the RP address information into the MVPN Source Active route using an Extended Community so this information can be shared with an existing MSDP infrastructure.

2.1. MVPN RPT-SPT Mode

For comparison, another method of supporting customer ASM is generally referred to "rpt-spt" mode. Section "13. Switching from a Shared C-Tree to a Source C-Tree" of [RFC6514] specifies the MVPN SA procedures for that mode, but those SA routes are replacement for PIM-ASM assert and (s,g,rpt) prune mechanisms, not for source discovery purpose. MVPN/MSDP SA interoperation for the "rpt-spt" mode is outside of the scope of this document. In the rest of the document, the "spt-only" mode is assumed.

3. Specification

When an MVPN PE advertises an MVPN SA route, it SHOULD attach an "MVPN SA RP-address Extended Community". This is a Transitive IPv4-Address-Specific Extended Community. The Local Administrative field is set to zero and the Global Administrative field is set to an RP address determined as the following:

- If the (C-S,C-G) is learnt as result of PIM Register mechanism, the local RP address in the VRF is used.

- If the (C-S,C-G) is learnt as result of incoming MSDP SA messages, the RP address in the selected MSDP SA message is used.

If an MVPN PE has one or more MSDP sessions and receives an MVPN SA route that is selected as the best MVPN SA route for a given (C-S,C-G), the PE generates an MSDP SA and transmits it to those MSDP peers. The Global Administrative field in the MVPN SA RP-address EC of the MVPN SA route is used to populate the RP address of the MSDP SA. If the MVPN SA route does not have the EC, the local RP address of the VRF is be used to populate the RP address field of the MSDP SA.

If an MVPN PE receives the withdraw of an MVPN SA route, a new best MVPN SA route for the (C-S,C-G) may be selected. A new MSDP SA message is advertised if the RP address determined according to the newly selected best MVPN SA route is different from before. If there is no MVPN SA route left for the (C-S,C-G), the previously advertised MSDP SA message will not be refreshed and will eventually time out.

4. IANA Considerations

This document introduces a new Transitive IPv4 Address Specific Extended Community "MVPN SA RP-address Extended Community". An IANA request is submitted for a subcode of 0x20 (pending approval and subject to change) in the Transitive IPv4-Address-Specific Extended Community Sub-Types registry.

5. Acknowledgements

The authors Eric Rosen for his review, comments, questions and suggestions for this document. The authors also thank Yajun Liu for her review and comments.

6. Normative References

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

[RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, DOI 10.17487/RFC3618, October 2003, <https://www.rfc-editor.org/info/rfc3618>.

[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>.

Authors’ Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Lenny Giuliano
Juniper Networks

EMail: lenny@juniper.net

BESS
Internet-Draft
Updates: 6514 (if approved)
Intended status: Standards Track
Expires: July 22, 2018

Z. Zhang
L. Giuliano
Juniper Networks

January 18, 2018

MVPN and MSDP SA Interoperation
draft-zzhang-bess-mvpn-msdp-sa-interoperation-01

Abstract

This document specifies the procedures for interoperation between MVPN Source Active routes and customer MSDP Source Active routes, which is useful for MVPN provider networks offering services to customers with an existing MSDP infrastructure. Without the procedures described in this document, VPN-specific MSDP sessions are required among the PEs that are customer MSDP peers.

Requirements Language

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

Status of This Memo

This Internet-Draft will expire on July 22, 2018.

Copyright Notice

Table of Contents

1. Terminologies

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

- ASM: Any source multicast.
- SPT: Source-specific Shortest-path Tree.
- C-S: A multicast source address, identifying a multicast source located at a VPN customer site.
- C-G: A multicast group address used by a VPN customer.
- C-RP: A multicast Rendezvous Point for a VPN customer.
- EC: Extended Community.

2. Introduction

Section "14. Supporting PIM-SM without Inter-Site Shared C-Trees" of [RFC6514] specifies the procedures for MVPN PEs to discover (C-S,C-G) via MVPN Source Active A-D routes and then send (C-S,C-G) C-multicast routes towards the ingress PEs, to establish SPTs for customer ASM flows for which they have downstream receivers. (C-*,C-G)

C-multicast routes are not sent among the PEs so inter-site shared C-Trees are not used and the method is generally referred to as "spt-only" mode.

With this mode, the MVPN Source Active routes are functionally similar to MSDP Source-Active messages [RFC3618]. One or more of the PEs, say PE1, either act as a C-RP and learn of (C-S,C-G) via PIM Register messages, or have MSDP sessions with some MSDP peers and learn (C-S,C-G) via MSDP SA messages. In either case, PE1 will then originate MVPN SA routes for other PEs to learn the (C-S,C-G).

[RFC6514] only specifies that a PE receiving the MVPN SA routes, say PE2, will advertise (C-S,C-G) C-multicast routes if it has corresponding (C-*,C-G) state learnt from its CE. PE2 may also have MSDP sessions with other C-RPs at its site, but [RFC6514] does not specify that it advertise MSDP SA messages to those MSDP peers for the (C-S,C-G) that it learns via MVPN SA routes. PE2 would need to have an MSDP session with PE1 (that advertised the MVPN SA messages) to learn the sources via MSDP SA messages, for it to advertise the MSDP SA to its local peers. To make things worse, unless blocked by policy control, PE2 would in turn advertise MVPN SA routes because of those MSDP SA messages that it receives from PE1, which are redundant and unnecessary. Also notice that the PE1-PE2 MSDP session is VPN-specific, while the BGP sessions over which the MVPN routes are advertised are not.

If a PE does advertise MSDP SA messages based on received MVPN SA routes, the VPN-specific MSDP sessions are no longer needed. Additionally, this MVPN/MSDP SA interoperation has the following inherent benefits for a BGP based solution.

o MSDP SA refreshes are replaced with BGP hard state.

o Route Reflectors can be used instead of having peer-to-peer sessions.

o VPN extranet mechanisms can be used to propagate (C-S,C-G) information across VPNs with flexible policy control.

While MSDP Source Active routes contain the source, group and RP address of a given multicast flow, MVPN Source Active routes only contain the source and group. MSDP requires the RP address information in order to perform peer-RPF. Therefore, this document describes how to convey the RP address information into the MVPN Source Active route using an Extended Community so this information can be shared with an existing MSDP infrastructure.

The procedures apply to Global Table Multicast (GTM) [RFC7716] as well.

2.1. MVPN RPT-SPT Mode

For comparison, another method of supporting customer ASM is generally referred to "rpt-spt" mode. Section "13. Switching from a Shared C-Tree to a Source C-Tree" of [RFC6514] specifies the MVPN SA procedures for that mode, but those SA routes are replacement for PIM-ASM assert and (s,g,rpt) prune mechanisms, not for source discovery purpose. MVPN/MSDP SA interoperation for the "rpt-spt" mode is outside of the scope of this document. In the rest of the document, the "spt-only" mode is assumed.

3. Specification

The MVPN PEs that act as customer RPs or have one or more MSDP sessions in a VPN (or the global table in case of GTM) are treated as an MSDP mesh group for that VPN (or the global table). In the rest of the document, it is referred to as the PE mesh group. It MUST not include other MSDP speakers, and is integrated into the rest of MSDP infrastructure for the VPN (or the global table) following normal MSDP rules and practices.

When an MVPN PE advertises an MVPN SA route following procedures in [RFC6514] for the "spt-only" mode, it SHOULD attach an "MVPN SA RP-address Extended Community". This is a Transitive IPv4-Address-Specific Extended Community. The Local Administrative field is set to zero and the Global Administrative field is set to an RP address determined as the following:

- If the (C-S,C-G) is learnt as result of PIM Register mechanism, the local RP address for the C-G is used.

- If the (C-S,C-G) is learnt as result of incoming MSDP SA messages, the RP address in the selected MSDP SA message is used.

In addition to procedures in [RFC6514], an MVPN PE may be provisioned to generate MSDP SA messages from received MVPN SA routes, with or without fine policy control. If a received MVPN SA route is to trigger MSDP SA message, it is treated as if a corresponding MSDP SA message was received from within the PE mesh group and normal MSDP procedure is followed (e.g. an MSDP SA message is advertised to other MSDP peers outside the PE mesh group). The (S,G) information comes from the (C-S,C-G) encoding in the MVPN SA NLRI and the RP address comes from the "MVPN SA RP-address EC" mentioned above. If the received MVPN SA route does not have the EC (this could be from a legacy PE that does not have the capability to attach the EC), the local RP address for the C-G is used. In that case, it is possible that receiving PE’s RP for the C-G is actually the MSDP peer to which the generated MSDP message is advertised, causing the peer to discard it due to RPF failure. To get around that problem the peer SHOULD use local policy to accept the MSDP SA message.

An MVPN PE MAY treat only the best MVPN SA route selected by BGP route selection process (instead of all MVPN SA routes) for a given (C-S,C-G) as a received MSDP SA message (and advertise corresponding MSDP message). In that case, if the selected best MVPN SA route does not have the "MVPN SA RP-address EC" but another route for the same (C-S, C-G) does, then the best route with the EC SHOULD be chosen. As a result, when/if the best MVPN SA route with the EC changes, a new MSDP SA message is advertised if the RP address determined according to the newly selected MVPN SA route is different from before. The previously advertised MSDP SA message with the older RP address will be timed out.

4. IANA Considerations

This document introduces a new Transitive IPv4 Address Specific Extended Community "MVPN SA RP-address Extended Community". An IANA request will be submitted for a subcode of 0x20 (pending approval and subject to change) in the Transitive IPv4-Address-Specific Extended Community Sub-Types registry.

5. Acknowledgements

The authors thank Eric Rosen and Vinod Kumar for their review, comments, questions and suggestions for this document. The authors also thank Yajun Liu for her review and comments.

6. References

6.1. Normative References

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

[RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, DOI 10.17487/RFC3618, October 2003, <https://www.rfc-editor.org/info/rfc3618>.

[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>.

6.2. Informative References

[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>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Lenny Giuliano
Juniper Networks

EMail: lenny@juniper.net