L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

A. Sajassi, Ed.
S. Salam
S. Thoria
Cisco
J. Drake
Juniper
J. Rabadan
Nokia
L. Yong
Huawei

Expires: August 8, 2017

February 8, 2017

Integrated Routing and Bridging in EVPN
draft-ietf-bess-evpn-inter-subnet-forwarding-03

Abstract

EVPN provides an extensible and flexible multi-homing VPN solution for intra-subnet connectivity among hosts/VMs over an MPLS/IP network. However, there are scenarios in which inter-subnet forwarding among hosts/VMs across different IP subnets is required, while maintaining the multi-homing capabilities of EVPN. This document describes an Integrated Routing and Bridging (IRB) solution based on EVPN to address such requirements.

http://www.ietf.org/shadow.html

Table of Contents

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].

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

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

IRB: Integrated Routing and Bridging

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

Bridge Table: An instantiation of a broadcast domain on a MAC-VRF

IP-VRF: A Virtual Routing and Forwarding table for IP addresses on a PE that is associated with one or more EVIs

IRB Interface: A virtual interface that connects a bridge table in a MAC-VRF to an IP-VRF in an NVE.

NVE: Network Virtualization Endpoint

TS: Tenant System

Ethernet NVO tunnel: It refers to Network Virtualization Overlay tunnels with Ethernet payload. Example of this type of tunnels are VxLAN and NvGRE.

IP NVO tunnel: It refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload). Examples of IP NVO tunnels are VxLAN GPE or MPLSoGRE (both with IP payload).

1 Introduction

EVPN provides an extensible and flexible multi-homing VPN solution for intra-subnet connectivity among Tenant Systems (TS's) over an MPLS/IP network; where, an IP subnet is represented by an EVI for a VLAN-based service or by an <EVI, VLAN> for a VLAN-aware bundle service. However, there are scenarios where, in addition to intra-subnet forwarding, inter-subnet forwarding is required among TS's across different IP subnets at EVPN PE nodes, also known as EVPN NVE nodes throughout this document, while maintaining the multi-homing capabilities of EVPN. This document describes an Integrated Routing and Bridging (IRB) solution based on EVPN to address such requirements.

The inter-subnet communication is traditionally achieved at centralized L3 Gateway (L3GW) nodes where all the inter-subnet communication policies are enforced. When two Tenant Systems (TS's) belonging to two different subnets connected to the same PE node, wanted to talk to each other, their traffic needed to be back hauled from the PE node all the way to the centralized gateway nodes where inter-subnet switching is performed and then back to the PE node. For today's large multi-tenant data center, this scheme is very inefficient and sometimes impractical.

In order to overcome the drawback of centralized approach, IRB functionality is needed on the PE nodes (i.e., NVE devices) as close to TS as possible to avoid hair pinning of user traffic unnecessarily. Under this design, all traffic between hosts attached to one NVE can be routed and bridged locally, thus avoiding traffic hair-pinning issue of the centralized L3GW.

There can be scenarios where both centralized and distributed approaches may be preferred simultaneously. For example, to allow NVEs to switch inter-subnet traffic belonging to one tenant or one security zone locally; whereas, to back haul inter-subnet traffic belonging to two different tenants or security zones to the centralized gateway nodes and perform switching there after the traffic is subjected to Firewall (FW) or Deep Packet Inspection (DPI).

Some TS's run non-IP protocols in conjunction with their IP traffic. Therefore, it is important to handle both kinds of traffic optimally - e.g., to bridge non-IP traffic and to route IP traffic.

Therefore, the solution needs to meet the following requirements:

R1: The solution MUST allow for inter-subnet traffic to be locally switched at NVEs.

R2: The solution MUST allow for both inter-subnet and intra-subnet traffic belonging to the same tenant to be locally routed and bridged respectively. The solution MUST provide IP routing for inter-subnet traffic and Ethernet Bridging for intra-subnet traffic.

R3: The solution MUST support bridging of non-IP traffic.

R4: The solution MUST allow inter-subnet switching to be disabled on a per VLAN basis on NVEs where the traffic needs to be back hauled to another node (i.e., for performing FW or DPI functionality).

2 Inter-Subnet Forwarding Scenarios

The inter-subnet forwarding scenarios performed by an EVPN NVE can be divided into the following five categories. The last scenario, along with its corresponding solution, are described in [EVPN-IPVPN-INTEROP]. The first four scenarios are covered in this document.

1. Switching among IP subnets within a DC using EVPN

2. Switching among IP subnets in different DCs using EVPN without GW

3. Switching among IP subnets in different DCs using EVPN with GW

4. Switching among IP subnets spread across IP-VPN and EVPN networks with GW

5. Switching among IP subnets spread across IP-VPN and EVPN networks without GW

In the above scenario, the term "GW" refers to the case where a node situated at the WAN edge of the data center network behaves as a default gateway (GW) for all the destinations that are outside the data center. The absence of GW refers to the scenario where NVEs within a data center maintain individual (host) routes that are outside of the data center.

In the case (4), the WAN edge node also performs route aggregation for all the destinations within its own data center, and acts as an interworking unit between EVPN and IP VPN (it implements both EVPN and IP-VPN functionality).

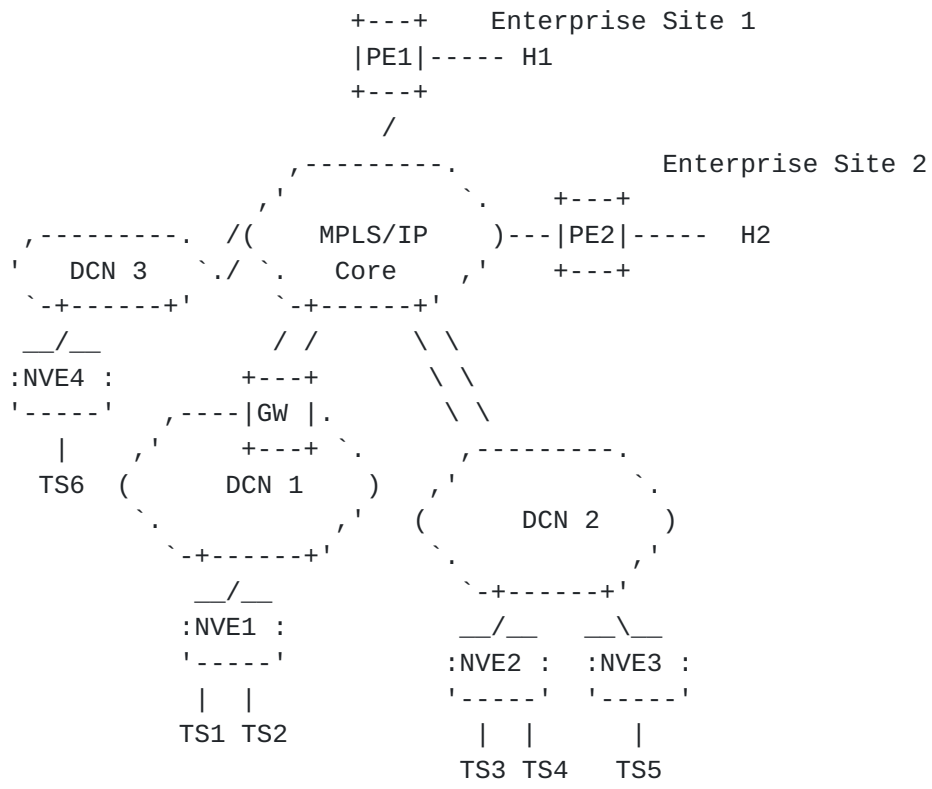


Figure 2: Interoperability Use-Cases

In what follows, we will describe scenarios 1 through 4 in more detail.

2.1 Switching among IP subnets within a DC

In this scenario, connectivity is required between TS's in the same data center, where those hosts belong to different IP subnets. All these subnets belong to the same tenant or are part of the same IP VPN. Each subnet is associated with a single EVI (or <EVI,VLAN>) realized by a collection of MAC-VRFs (one per NVE) residing on the NVEs configured for that EVI.

As an example, consider TS3 and TS5 of Figure 2 above. Assume that connectivity is required between these two TS's where TS3 belongs to the IP-subnet 3 (SN3) whereas TS5 belongs to the IP-subnet 5 (SN5). Both SN3 and SN5 subnets belong to the same tenant. NVE2 has an EVI3 associated with the SN3 and this EVI is represented by a MAC-VRF which is associated with an IP-VRF (for that tenant) via an IRB interface. NVE3 respectively has an EVI5 associated with the SN5 and this EVI is represented by an MAC-VRF which is associated with the same IP-VRF via a different IRB interface.

2.2 Switching among IP subnets in different DCs without GW

This case is similar to that of section 2.1 above albeit for the fact that the TS's belong to different data centers that are interconnected over a WAN (e.g. MPLS/IP PSN). The data centers in question here are seamlessly interconnected to the WAN, i.e., the WAN edge devices do not maintain any TS-specific addresses in the forwarding path - e.g., there is no WAN edge GW(s) between these DCs.

As an example, consider TS3 and TS6 of Figure 2 above. Assume that connectivity is required between these two TS's where TS3 belongs to the SN3 whereas TS6 belongs to the SN6. NVE2 has an EVI3 associated with SN3 and NVE4 has an EVI6 associated with the SN6. Both SN3 and SN6 are part of the same IP-VRF.

2.3 Switching among IP subnets in different DCs with GW

In this scenario, connectivity is required between TS's in different data centers, and those hosts belong to different IP subnets. What makes this case different from that of Section 2.2 is that at least one of the data centers has a gateway as the WAN edge switch. Because of that, the NVE's IP-VRF within that data center need not maintain (host) routes to individual TS's outside of that data center.

As an example, consider a tenant with TS1 and TS5 of Figure 2 above. Assume that connectivity is required between these two TS's where TS1 belongs to the SN1 whereas TS5 belongs to the SN5. NVE3 has an EVI5 associated with the SN5 and this EVI is represented by the MAC-VRF which is connected to the IP-VRF via an IRB interface. NVE1 has an EVI1 associated with the SN1 and this EVI is represented by the MAC-VRF which is connected to the IP-VRF representing the same tenant. Due to the gateway at the edge of DCN 1, NVE1's IP-VRF does not need to have the address of TS5 but instead it has a default route in its IP-VRF with the next-hop being the GW.

2.4 Switching among IP subnets spread across IP-VPN and EVPN networks with GW

In this scenario, connectivity is required between TS's in a data center and hosts in an enterprise site that belongs to a given IP-VPN. The NVE within the data center is an EVPN NVE, whereas the enterprise site has an IP-VPN PE. Furthermore, the data center in question has a gateway as the WAN edge switch. Because of that, the NVE in the data center does not need to maintain individual IP prefixes advertised by enterprise sites (by IP-VPN PEs).

As an example, consider end-station H1 and TS2 of Figure 2. Assume

that connectivity is required between the end-station and the TS, where TS2 belongs to the SN2 that is realized using EVPN, whereas H1 belongs to an IP VPN site connected to PE1 (PE1 maintains an IP-VRF associated with that IP VPN). NVE1 has an EVI2 associated with the SN2. Moreover, EVI2 on NVE1 is connected to an IP-VRF associated with that IP VPN. PE1 originates a VPN-IP route that covers H1. The gateway at the edge of DCN1 performs interworking function between IP-VPN and EVPN. As a result of this, a default route in the IP-VRF on the NVE1, pointing to the gateway as the next hop, and a route to the TS2 (or maybe SN2) on the PE1's IP-VRF are sufficient for the connectivity between H1 and TS2. In this scenario, the NVE1's IP-VRF does not need to maintain a route to H1 because it has the default route to the gateway.

3 Default L3 Gateway for Tenant System

3.1 Homogeneous Environment

This is an environment where all NVEs to which an EVPN instance could potentially be attached (or moved), perform inter-subnet switching. Therefore, inter-subnet traffic can be locally switched by the EVPN NVE connecting the TS's belonging to different subnets.

To support such inter-subnet forwarding, the NVE behaves as an IP Default Gateway from the perspective of the attached TS's. Two models are possible:

1. All the NVEs of a given EVPN instance use the same anycast default gateway IP address and the same anycast default gateway MAC address. On each NVE, this default gateway IP/MAC address correspond to the IRB interface connecting the MAC-VRF of that EVI to the corresponding IP-VRF.

2. Each NVE of a given EVPN instance uses its own default gateway IP and MAC addresses, and these addresses are aliased to the same conceptual gateway through the use of the Default Gateway extended community as specified in [EVPN], which is carried in the EVPN MAC Advertisement routes. On each NVE, this default gateway IP/MAC address correspond to the IRB interface connecting the MAC-VRF of that EVI to the corresponding IP-VRF.

Both of these models enable a packet forwarding paradigm for both symmetric and asymmetric IRB forwarding. In case of asymmetric IRB, a packet is forwarded through the MAC-VRF followed by the IP-VRF on the ingress NVE, and then forwarded through the the MAC-VRF on the egress (disposition) NVE. The egress NVE merely needs to perform a lookup in the associated MAC-VRF and forward the Ethernet frames unmodified, i.e. without rewriting the source MAC address. This is different

from symmetric IRB forwarding where a packet is forwarded through the MAC-VRF followed by the IP-VRF on the ingress NVE, and then forwarded through the IP-VRF followed by the MAC-VRF on the egress NVE.

It is worth noting that if the applications that are running on the TS's are employing or relying on any form of MAC security, then the first model (i.e. using anycast addresses) would be required to ensure that the applications receive traffic from the same source MAC address that they are sending to.

3.2 Heterogeneous Environment

For large data centers with thousands of servers and ToR (or Access) switches, some of them may not have the capability of maintaining or enforcing policies for inter-subnet switching. Even though policies among multiple subnets belonging to same tenant can be simpler, hosts belonging to one tenant can also send traffic to peers belonging to different tenants or security zones. In such scenarios, a WAN edge PE (e.g., L3GW) may not only need to enforce policies for communication among subnets belonging to a single tenant, but also it may need to know how to handle traffic destined towards peers in different tenants. Therefore, there can be a mixed environment where an NVE performs inter-subnet switching for some EVPN instances and the L3GW for others.

4 Operational Models for Asymmetric Inter-Subnet Forwarding

4.1 Among EVPN NVEs within a DC

When an EVPN MAC/IP advertisement route is received by a NVE, the IP address associated with the route is used to populate the IP-VRF table, whereas the MAC address associated with the route is used to populate both the MAC-VRF table, as well as the adjacency associated with the IP route in the IP-VRF table (i.e., ARP table).

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated MAC-VRF for that EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup identifies an adjacency that contains a MAC rewrite and in turn the next-hop (i.e., egress) NVE to which the packet must be forwarded and the associated MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the EVPN MAC route), instead of the MAC address of the next-hop NVE. The ingress NVE then rewrites the

destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) NVE after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the EVPN label that was advertised by the egress NVE. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVPN label to identify the MAC-VRF table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 2 below depicts the packet flow, where NVE1 and NVE2 are the ingress and egress NVEs, respectively.

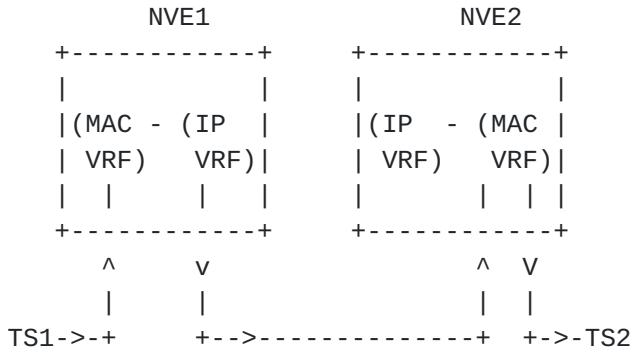


Figure 2: Inter-Subnet Forwarding Among EVPN NVEs within a DC

Note that the forwarding behavior on the egress NVE is similar to EVPN intra-subnet forwarding. In other words, all the packet processing associated with the inter-subnet forwarding semantics is confined to the ingress NVE and that is why it is called Asymmetric IRB.

It should also be noted that [EVPN] provides different level of granularity for the EVPN label. Besides identifying bridge domain table, it can be used to identify the egress interface or a destination MAC address on that interface. If EVPN label is used for egress interface or destination MAC address identification, then no MAC lookup is needed in the egress EVI and the packet can be directly forwarded to the egress interface just based on EVPN label lookup.

4.2 Among EVPN NVEs in Different DCs Without GW

When an EVPN MAC advertisement route is received by a NVE, the IP address associated with the route is used to populate the IP-VRF table, whereas the MAC address associated with the route is used to populate both the MAC-VRF table, as well as the adjacency associated

with the IP route in the IP-VRF table (i.e., ARP table).

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated EVI. If the MAC address corresponds to its IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup identifies an adjacency that contains a MAC rewrite and in turn the next-hop (i.e. egress) Gateway to which the packet must be forwarded along with the associated MPLS label stack. The MAC rewrite holds the MAC address associated with the destination host (as populated by the EVPN MAC route), instead of the MAC address of the next-hop Gateway. The ingress NVE then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address with its IRB Interface MAC address. The ingress NVE, then, forwards the frame to the next-hop (i.e. egress) Gateway after encapsulating it with the MPLS label stack.

Note that this label stack includes the LSP label as well as an EVPN label. The EVPN label could be either advertised by the ingress Gateway, if inter-AS option B is used, or advertised by the egress NVE, if inter-AS option C is used. When the MPLS encapsulated packet is received by the ingress Gateway, the processing again differs depending on whether inter-AS option B or option C is employed: in the former case, the ingress Gateway swaps the EVPN label in the packets with the EVPN label value received from the egress Gateway. In the latter case, the ingress Gateway does not modify the EVPN label and performs normal label switching on the LSP label. Similarly on the egress Gateway, for option B, the egress Gateway swaps the EVPN label with the value advertised by the egress NVE. Whereas, for option C, the egress Gateway does not modify the EVPN label, and performs normal label switching on the LSP label. When the MPLS encapsulated packet is received by the egress NVE, it uses the EVPN label to identify the bridge-domain table. It then performs a MAC lookup in that table, which yields the outbound interface to which the Ethernet frame must be forwarded. Figure 3 below depicts the packet flow.

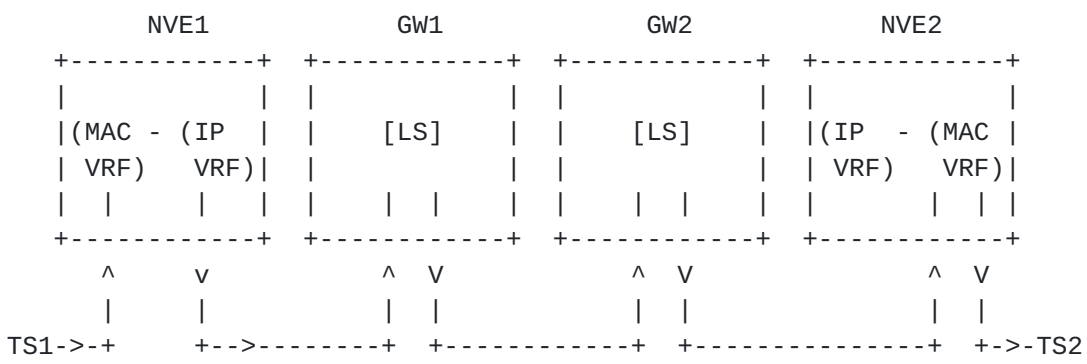


Figure 3: Inter-Subnet Forwarding Among EVPN NVEs in Different DCs without GW

4.3 Among EVPN NVEs in Different DCs with GW

In this scenario, the NVEs within a given data center do not have entries for the MAC/IP addresses of hosts in remote data centers. Rather, the NVEs have a default IP route pointing to the WAN gateway for each VRF. This is accomplished by the WAN gateway advertising for a given EVPN that spans multiple DC a default VPN-IP route that is imported by the NVEs of that VPN that are in the gateway's own DC.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated MAC-VRF table. If the MAC address corresponds to the IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup, in this case, matches the default host route which points to the local WAN gateway. The ingress NVE then rewrites the destination MAC address in the packet with the router's MAC address of the local WAN gateway. It also rewrites the source MAC address with its own IRB Interface MAC address. The ingress NVE, then, forwards the frame to the WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the label for default host route that was advertised by the local WAN gateway. When the MPLS encapsulated packet is received by the local WAN gateway, it uses the default host route label to identify the IP-VRF table. It then performs an IP lookup in that table. The lookup identifies an adjacency that contains a MAC rewrite and in turn the remote WAN gateway (of the remote data center) to which the packet must be forwarded along with the associated MPLS label stack. The MAC rewrite holds the MAC address associated with the ultimate destination host (as populated by the EVPN MAC route). The local WAN gateway then rewrites the destination MAC address in the packet with the address specified in the adjacency. It also rewrites the source MAC address

with its router's MAC address. The local WAN gateway, then, forwards the frame to the remote WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as a EVPN label that was advertised by the remote WAN gateway. When the MPLS encapsulated packet is received by the remote WAN gateway, it simply swaps the EVPN label and forwards the packet to the egress NVE. This implies that the GW1 needs to keep the remote host MAC addresses along with the corresponding EVPN labels in the adjacency entries of the IP-VRF table (i.e., its ARP table). The remote WAN gateway then forward the packet to the egress NVE. The egress NVE then performs a MAC lookup in the MAC-VRF (identified by the received EVPN label) to determine the outbound port to send the traffic on.

Figure 4 below depicts the forwarding model.

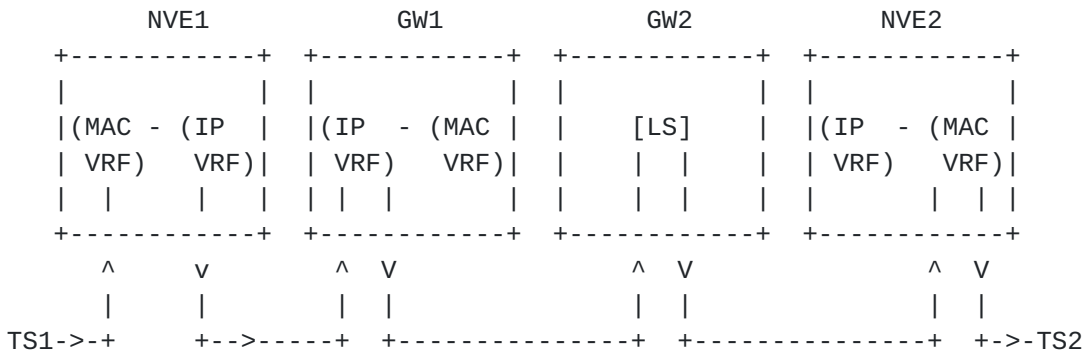


Figure 4: Inter-Subnet Forwarding Among EVPN NVEs in Different DCs with GW

4.4 Among IP-VPN Sites and EVPN NVEs with GW

In this scenario, the NVEs within a given data center do not have entries for the IP addresses of hosts in remote enterprise sites. Rather, the NVEs have a default IP route pointing the WAN gateway for each IP-VRF.

When an Ethernet frame is received by an ingress NVE, it performs a lookup on the destination MAC address in the associated MAC-VRF table. If the MAC address corresponds to the IRB Interface MAC address, the ingress NVE deduces that the packet MUST be inter-subnet routed. Hence, the ingress NVE performs an IP lookup in the associated IP-VRF table. The lookup, in this case, matches the default route which points to the local WAN gateway. The ingress NVE then rewrites the destination MAC address in the packet with the router's MAC address of the local WAN gateway. It also rewrites the

source MAC address with its own IRB Interface MAC address. The ingress NVE, then, forwards the frame to the local WAN gateway after encapsulating it with the MPLS label stack. Note that this label stack includes the LSP label as well as the default host route label that was advertised by the local WAN gateway. When the MPLS encapsulated packet is received by the local WAN gateway, it uses the default host route label to identify the IP-VRF table. It then performs an IP lookup in that table. The lookup identifies the next hop ASBR to which the packet must be forwarded. The local gateway in this case strips the Ethernet encapsulation and perform an IP lookup in its IP-VRF and forwards the IP packet to the ASBR using a label stack comprising of an LSP label and an IP-VPN label that was advertised by the ASBR. When the MPLS encapsulated packet is received by the ASBR, it simply swaps the IP-VPN label with the one advertised by the egress PE. The ASBR then forwards the packet to the egress PE. The egress PE then performs an IP lookup in the IP-VRF (identified by the received IP-VPN label) to determine where to forward the traffic.

Figure 5 below depicts the forwarding model.

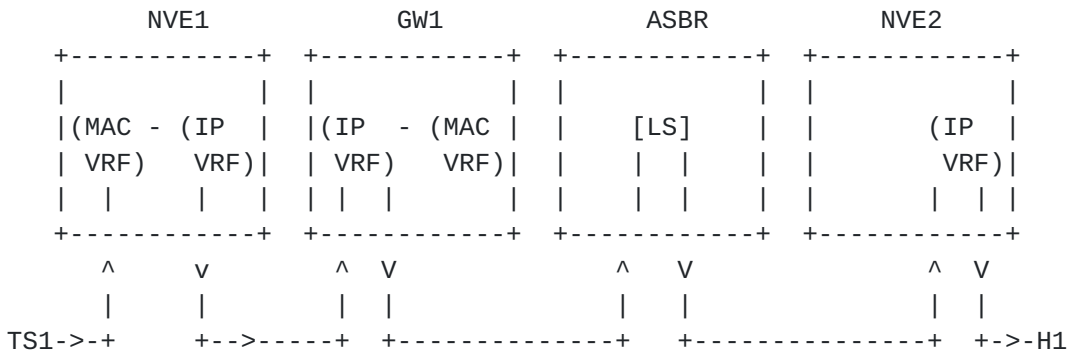


Figure 5: Inter-Subnet Forwarding Among IP-VPN Sites and EVPN NVEs with GW

4.5 Use of Centralized Gateway

In this scenario, the NVEs within a given data center need to forward traffic in L2 to a centralized L3GW for a number of reasons: a) they don't have IRB capabilities or b) they don't have required policy for switching traffic between different tenants or security zones. The centralized L3GW performs both the IRB function for switching traffic among different EVPN instances as well as it performs interworking function when the traffic needs to be switched between IP-VPN sites and EVPN instances.

5 Operational Models for Symmetric Inter-Subnet Forwarding

The following sections describe several main symmetric IRB forwarding scenarios.

5.1 IRB forwarding on NVEs for Tenant Systems

This section covers the symmetric IRB procedures for the scenario where each Tenant System (TS) is attached to one or more NVEs and its host IP and MAC addresses are learned by the attached NVEs and are distributed to all other NVEs that are interested in participating in both intra-subnet and inter-subnet communications with that TS.

In this scenario, for a given tenant (e.g., an IP-VPN instance), an NVE has typically one MAC-VRF for each tenant's subnet (VLAN) that is configured for, Assuming VLAN-based service which is typically the case for VxLAN and NVGRE encapsulation, each MAC-VRF consists of a single bridge domain. In case of MPLS encapsulation with VLAN-aware bundling, then each MAC-VRF consists of multiple bridge domains (one bridge domain per VLAN). The MAC-VRFs on an NVE for a given tenant are associated with an IP-VRF corresponding to that tenant (or IP-VPN instance) via their IRB interfaces.

Each NVE MUST support QoS, Security, and OAM policies per IP-VRF to/from the core network. This is not to be confused with the QoS, Security, and OAM policies per Attachment Circuits (AC) to/from the Tenant Systems. How this requirement is met is an implementation choice and it is outside the scope of this document.

Since VxLAN and NVGRE encapsulations require inner Ethernet header (inner MAC SA/DA), and since for inter-subnet traffic, TS MAC address cannot be used, the ingress NVE's MAC address is used as inner MAC SA. The NVE's MAC address is the device MAC address and it is common across all MAC-VRFs and IP-VRFs. This MAC address is advertised using the new EVPN Router's MAC Extended Community (section 6.1).

Figure below illustrates this scenario where a given tenant (e.g., an IP-VPN instance) has three subnets represented by MAC-VRF1, MAC-VRF2, and MAC-VRF3 across two NVEs. There are five TS's that are associated with these three MAC-VRFs - i.e., TS1, TS4, and TS5 are sitting on the same subnet (e.g., same MAC-VRF/VLAN);where, TS1 and TS5 are associated with MAC-VRF1 on NVE1, TS4 is associated with MAC-VRF1 on NVE2. TS2 is associated with MAC-VRF2 on NVE1, and TS3 is associated with MAC-VRF3 on NVE2. MAC-VRF1 and MAC-VRF2 on NVE1 are in turn associated with IP-VRF1 on NVE1 and MAC-VRF1 and MAC-VRF3 on NVE2 are associated with IP-VRF1 on NVE2. When TS1, TS5, and TS4 exchange traffic with each other, only L2 forwarding (bridging) part of the IRB solution is exercised because all these TS's sit on the same

subnet. However, when TS1 wants to exchange traffic with TS2 or TS3 which belong to different subnets, then both bridging and routing parts of the IRB solution are exercised. The following subsections describe the control and data planes operations for this IRB scenario in details.

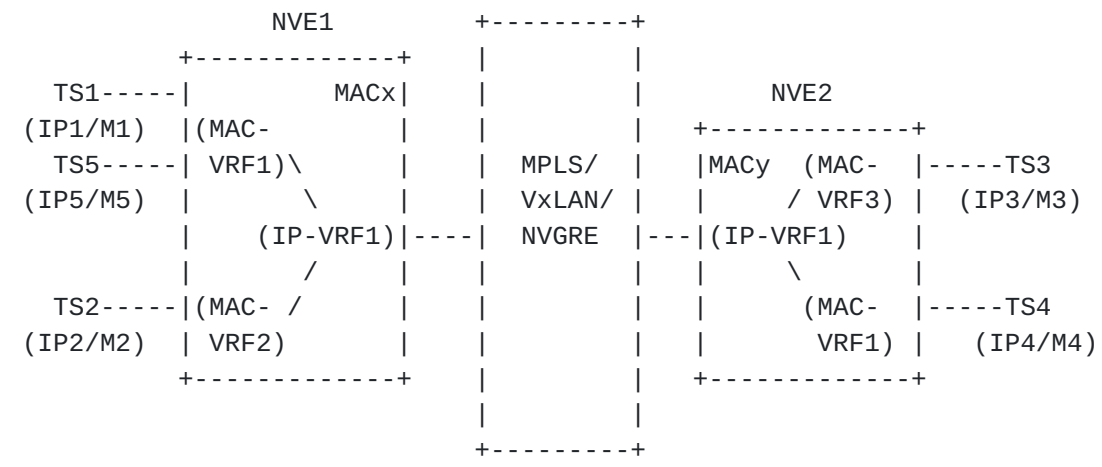


Figure 6: IRB forwarding on NVEs for Tenant Systems

5.1.1 Control Plane Operation

Each NVE advertises a Route Type-2 (RT-2, MAC/IP Advertisement Route) for each of its TS's with the following field set:

- RD and ESI per [EVPN]
- Ethernet Tag = 0; assuming VLAN-based service
- MAC Address Length = 48
- MAC Address = Mi ; where i = 1,2,3,4, or 5 in the above example
- IP Address Length = 32 or 128
- IP Address = IPi ; where i = 1,2,3,4, or 5 in the above example
- Label-1 = MPLS Label or VNID corresponding to MAC-VRF
- Label-2 = MPLS Label or VNID corresponding to IP-VRF

Each NVE advertises an RT-2 route with two Route Targets (one corresponding to its MAC-VRF and the other corresponding to its IP-VRF. Furthermore, the RT-2 is advertised with two BGP Extended Communities. The first BGP Extended Community identifies the tunnel type per section 4.5 of [TUNNEL-ENCAP] and the second BGP Extended Community includes the MAC address of the NVE (e.g., MACx for NVE1 or MACy for NVE2) as defined in section 6.1. This second Extended Community (for the MAC address of NVE) is only required when Ethernet NVO tunnel type is used. If IP NVO tunnel type is used, then there is no need to send this second Extended Community.

Upon receiving this advertisement, the receiving NVE performs the following:

- It uses Route Targets corresponding to its MAC-VRF and IP-VRF for identifying these tables and subsequently importing this route into them.

- It imports the MAC address into the MAC-VRF with BGP Next Hop address as underlay tunnel destination address (e.g., VTEP DA for VxLAN encapsulation) and Label-1 as VNID for VxLAN encapsulation or EVPN label for MPLS encapsulation.

- If the route carries the new Router's MAC Extended Community, and if the receiving NVE is using Ethernet NVO tunnel, then the receiving NVE imports the IP address into IP-VRF with NVE's MAC address (from the new Router's MAC Extended Community) as inner MAC DA and BGP Next Hop address as underlay tunnel destination address, VTEP DA for VxLAN encapsulation and Label-2 as IP-VPN VNID for VxLAN encapsulation.

- If the receiving NVE is going to use MPLS encapsulation, then the receiving NVE imports the IP address into IP-VRF with BGP Next Hop address as underlay tunnel destination address, and Label-2 as IP-VPN label for MPLS encapsulation.

If the receiving NVE receives a RT-2 with only a single Route Target corresponding to IP-VRF and Label-1, then it must discard this route and log an error. If the receiving NVE receives a RT-2 with only a single Route Target corresponding to MAC-VRF but with both Label-1 and Label-2, then it must discard this route and log an error. If the receiving NVE receives a RT-2 with MAC Address Length of zero, then it must discard this route and log an error.

5.1.2 Data Plane Operation - Inter Subnet

The following description of the data-plane operation describes just the logical functions and the actual implementation may differ. Lets consider data-plane operation when TS1 in subnet-1 (MAC-VRF1) on NVE1 wants to send traffic to TS3 in subnet-3 (MAC-VRF3) on NVE2.

- TS1 send a packet with MAC DA corresponding to the MAC-VRF1 IRB interface on NVE1 (the interface between MAC-VRF1 and IP-VRF1), and VLAN-tag corresponding to MAC-VRF1.

- Upon receiving the packet, the NVE1 uses VLAN-tag to identify the MAC-VRF1. It then looks up the MAC DA and forwards the frame to its IRB interface.

- The Ethernet header of the packet is stripped and the packet is fed to the IP-VRF where IP lookup is performed on the destination address. This lookup yields an outgoing interface and the required encapsulation. If the encapsulation is for Ethernet NVO tunnel, then it includes a MAC address to be used as inner MAC DA, an IP address to be used as VTEP DA, and a VPN-ID to be used as VNID.

- The packet is then encapsulated with the proper header based on the above info. The inner MAC SA and VTEP SA is set to NVE's MAC and IP addresses respectively. The packet is then forwarded to the egress NVE.

- On the egress NVE, if the packet arrives on Ethernet NOV tunnel (e.g., it is VxLAN encapsulated), then the VxLAN header is removed. Since the inner MAC DA is the egress NVE's MAC address, the egress NVE knows that it needs to perform an IP lookup. It uses VNID to identify the IP-VRF table and then performs an IP lookup for the destination TS (TS3) which results in access-facing IRB interface over which the packet is sent. Before sending the packet over this interface, the ARP table is consulted to get the destination TS's MAC address.

- The IP packet is encapsulated with an Ethernet header with MAC SA set to that of IRB interface MAC address and MAC DA set to that of destination TS (TS3) MAC address. The packet is sent to the corresponding MAC-VRF3 and after a lookup of MAC DA, is forwarded to the destination TS (TS3) over the corresponding interface.

In this symmetric IRB scenario, inter-subnet traffic between NVEs will always use the IP-VRF VNID/MPLS label. For instance, traffic from TS2 to TS4 will be encapsulated by NVE1 using NVE2's IP-VRF VNID/MPLS label, as long as TS4's host IP is present in NVE1's IP-VRF.

5.1.3 TS Move Operation

When a TS move from one NVE to other, it is important that the MAC mobility procedures are properly executed and the corresponding MAC-VRF and IP-VRF tables on all participating NVEs are updated. [EVPN] describes the MAC mobility procedures for L2-only services for both single-homed TS and multi-homed TS. This section describes the incremental procedures and BGP Extended Communities needed to handle the MAC mobility for a mixed of L2 and L3 connectivity (aka IRB). In order to place the emphasis on the differences between L2-only versus L2-and-L3 use cases, the incremental procedure is described for single-homed TS with the expectation that the reader can easily extrapolate multi-homed TS based on the procedures described in section 15 of [EVPN].

Lets consider TS1 in figure-6 above where it moves from NVE1 to NVE2. In such move, NVE2 discovers IP1/MAC1 of TS1 and realizes that it is a MAC move and it advertises a MAC/IP route per section 5.1.1 above with MAC Mobility Extended Community. In this IRB use case, both MAC and IP addresses of the TS along with their corresponding VNI/MPLS labels are included in the EVPN MAC/IP Advertisement route. Furthermore, besides MAC mobility Extended Community and Route Target corresponding to the MAC-VRF, the following additional BGP Extended Communities are advertised along with the MAC/IP Advertisement route:

- Route Target associated with IP-VRF
- Router's MAC Extended Community
- Tunnel Type Extended Community

Since NVE2 learns TS1's MAC/IP addresses locally, it updates its MAC-VRF1 and IP-VRF1 for TS1 with its local interface.

If the local learning at NVE1 is performed using control or management planes, then these interactions serve as the trigger for NVE1 to withdraw the MAC/IP addresses associated with TS1. However, if the local learning at NVE1 is performed using data-plane learning, then the reception of the MAC/IP Advertisement route (for TS1) from NVE2 with MAC Mobility extended community serve as the trigger for NVE1 to withdraw the MAC/IP addresses associated with TS1.

All other remote NVE devices upon receiving the MAC/IP advertisement route for TS1 from NVE2 with MAC Mobility extended community compare the sequence number in this advertisement with the one previously received. If the new sequence number is greater than the old one, then they update the MAC/IP addresses of TS1 in their corresponding MAC-VRFs and IP-VRFs to point to NVE2. Furthermore, upon receiving the MAC/IP withdraw for TS1 from NVE1, these remote PEs perform the cleanups for their BGP tables.

5.2 IRB forwarding on NVEs for Subnets behind Tenant Systems

This section covers the symmetric IRB procedures for the scenario where some Tenant Systems (TS's) support one or more subnets and these TS's are associated with one ore more NVEs. Therefore, besides the advertisement of MAC/IP addresses for each TS which can be in the presence of All-Active multi-homing, the associated NVE needs to also advertise the subnets behind each TS.

The main difference between this scenario and the previous one is the additional advertisement corresponding to each subnet. These subnet advertisements are accomplished using EVPN IP Prefix route defined in [EVPN-PREFIX]. These subnet prefixes are advertised with the IP

address of their associated TS (which is in overlay address space) as their next hop. The receiving NVEs perform recursive route resolution to resolve the subnet prefix with its associated ingress NVE so that they know which NVE to forward the packets to when they are destined for that subnet prefix.

The advantage of this recursive route resolution is that when a TS moves from one NVE to another, there is no need to re-advertise any of the subnet prefixes for that TS. All it is needed is to advertise the IP/MAC addresses associated with the TS itself and exercise MAC mobility procedures for that TS. The recursive route resolution automatically takes care of the updates for the subnet prefixes of that TS.

Figure below illustrates this scenario where a given tenant (e.g., an IP-VPN service) has three subnets represented by MAC-VRF1, MAC-VRF2, and MAC-VRF3 across two NVEs. There are four TS's associated with these three MAC-VRFs - i.e., TS1, TS5 are connected to MAC-VRF1 on NVE1, TS2 is connected to MAC-VRF2 on NVE1, TS3 is connected to MAC-VRF3 on NVE2, and TS4 is connected to MAC-VRF1 on NVE2. TS1 has two subnet prefixes (SN1 and SN2) and TS3 has a single subnet prefix, SN3. The MAC-VRFs on each NVE are associated with their corresponding IP-VRF using their IRB interfaces. When TS4 and TS1 exchange intra-subnet traffic, only L2 forwarding (bridging) part of the IRB solution is used (i.e., the traffic only goes through their MAC-VRFs); however, when TS3 wants to forward traffic to SN1 or SN2 sitting behind TS1 (inter-subnet traffic), then both bridging and routing parts of the IRB solution are exercised (i.e., the traffic goes through the corresponding MAC-VRFs and IP-VRFs). The following subsections describe the control and data planes operations for this IRB scenario in details.

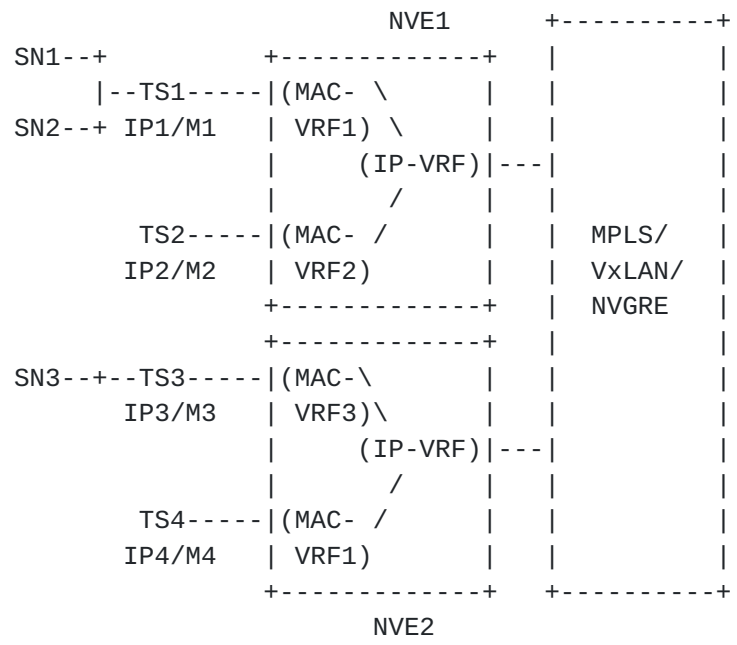


Figure 7: IRB forwarding on NVEs for Tenant Systems with configured subnets

5.2.1 Control Plane Operation

Each NVE advertises a Route Type-5 (RT-5, IP Prefix Route defined in [EVPN-PREFIX]) for each of its subnet prefixes with the IP address of its TS as the next hop (gateway address field) as follow:

- RD per VPN
- ESI = 0
- Ethernet Tag = 0;
- IP Prefix Length = 32 or 128
- IP Prefix = SNi
- Gateway Address = IPi; IP address of TS
- Label = 0

This RT-5 is advertised with a Route Target corresponding to the IP-VPN service.

Each NVE also advertises an RT-2 (MAC/IP Advertisement Route) along with their associated Route Targets and Extended Communities for each of its TS's exactly as described in section 5.1.1.

Upon receiving the RT-5 advertisement, the receiving NVE performs the following:

- It uses the Route Target to identify the corresponding IP-VRF

- It imports the IP prefix into its corresponding IP-VRF with the IP address of the associated TS as its next hop.

Upon receiving the RT-2 advertisement, the receiving NVE imports MAC/IP addresses of the TS into the corresponding MAC-VRF and IP-VRF per section 5.1.1. Furthermore, it performs recursive route resolution to resolve the IP prefix (received in RT-5) to its corresponding NVE's IP address (e.g., its BGP next hop). BGP next hop will be used as underlay tunnel destination address (e.g., VTEP DA for VxLAN encapsulation) and Router's MAC will be used as inner MAC for VxLAN encapsulation.

5.2.2 Data Plane Operation

The following description of the data-plane operation describes just the logical functions and the actual implementation may differ. Lets consider data-plane operation when a host on SN1 sitting behind TS1 wants to send traffic to a host sitting behind SN3 behind TS3.

- TS1 send a packet with MAC DA corresponding to the MAC-VRF1 IRB interface of NVE1, and VLAN-tag corresponding to MAC-VRF1.

- Upon receiving the packet, the ingress NVE1 uses VLAN-tag to identify the MAC-VRF1. It then looks up the MAC DA and forwards the frame to its IRB interface just like section 5.1.1.

- The Ethernet header of the packet is stripped and the packet is fed to the IP-VRF; where, IP lookup is performed on the destination address. This lookup yields the fields needed for VxLAN encapsulation with NVE2's MAC address as the inner MAC DA, NVE'2 IP address as the VTEP DA, and the VNID. MAC SA is set to NVE1's MAC address and VTEP SA is set to NVE1's IP address.

- The packet is then encapsulated with the proper header based on the above info and is forwarded to the egress NVE (NVE2).

- On the egress NVE (NVE2), assuming the packet is VxLAN encapsulated, the VxLAN and the inner Ethernet headers are removed and the resultant IP packet is fed to the IP-VRF associated with that the VNID.

- Next, a lookup is performed based on IP DA (which is in SN3) in the associated IP-VRF of NVE2. The IP lookup yields the access-facing IRB interface over which the packet needs to be sent. Before sending the packet over this interface, the ARP table is consulted to get the destination TS (TS3) MAC address.

- The IP packet is encapsulated with an Ethernet header with the MAC SA set to that of the access-facing IRB interface of the egress NVE (NVE2) and the MAC DA is set to that of destination TS (TS3) MAC address. The packet is sent to the corresponding MAC-VRF3 and after a lookup of MAC DA, is forwarded to the destination TS (TS3) over the corresponding interface.

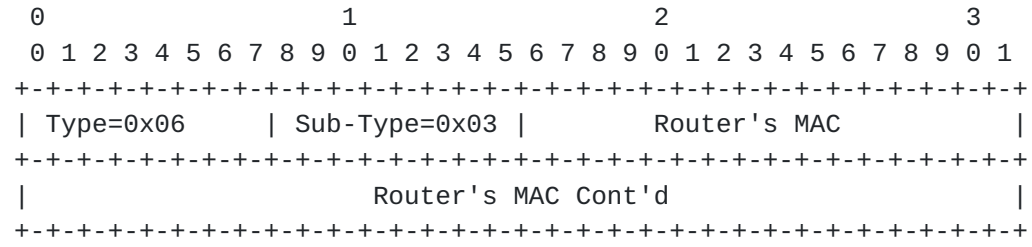
6 BGP Encoding

This document defines one new BGP Extended Community for EVPN.

6.1 Router's MAC Extended Community

A new EVPN BGP Extended Community called Router's MAC is introduced here. This new extended community is a transitive extended community with the Type field of 0x06 (EVPN) and the Sub-Type of 0x03. It may be advertised along with BGP Encapsulation Extended Community define in section 4.5 of [TUNNEL-ENCAP].

The Router's MAC Extended Community is encoded as an 8-octet value as follows:



This extended community is used to carry the NVE's MAC address for symmetric IRB scenarios and it is sent with RT-2 as described in section 5.1.1 and 5.2.1.

7 TS Mobility

7.1 TS Mobility & Optimum Forwarding for TS Outbound Traffic

Optimum forwarding for the TS outbound traffic, upon TS mobility, can be achieved using either the anycast default Gateway MAC and IP addresses, or using the address aliasing as discussed in [DC-MOBILITY].

7.2 TS Mobility & Optimum Forwarding for TS Inbound Traffic

For optimum forwarding of the TS inbound traffic, upon TS mobility, all the NVEs and/or IP-VPN PEs need to know the up to date location of the TS. Two scenarios must be considered, as discussed next.

In what follows, we use the following terminology:

- source NVE refers to the NVE behind which the TS used to reside prior to the TS mobility event.

- target NVE refers to the new NVE behind which the TS has moved after the mobility event.

7.2.1 Mobility without Route Aggregation

In this scenario, when a target NVE detects that a MAC mobility event has occurred, it initiates the MAC mobility handshake in BGP as specified in section 5.1.3. The WAN Gateways, acting as ASBRs in this case, re-advertise the MAC route of the target NVE with the MAC Mobility extended community attribute unmodified. Because the WAN Gateway for a given data center re-advertises BGP routes received from the WAN into the data center, the source NVE will receive the MAC Advertisement route of the target NVE (with the next hop attribute adjusted depending on which inter-AS option is employed). The source NVE will then withdraw its original MAC Advertisement route as a result of evaluating the Sequence Number field of the MAC Mobility extended community in the received MAC Advertisement route. This is per the procedures already defined in [EVPN].

8 Acknowledgements

The authors would like to thank Sami Boutros for his valuable comments.

9 Security Considerations

The security considerations discussed in [EVPN] apply to this document.

10 IANA Considerations

IANA has allocated a new transitive extended community Type of 0x06 and Sub-Type of 0x03 for EVPN Router's MAC Extended Community.

11 References

11.1 Normative References

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

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

[TUNNEL-ENCAP] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-03, November 2016.

[EVPN-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-03, September, 2016.

11.2 Informative References

[802.1Q] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q(tm), 2014 Edition, November 2014.

[EVPN-IPVPN-INTEROP] Sajassi et al., "EVPN Seamless Interoperability with IP-VPN", draft-sajassi-l2vpn-evpn-ipvpn-interop-01, work in progress, October, 2012.

[DC-MOBILITY] Aggarwal et al., "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-05.txt, work in progress, June, 2013.

12 Contributors

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

Samer Salam
Florin Balus
Cisco

Yakov Rekhter
Juniper

Wim Henderickx
Nokia

Linda Dunbar
Huawei

Dennis Cai
Alibaba

Authors' Addresses

Ali Sajassi (Editor)
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
Email: sslam@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com

John E. Drake
Juniper Networks
Email: jdrake@juniper.net

Lucy Yong
Huawei Technologies
Email: lucy.yong@huawei.com

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