BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: August 9, 2021

A. Sajassi
S. Salam
S. Thoria
Cisco Systems
J. Drake
Juniper
J. Rabadan
Nokia

February 5, 2021

Integrated Routing and Bridging in EVPN
draft-ietf-bess-evpn-inter-subnet-forwarding-12

Abstract

Ethernet VPN (EVPN) provides an extensible and flexible multi-homing VPN solution over an MPLS/IP network for intra-subnet connectivity among Tenant Systems and End Devices that can be physical or virtual. However, there are scenarios for which there is a need for a dynamic and efficient inter-subnet connectivity among these Tenant Systems and End Devices 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.

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 RFC 2119 [RFC2119] and RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 9, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Terminology 3
2. Introduction 4
3. EVPN PE Model for IRB Operation 6
4. Symmetric and Asymmetric IRB 7
4.1. IRB Interface and its MAC and IP addresses 10
5. Symmetric IRB Procedures 12
5.1. Control Plane - Advertising PE 12
5.2. Control Plane - Receiving PE 13
5.3. Subnet route advertisement 14
5.4. Data Plane - Ingress PE 15
5.5. Data Plane - Egress PE 15
6. Asymmetric IRB Procedures 16
6.1. Control Plane - Advertising PE 16
6.2. Control Plane - Receiving PE 17
6.3. Data Plane - Ingress PE 18
6.4. Data Plane - Egress PE 18
7. Mobility Procedure 19
7.1. Initiating a gratutious ARP upon a Move 20
7.2. Sending Data Traffic without an ARP Request 21
7.3. Silent Host 22
8. BGP Encoding 23
8.1. Router's MAC Extended Community 23
9. Operational Models for Symmetric Inter-Subnet Forwarding . . 24
9.1. IRB forwarding on NVEs for Tenant Systems 24
9.1.1. Control Plane Operation 25
9.1.2. Data Plane Operation 27
9.2. IRB forwarding on NVEs for Subnets behind Tenant Systems 28
9.2.1. Control Plane Operation 29
9.2.2. Data Plane Operation 30

10. Acknowledgements 31
11. Security Considerations 32
12. IANA Considerations 32
13. References 32
13.1. Normative References 33
13.2. Informative References 34
Authors' Addresses 34

1. Terminology

AC: Attachment Circuit

ARP: Address Resolution Protocol

BD: Broadcast Domain. As per [RFC7432], an EVI consists of a single or multiple BDs. In the case of VLAN-bundle and VLAN-based service models (see [RFC7432]), a BD is equivalent to an EVI. In the case of VLAN-aware bundle service model, an EVI contains multiple BDs. Also, in this document, BD and subnet are equivalent terms and wherever "subnet" is used, it means "IP subnet"

BD Route Target: refers to the Broadcast Domain assigned Route Target [RFC4364]. In the case of VLAN-aware bundle service model, all the BD instances in the MAC-VRF share the same Route Target

BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per [RFC7432].

Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload as specified for VxLAN in [RFC7348] and for NVGRE in [RFC7637].

EVI: EVPN Instance spanning the NVE/PE devices that are participating on that EVPN, as per [RFC7432].

EVPN: Ethernet Virtual Private Networks, as per [RFC7432].

IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload) as specified for GPE in [I-D.ietf-nvo3-vxlan-gpe].

IP-VRF: A Virtual Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by EVPN and IP-VPN address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.

IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on an NVE/PE, as per [RFC7432]. A MAC-VRF is also an instantiation of an EVI in an NVE/PE.

ND: Neighbor Discovery Protocol

NVE: Network Virtualization Edge

NVGRE: Network Virtualization Generic Routing Encapsulation, [RFC7637]

NVO: Network Virtualization Overlays

RT-2: EVPN route type 2, i.e., MAC/IP Advertisement route, as defined in [RFC7432]

RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [I-D.ietf-bess-evpn-prefix-advertisement]

TS: Tenant System

VA: Virtual Appliance

VNI: Virtual Network Identifier. As in [RFC8365], the term is used as a representation of a 24-bit NVO instance identifier, with the understanding that VNI will refer to a VXLAN Network Identifier in VXLAN, or Virtual Subnet Identifier in NVGRE, etc. unless it is stated otherwise.

VTEP: VXLAN Termination End Point, as in [RFC7348].

VXLAN: Virtual Extensible LAN, as in [RFC7348].

This document also assumes familiarity with the terminology of [RFC7432], [RFC8365] and [RFC7365].

2. Introduction

EVPN [RFC7432] provides an extensible and flexible multi-homing VPN solution over an MPLS/IP network for intra-subnet connectivity among Tenant Systems (TSes) and End Devices that can be physical or virtual; where an IP subnet is represented by an EVPN Instance (EVI) for a VLAN-based service or by an (EVI, VLAN) for a VLAN-aware bundle service. However, there are scenarios for which there is a need for a dynamic and efficient inter-subnet connectivity among these Tenant Systems and End Devices 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) devices where all the inter-subnet forwarding is performed and all the inter-subnet communication policies are enforced. When two TSes belonging to two different subnets connected to the same PE wanted to communicate with each other, their traffic needed to be backhauled from the PE all the way to the centralized gateway where inter-subnet switching is performed and then back to the PE. For today's large multi-tenant data center, this scheme is very inefficient and sometimes impractical.

In order to overcome the drawback of the centralized layer-3 GW approach, IRB functionality is needed on the PEs (also referred to as EVPN NVEs) attached to TSes in order to avoid inefficient forwarding of tenant traffic (i.e., avoid back-hauling and hair-pinning). When a PE with IRB capability receives tenant traffic over an Attachment Circuit (AC), it can not only locally bridge the tenant intra-subnet traffic but also can locally route the tenant inter-subnet traffic on a packet by packet basis thus meeting the requirements for both intra and inter-subnet forwarding and avoiding non-optimal traffic forwarding associated with centralized layer-3 GW approach.

Some TSes 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 and intra-subnet traffic and to route inter-subnet IP traffic. Therefore, the solution needs to meet the following requirements:

R1: 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. It should be noted that if an IP-VRF in a NVE is configured for IPv6 and that NVE receives IPv4 traffic on the corresponding VLAN, then the IPv4 traffic is treated as L2 traffic and it is bridged. Also vise versa, if an IP-VRF in a NVE is configured for IPv4 and that NVE receives IPv6 traffic on the corresponding VLAN, then the IPv6 traffic is treated as L2 traffic and it is bridged.

R2: The solution must support bridging for non-IP traffic.

R3: The solution must allow inter-subnet switching to be disabled on a per VLAN basis on PEs where the traffic needs to be backhauled to another node (i.e., for performing FW or DPI functionality).

3. EVPN PE Model for IRB Operation

Since this document discusses IRB operation in relationship to EVPN MAC-VRF, IP-VRF, EVI, Bridge Domain (BD), Bridge Table (BT), and IRB interfaces, it is important to understand the relationship between these components. Therefore, the following PE model is illustrated below to a) describe these components and b) illustrate the relationship among them.

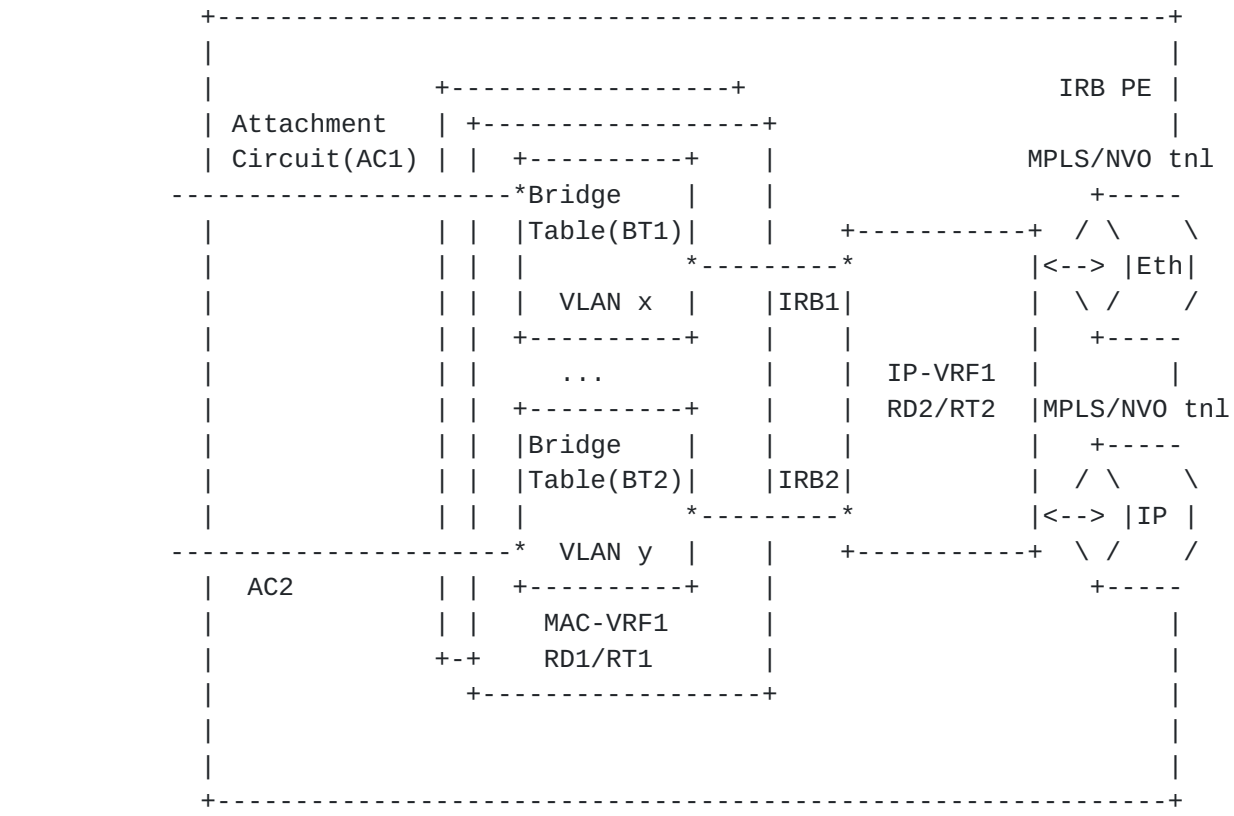


Figure 1: EVPN IRB PE Model

A tenant needing IRB services on a PE, requires an IP Virtual Routing and Forwarding table (IP-VRF) along with one or more MAC Virtual Routing and Forwarding tables (MAC-VRFs). An IP-VRF, as defined in [RFC4364], is the instantiation of an IPVPN instance in a PE. A MAC-VRF, as defined in [RFC7432], is the instantiation of an EVI (EVPN Instance) in a PE. A MAC-VRF consists of one or more Bridge Tables (BTs) where each BT corresponds to a VLAN (broadcast domain - BD). If service interfaces for an EVPN PE are configured in VLAN- Based mode (i.e., section 6.1 of RFC7432), then there is only a single BT per MAC-VRF (per EVI) - i.e., there is only one tenant VLAN per EVI. However, if service interfaces for an EVPN PE are configured in VLAN-

Aware Bundle mode (i.e., section 6.3 of RFC7432), then there are several BTs per MAC-VRF (per EVI) - i.e., there are several tenant VLANs per EVI.

Each BT is connected to an IP-VRF via an L3 interface called IRB interface. Since a single tenant subnet is typically (and in this document) represented by a VLAN (and thus supported by a single BT), for a given tenant there are as many BTs as there are subnets and thus there are also as many IRB interfaces between the tenant IP-VRF and the associated BTs as shown in the PE model above.

IP-VRF is identified by its corresponding route target and route distinguisher and MAC-VRF is also identified by its corresponding route target and route distinguisher. If operating in EVPN VLAN-Based mode, then a receiving PE that receives an EVPN route with MAC-VRF route target can identify the corresponding BT; however, if operating in EVPN VLAN-Aware Bundle mode, then the receiving PE needs both the MAC-VRF route target and VLAN ID in order to identify the corresponding BT.

4. Symmetric and Asymmetric IRB

This document defines and describes two types of IRB solutions - namely symmetric and asymmetric IRB. The description of symmetric and asymmetric IRB procedures relating to data path operations and tables in this document is a logical view of data path lookups and related tables. Actual implementations, while following this logical view, may not strictly adhere to it for performance tradeoffs. Specifically,

- references to ARP table in the context of asymmetric IRB is a logical view of a forwarding table that maintains an IP to MAC binding entry on a layer 3 interface for both IPv4 and IPv6. These entries are not subject to ARP or ND protocol. For IP to MAC bindings learnt via EVPN, an implementation may choose to import these bindings directly to the respective forwarding table (such as an adjacency/next-hop table) as opposed to importing them to ARP or ND protocol tables.

- references to host IP lookup followed by a host MAC lookup in the context of asymmetric IRB MAY be collapsed into a single IP lookup in a hardware implementation.

In symmetric IRB as its name implies, the lookup operation is symmetric at both ingress and egress PEs - i.e., both ingress and egress PEs perform lookups on both MAC and IP addresses. The ingress PE performs a MAC lookup followed by an IP lookup and the egress PE

performs an IP lookup followed by a MAC lookup as depicted in the following figure.

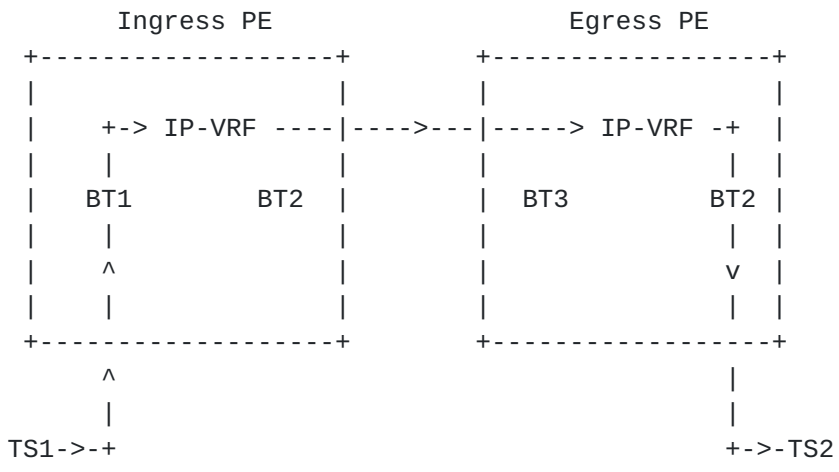


Figure 2: Symmetric IRB

In symmetric IRB as shown in figure-2, the inter-subnet forwarding between two PEs is done between their associated IP-VRFs. Therefore, the tunnel connecting these IP-VRFs can be either IP-only tunnel (e.g., in case of MPLS or GPE encapsulation) or Ethernet NVO tunnel (e.g., in case of VxLAN encapsulation). If it is an Ethernet NVO tunnel, the TS1's IP packet is encapsulated in an Ethernet header consisting of ingress and egress PEs MAC addresses - i.e., there is no need for ingress PE to use the destination TS2's MAC address. Therefore, in symmetric IRB, there is no need for the ingress PE to maintain ARP entries for destination TS2's IP and MAC addresses association in its ARP table. Each PE participating in symmetric IRB only maintains ARP entries for locally connected hosts and maintains MAC-VRFs/BTs for only locally configured subnets.

In asymmetric IRB, the lookup operation is asymmetric and the ingress PE performs three lookups; whereas the egress PE performs a single lookup - i.e., the ingress PE performs a MAC lookup, followed by an IP lookup, followed by a MAC lookup again; whereas, the egress PE performs just a single MAC lookup as depicted in figure 3 below.

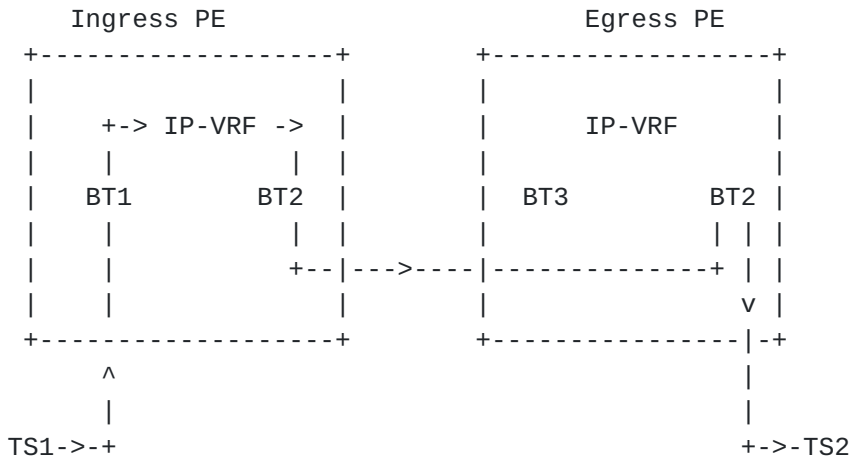


Figure 3: Asymmetric IRB

In asymmetric IRB as shown in figure-3, the inter-subnet forwarding between two PEs is done between their associated MAC-VRFs/BTs. Therefore, the MPLS or NVO tunnel used for inter-subnet forwarding MUST be of type Ethernet. Since only MAC lookup is performed at the egress PE (e.g., no IP lookup), the TS1's IP packets need to be encapsulated with the destination TS2's MAC address. In order for ingress PE to perform such encapsulation, it needs to maintain TS2's IP and MAC address association in its ARP table. Furthermore, it needs to maintain destination TS2's MAC address in the corresponding BT even though it may not have any TSes of the corresponding subnet locally attached. In other words, each PE participating in asymmetric IRB MUST maintain ARP entries for remote hosts (hosts connected to other PEs) as well as maintain MAC-VRFs/BTs and IRB interfaces for ALL subnets in an IP VRF including subnets that may not be locally attached. Therefore, careful consideration of PE scale aspects for its ARP table size, its IRB interfaces, number and size of its bridge tables should be given for the application of asymmetric IRB.

It should be noted that whenever a PE performs a host IP lookup for a packet that is routed, IPv4 TTL or IPv6 hop limit for that packet is decremented by one and if it reaches zero, the packet is discarded. In the case of symmetric IRB, the TTL/hop limit is decremented by both ingress and egress PEs (once by each); whereas, in the case of asymmetric IRB, the TTL/hop limit is decremented only once by the ingress PE.

The following subsection defines the control and data planes procedures for symmetric and asymmetric IRB on ingress and egress PEs. The following figure is used to describe these procedures where it shows a single IP-VRF and a number of BTs on each PE for a given tenant. The IP-VRF of the tenant (i.e., IP-VRF1) is connected to

each BT via its associated IRB interface. Each BT on a PE is associated with a unique VLAN (e.g., with a BD) where in turn it is associated with a single MAC-VRF in the case of VLAN-Based mode or a number of BTs can be associated with a single MAC-VRF in the case of VLAN-Aware Bundle mode. Whether the service interface on a PE is VLAN-Based or VLAN-Aware Bundle mode does not impact the IRB operation and procedures. It mainly impacts the setting of Ethernet tag field in EVPN BGP routes as described in section 6 of [RFC7432].

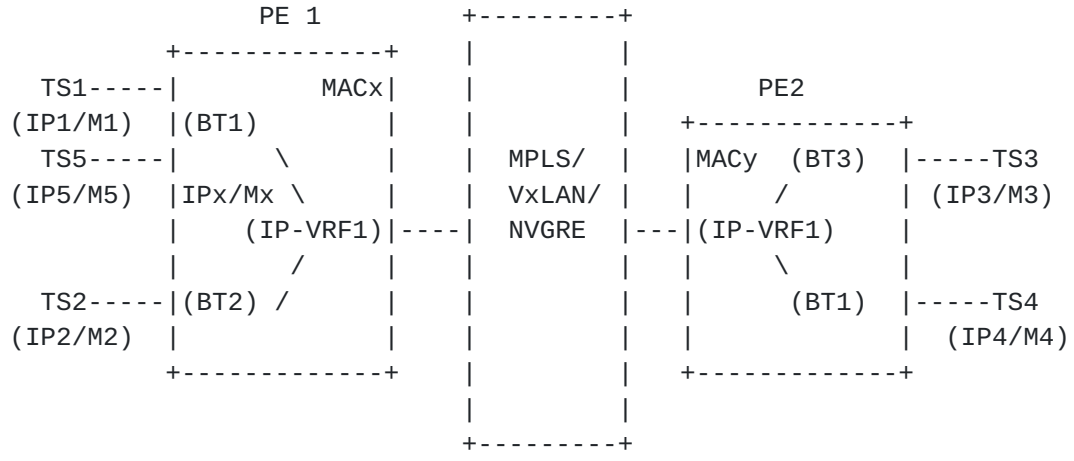


Figure 4: IRB forwarding

4.1. IRB Interface and its MAC and IP addresses

To support inter-subnet forwarding on a PE, the PE acts as an IP Default Gateway from the perspective of the attached Tenant Systems where default gateway MAC and IP addresses are configured on each IRB interface associated with its subnet and falls into one of the following two options:

1. All the PEs for a given tenant subnet use the same anycast default gateway IP and MAC addresses. On each PE, this default gateway IP and MAC addresses correspond to the IRB interface connecting the BT associated with the tenant's VLAN to the corresponding tenant's IP-VRF.

2. Each PE for a given tenant subnet uses the same anycast default gateway IP address but its own MAC address. These MAC addresses are aliased to the same anycast default gateway IP address through the use of the Default Gateway extended community as specified in [RFC7432], which is carried in the EVPN MAC/IP Advertisement routes. On each PE, this default gateway IP address along with its associated MAC addresses correspond to the

IRB interface connecting the BT associated with the tenant's VLAN to the corresponding tenant's IP-VRF.

It is worth noting that if the applications that are running on the TSes are employing or relying on any form of MAC security, then the first option (i.e. using anycast MAC address) should be used to ensure that the applications receive traffic from the same IRB interface MAC address that they are sending to. If the second option is used, then the IRB interface MAC address MUST be the one used in the initial ARP reply or ND Neighbor Advertisement (NA)for that TS.

Although both of these options are applicable to both symmetric and asymmetric IRB, the option-1 is recommended because of the ease of anycast MAC address provisioning on not only the IRB interface associated with a given subnet across all the PEs corresponding to that VLAN but also on all IRB interfaces associated with all the tenant's subnets across all the PEs corresponding to all the VLANs for that tenant. Furthermore, it simplifies the operation as there is no need for Default Gateway extended community advertisement and its associated MAC aliasing procedure. Yet another advantage is that following host mobility, the host does not need to refresh the default GW ARP/ND entry.

If option-1 is used, an implementation MAY choose to auto-derive the anycast MAC address. If auto-derivation is used, the anycast MAC MUST be auto-derived out of the following ranges (which are defined in [RFC5798]):

- Anycast IPv4 IRB case: 00-00-5E-00-01-{VRID} (in hex, in Internet standard bit-order)

- Anycast IPv6 IRB case: 00-00-5E-00-02-{VRID} (in hex, in Internet standard bit-order)

Where the last octet is generated based on a configurable Virtual Router ID (VRID, range 1-255)). If not explicitly configured, the default value for the VRID octet is '1'. Auto-derivation of the anycast MAC can only be used if there is certainty that the auto-derived MAC does not collide with any customer MAC address.

In addition to IP anycast addresses, IRB interfaces can be configured with non-anycast IP addresses for the purpose of OAM (such as traceroute/ping to these interfaces) for both symmetric and asymmetric IRB. These IP addresses need to be distributed as VPN routes when PEs operate in symmetric IRB mode. However, they don't need to be distributed if the PEs are operating in asymmetric IRB mode as the non-anycast IP addresses are configured along with their

individual MACs and they get distributed via EVPN route type-2 advertisement.

For option-1, irrespective of using only the anycast MAC address or both anycast and non-anycast MAC addresses (where the latter one is used for the purpose of OAM) on the same IRB, when a TS sends an ARP request or ND Neighbor Solicitation (NS) to the PE that is attached to, the request is sent for the anycast IP address of the IRB interface associated with the TS's subnet and then the reply will use anycast MAC address (in both Source MAC in the Ethernet header and Sender hardware address in the payload). For example, in figure 4, TS1 is configured with the anycast IPx address as its default gateway IP address and thus when it sends an ARP request for IPx (anycast IP address of the IRB interface for BT1), the PE1 sends an ARP reply with the MACx which is the anycast MAC address of that IRB interface. Traffic routed from IP-VRF1 to TS1 uses the anycast MAC address as source MAC address.

5. Symmetric IRB Procedures

5.1. Control Plane - Advertising PE

When a PE (e.g., PE1 in figure 4 above) learns MAC and IP address of a TS (e.g., via an ARP request or Neighbor Solicitation), it adds the MAC address to the corresponding MAC-VRF/BT of that tenant's subnet and adds the IP address to the IP-VRF for that tenant. Furthermore, it adds this TS's MAC and IP address association to its ARP table or NDP cache. It then builds an EVPN MAC/IP Advertisement route (type 2) as follows and advertises it to other PEs participating in that tenant's VPN.

- The Length field of the BGP EVPN NLRI for an EVPN MAC/IP Advertisement route MUST be either 40 (if IPv4 address is carried) or 52 (if IPv6 address is carried).

- Route Distinguisher (RD), Ethernet Segment Identifier, Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, IP Address, and MPLS Label1 fields MUST be set per [RFC7432] and [RFC8365].

- The MPLS Label2 field is set to either an MPLS label or a VNI corresponding to the tenant's IP-VRF. In the case of an MPLS label, this field is encoded as 3 octets, where the high-order 20 bits contain the label value.

Just as in [RFC7432], the RD, Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, and IP Address fields are part of the

route key used by BGP to compare routes. The rest of the fields are not part of the route key.

This route is advertised along with the following two extended communities:

1. Encapsulation Extended Community

2. Router's MAC Extended Community

For symmetric IRB mode, Router's MAC EC is needed to carry the PE's overlay MAC address (e.g., inner MAC address in NVO encapsulation) which is used for IP-VRF to IP-VRF communications with Ethernet NVO tunnel. If MPLS or IP-only NVO tunnel is used, then there is no need to send Router's MAC Extended Community along with this route.

This route MUST be advertised with two route targets, one corresponding to the MAC-VRF of the tenant's subnet and another corresponding to the tenant's IP-VRF.

5.2. Control Plane - Receiving PE

When a PE (e.g., PE2 in figure 4 above) receives this EVPN MAC/IP Advertisement route, it performs the following:

- Using MAC-VRF Route Target (and Ethernet Tag if different from zero), it identifies the corresponding MAC-VRF (and BT). If the MAC- VRF (and BT) exists (e.g., it is locally configured) then it imports the MAC address into it. Otherwise, it does not import the MAC address.

- Using IP-VRF route target, it identifies the corresponding IP-VRF and imports the IP address into it.

The inclusion of MPLS label2 field in this route signals to the receiving PE that this route is for symmetric IRB mode and MPLS label2 needs to be installed in forwarding path to identify the corresponding IP-VRF.

If the receiving PE receives this route with both the MAC-VRF and IP-VRF route targets but the MAC/IP Advertisement route does not include MPLS label2 field and if the receiving PE supports asymmetric IRB mode, then the receiving PE installs the MAC address in the corresponding MAC-VRF and (IP, MAC) association in the ARP table for that tenant (identified by the corresponding IP-VRF route target).

If the receiving PE receives this route with both the MAC-VRF and IP-VRF route targets and if the receiving PE does not support either

asymmetric or symmetric IRB modes, then if it has the corresponding MAC-VRF, it only imports the MAC address. Otherwise, if it doesn't have the corresponding MAC-VRF, it must not import this route.

If the receiving PE receives this route with both the MAC-VRF and IP-VRF route targets and the MAC/IP Advertisement route includes MPLS label2 field but the receiving PE only supports asymmetric IRB mode, then the receiving PE MUST ignore MPLS label2 field and install the MAC address in the corresponding MAC-VRF and (IP, MAC) association in the ARP table for that tenant (identified by the corresponding IP-VRF route target).

5.3. Subnet route advertisement

In the case of symmetric IRB, a layer-3 subnet and IRB interface corresponding to a MAC-VRF/BT is required to be provisioned at a PE only if that PE has locally attached hosts in that subnet. In order to enable inter-subnet routing across PEs in a deployment where not all subnets are provisioned at all PEs participating in an EVPN IRB instance, PEs MUST advertise local subnet routes as EVPN RT-5. These subnet routes are required for bootstrapping host (MAC,IP) learning using gleaning procedures initiated by an inter-subnet data packet.

Consider a subnet A that is locally attached to PE1 and subnet B that is locally attached to PE2 and to PE3. Host A in subnet A, that is attached to PE1 initiates a data packet destined to host B in subnet B that is attached to PE3. If host B's (MAC, IP) has not yet been learnt either via a gratuitous ARP OR via a prior gleaning procedure, a new gleaning procedure MUST be triggered for host B's (MAC, IP) to be learnt and advertised across the EVPN network. Since host B's subnet is not local to PE1, an IP lookup for host B at PE1 will not trigger this gleaning procedure for host B's (MAC, IP). Therefore, PE1 MUST learn subnet B's prefix route via EVPN RT-5 advertised from PE2 and PE3, so it can route the packet to one of the PEs that have subnet B locally attached. Once the packet is received at PE2 OR PE3, and the route lookup yields a glean result, an ARP request is triggered and flooded across the layer-2 overlay. This ARP request would be received and replied to by host B, resulting in host B (MAC, IP) learning at PE3, and its advertisement across the EVPN network. Packets from host A to host B can now be routed directly from PE1 to PE3. Advertisement of local subnet EVPN RT-5 for an IP VRF MAY typically be achieved via provisioning connected route redistribution to BGP.

5.4. Data Plane - Ingress PE

When an Ethernet frame is received by an ingress PE (e.g., PE1 in figure 4 above), the PE uses the AC ID (e.g., VLAN ID) to identify the associated MAC-VRF/BT and it performs a lookup on the destination MAC address. If the MAC address corresponds to its IRB Interface MAC address, the ingress PE deduces that the packet must be inter-subnet routed. Hence, the ingress PE performs an IP lookup in the associated IP-VRF table. The lookup identifies BGP next hop of egress PE along with the tunnel/encapsulation type and the associated MPLS/VNI values. The ingress PE also decrements the TTL/hop limit for that packet by one and if it reaches zero, the ingress PE discards the packet.

If the tunnel type is that of MPLS or IP-only NVO tunnel, then TS's IP packet is sent over the tunnel without any Ethernet header. However, if the tunnel type is that of Ethernet NVO tunnel, then an Ethernet header needs to be added to the TS's IP packet. The source MAC address of this inner Ethernet header is set to the ingress PE's router MAC address and the destination MAC address of this inner Ethernet header is set to the egress PE's router MAC address learnt via Router's MAC extended community attached to the route. MPLS VPN label is set to the received label2 in the route. In the case of Ethernet NVO tunnel type, VNI may be set one of two ways:

- downstream mode: VNI is set to the received label2 in the route which is downstream assigned.

- global mode: VNI is set to the received label2 in the route which is domain-wide assigned. This VNI value from received label2 MUST be the same as the locally configured VNI for the IP VRF as all PEs in the NVO MUST be configured with the same IP VRF VNI for this mode of operation.

PEs may be configured to operate in one of these two modes depending on the administrative domain boundaries across PEs participating in the NVO, and PE's capability to support downstream VNI mode.

In the case of NVO tunnel encapsulation, the outer source and destination IP addresses are set to the ingress and egress PE BGP next-hop IP addresses respectively.

5.5. Data Plane - Egress PE

When the tenant's MPLS or NVO encapsulated packet is received over an MPLS or NVO tunnel by the egress PE, the egress PE removes NVO tunnel encapsulation and uses the VPN MPLS label (for MPLS encapsulation) or VNI (for NVO encapsulation) to identify the IP-VRF in which IP lookup

needs to be performed. If the VPN MPLS label or VNI identifies a MAC- VRF instead of an IP-VRF, then the procedures in section 6.4 for asymmetric IRB are executed.

The lookup in the IP-VRF identifies a local adjacency to the IRB interface associated with the egress subnet's MAC-VRF/BT. The egress PE also decrements the TTL/hop limit for that packet by one and if it reaches zero, the egress PE discards the packet.

The egress PE gets the destination TS's MAC address for that TS's IP address from its ARP table or NDP cache, it encapsulates the packet with that destination MAC address and a source MAC address corresponding to that IRB interface and sends the packet to its destination subnet MAC-VRF/BT.

The destination MAC address lookup in the MAC-VRF/BT results in local adjacency (e.g., local interface) over which the Ethernet frame is sent on.

6. Asymmetric IRB Procedures

6.1. Control Plane - Advertising PE

When a PE (e.g., PE1 in figure 4 above) learns MAC and IP address of an attached TS (e.g., via an ARP request or ND Neighbor Solicitation), it populates its MAC-VRF/BT, IP-VRF, and ARP table or NDP cache just as in the case for symmetric IRB. It then builds an EVPN MAC/IP Advertisement route (type 2) as follows and advertises it to other PEs participating in that tenant's VPN.

- The Length field of the BGP EVPN NLRI for an EVPN MAC/IP Advertisement route MUST be either 37 (if IPv4 address is carried) or 49 (if IPv6 address is carried).

- Route Distinguisher (RD), Ethernet Segment Identifier, Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, IP Address, and MPLS Label1 fields MUST be set per [RFC7432] and [RFC8365].

- The MPLS Label2 field MUST NOT be included in this route.

Just as in [RFC7432], the RD, Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, and IP Address fields are part of the route key used by BGP to compare routes. The rest of the fields are not part of the route key.

This route is advertised along with the following extended community:

- Tunnel Type Extended Community

For asymmetric IRB mode, Router's MAC EC is not needed because forwarding is performed using destination TS's MAC address which is carried in this EVPN route type-2 advertisement.

This route MUST always be advertised with the MAC-VRF route target. It MAY also be advertised with a second route target corresponding to the IP-VRF.

6.2. Control Plane - Receiving PE

When a PE (e.g., PE2 in figure 4 above) receives this EVPN MAC/IP Advertisement route, it performs the following:

- Using MAC-VRF route target, it identifies the corresponding MAC-VRF and imports the MAC address into it. For asymmetric IRB mode, it is assumed that all PEs participating in a tenant's VPN are configured with all subnets (i.e., all VLANs) and corresponding MAC-VRFs/BTs even if there are no locally attached TSes for some of these subnets. The reason for this is because ingress PE needs to do forwarding based on destination TS's MAC address and perform NVO tunnel encapsulation as a property of a lookup in MAC-VRF/BT.

- If only MAC-VRF route target is used, then the receiving PE uses the MAC-VRF route target to identify the corresponding IP-VRF -- i.e., many MAC-VRF route targets map to the same IP-VRF for a given tenant. In this case, MAC-VRF may be used by the receiving PE to identify the corresponding IP VRF via the IRB interface associated with the subnet MAC-VRF/BT. This would equivalent to how ARP table or NDP cache entries are typically mapped to IRB interface of an IP VRF for installing attached host routes in an IP VRF. Since in asymmetric IRB mode, each PE is configured with all VLANs of a tenant, indirect import to IP VRF via the corresponding MAC-VRF route target is a viable alternative.

- Using MAC-VRF route target, the receiving PE identifies the corresponding ARP table or NDP cache for the tenant and it adds an entry to the ARP table or NDP cache for the TS's MAC and IP address association. It should be noted that the tenant's ARP table or NDP cache at the receiving PE is identified by all the MAC- VRF route targets for that tenant.

- If IP-VRF route target is included, it may be used to import the route to IP-VRF. If IP-VRF route-target is not included, MAC-VRF is used to derive corresponding IP-VRF for import, as explained in the prior section. In both cases, IP-VRF route is installed with the TS MAC binding included in the received route.

If the receiving PE receives the MAC/IP Advertisement route with MPLS label2 field but the receiving PE only supports asymmetric IRB mode, then the receiving PE MUST ignore MPLS label2 field and install the MAC address in the corresponding MAC-VRF and (IP, MAC) association in the ARP table or NDP cache for that tenant (with IRB interface identified by the MAC-VRF).

If the receiving PE receives the MAC/IP Advertisement route with MPLS label2 field and it uses symmetric IRB mode, then it should use the MAC-VRF route target to identify its corresponding MAC-VRF table and import the MAC address. It should use the IP-VRF route target to identify the corresponding IP-VRF table and import the IP address, as specified in symmetric IRB handling. It MUST NOT import (IP, MAC) association into its ARP table or NDP cache.

6.3. Data Plane - Ingress PE

When an Ethernet frame is received by an ingress PE (e.g., PE1 in figure 4 above), the PE uses the AC ID (e.g., VLAN ID) to identify the associated MAC-VRF/BT and it performs a lookup on the destination MAC address. If the MAC address corresponds to its IRB Interface MAC address, the ingress PE deduces that the packet must be inter-subnet routed. Hence, the ingress PE performs an IP lookup in the associated IP-VRF table. The lookup identifies a local adjacency to the IRB interface associated with the egress subnet's MAC-VRF/BT. The ingress PE also decrements the TTL/hop limit for that packet by one and if it reaches zero, the ingress PE discards the packet.

The ingress PE gets the destination TS's MAC address for that TS's IP address from its ARP table or NDP cache, it encapsulates the packet with that destination MAC address and a source MAC address corresponding to that IRB interface and sends the packet to its destination subnet MAC-VRF/BT.

The destination MAC address lookup in the MAC-VRF/BT results in BGP next hop address of egress PE along with label1 (L2 VPN MPLS label or VNI). The ingress PE encapsulates the packet using Ethernet NVO tunnel of the choice (e.g., VxLAN or NVGRE) and sends the packet to the egress PE. Because the packet forwarding is between ingress PE's MAC-VRF/BT and egress PE's MAC-VRF/BT, the packet encapsulation procedures follow that of [RFC7432] for MPLS and [RFC8365] for VxLAN encapsulations.

6.4. Data Plane - Egress PE

When a tenant's Ethernet frame is received over an NVO tunnel by the egress PE, the egress PE removes NVO tunnel encapsulation and uses the VPN MPLS label (for MPLS encapsulation) or VNI (for NVO

encapsulation) to identify the MAC-VRF/BT in which MAC lookup needs to be performed.

The MAC lookup results in local adjacency (e.g., local interface) over which the packet needs to get sent.

Note that the forwarding behavior on the egress PE is the same as EVPN intra-subnet forwarding described in [RFC7432] for MPLS and [RFC8365] for NVO networks. In other words, all the packet processing associated with the inter-subnet forwarding semantics is confined to the ingress PE for asymmetric IRB mode.

It should also be noted that [RFC7432] provides a different level of granularity for the EVPN label. Besides identifying the 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 individual MAC address identification, then no MAC lookup is needed in the egress PE for MPLS encapsulation and the packet can be directly forwarded to the egress interface just based on EVPN label lookup.

7. Mobility Procedure

When a TS moves from one NVE (aka source NVE) to another NVE (aka target NVE), 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. [RFC7432] 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 IRB. In order to place the emphasis on the differences between L2-only and IRB use cases, the incremental procedure is described for single-homed TS with the expectation that the additional steps needed for multi-homed TS, can be extended per section 15 of [RFC7432]. This section describes mobility procedures for both symmetric and asymmetric IRB. Although the language used in this section is for IPv4 ARP, it equally applies to IPv6 ND.

When a TS moves from a source NVE to a target NVE, it can behave in one of the following three ways:

1. TS initiates an ARP request upon a move to the target NVE

2. TS sends data packet without first initiating an ARP request to the target NVE

3. TS is a silent host and neither initiates an ARP request nor sends any packets

Depending on the expexted TS's behavior, an NVE needs to handle at least the first bullet and should be able to handle the 2nd and the 3rd bullet. The following subsections describe the procedures for each of them where it is assumed that the MAC and IP addresses of a TS have one-to-one relationship (i.e., there is one IP address per MAC address and vice versa). If there is many- to-one relationship such that there are many host IP addresses corresponding to a single host MAC address or there are many host MAC addresses corresponding to a single IP address, then to detect host mobility, the procedures in [I-D.ietf-bess-evpn-irb-extended-mobility] must be exercised followed by the procedures described below.

7.1. Initiating a gratutious ARP upon a Move

In this scenario when a TS moves from a source NVE to a target NVE, the TS initiates a gratuitous ARP upon the move to the target NVE.

The target NVE upon receiving this ARP message, updates its MAC-VRF, IP-VRF, and ARP table with the host MAC, IP, and local adjacency information (e.g., local interface).

Since this NVE has previously learned the same MAC and IP addresses from the source NVE, it recognizes that there has been a MAC move and it initiates MAC mobility procedures per [RFC7432] by advertising an EVPN MAC/IP Advertisement route with both the MAC and IP addresses filled in (per sections 5.1 and 6.1) along with MAC Mobility Extended Community with the sequence number incremented by one. The target NVE also exercises the MAC duplication detection procedure in section 15.1 of [RFC7432].

The source NVE upon receiving this MAC/IP Advertisement route, realizes that the MAC has moved to the target NVE. It updates its MAC-VRF and IP-VRF table accordingly with the adjacency information of the target NVE. In the case of the asymmetric IRB, the source NVE also updates its ARP table with the received adjacency information and in the case of the symmetric IRB, the source NVE removes the entry associated with the received (MAC, IP) from its local ARP table. It then withdraws its EVPN MAC/IP Advertisement route. Furthermore, it sends an ARP probe locally to ensure that the MAC is gone. If an ARP response is received, the source NVE updates its ARP entry for that (IP, MAC) and re-advertises an EVPN MAC/IP Advertisement route for that (IP, MAC) along with MAC Mobility Extended Community with the sequence number incremented by one. The source NVE also exercises the MAC duplication detection procedure in section 15.1 of [RFC7432].

All other remote NVE devices upon receiving the MAC/IP Advertisement route 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 the TS in their corresponding MAC-VRF and IP-VRF tables to point to the target NVE. Furthermore, upon receiving the MAC/IP withdraw for the TS from the source NVE, these remote PEs perform the cleanups for their BGP tables.

7.2. Sending Data Traffic without an ARP Request

In this scenario when a TS moves from a source NVE to a target NVE, the TS starts sending data traffic without first initiating an ARP request.

The target NVE upon receiving the first data packet, learns the MAC address of the TS in the data plane and updates its MAC-VRF table with the MAC address and the local adjacency information (e.g., local interface) accordingly. The target NVE realizes that there has been a MAC move because the same MAC address has been learned remotely from the source NVE.

If EVPN-IRB NVEs are configured to advertise MAC-only routes in addition to MAC-and-IP EVPN routes, then the following steps are taken:

- The target NVE upon learning this MAC address in the data plane, updates this MAC address entry in the corresponding MAC-VRF with the local adjacency information (e.g., local interface). It also recognizes that this MAC has moved and initiates MAC mobility procedures per [RFC7432] by advertising an EVPN MAC/IP Advertisement route with only the MAC address filled in along with MAC Mobility Extended Community with the sequence number incremented by one.

- The source NVE upon receiving this MAC/IP Advertisement route, realizes that the MAC has moved to the new NVE. It updates its MAC-VRF table with the adjacency information for that MAC address to point to the target NVE and withdraws its EVPN MAC/IP Advertisement route that has only the MAC address (if it has advertised such route previously). Furthermore, it searches for the corresponding MAC-IP entry and sends an ARP probe for this (MAC,IP) pair. The ARP request message is sent both locally to all attached TSes in that subnet as well as it is sent to other NVEs participating in that subnet including the target NVE. Note that the PE needs to maintain a correlation between MAC and MAC-IP route entries in the MAC-VRF to accomplish this.

- The target NVE passes the ARP request to its locally attached TSes and when it receives the ARP response, it updates its IP-VRF and

ARP table with the host (MAC, IP) information. It also sends an EVPN MAC/IP Advertisement route with both the MAC and IP addresses filled in along with MAC Mobility Extended Community with the sequence number set to the same value as the one for MAC-only advertisement route it sent previously.

- When the source NVE receives the EVPN MAC/IP Advertisement route, it updates its IP-VRF table with the new adjacency information (pointing to the target NVE). In the case of the asymmetric IRB, the source NVE also updates its ARP table with the received adjacency information and in the case of the symmetric IRB, the source NVE removes the entry associated with the received (MAC, IP) from its local ARP table. Furthermore, it withdraws its previously advertised EVPN MAC/IP route with both the MAC and IP address fields filled in.

- All other remote NVE devices upon receiving the MAC/IP advertisement route 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 the TS in their corresponding MAC-VRF, IP-VRF, and ARP tables (in the case of asymmetric IRB) to point to the new NVE. Furthermore, upon receiving the MAC/IP withdraw for the TS from the old NVE, these remote PEs perform the cleanups for their BGP tables.

If EVPN-IRB NVEs are configured not to advertise MAC-only routes, then upon receiving the first data packet, it learns the MAC address of the TS and updates the MAC entry in the corresponding MAC-VRF table with the local adjacency information (e.g., local interface). It also realizes that there has been a MAC move because the same MAC address has been learned remotely from the source NVE. It uses the local MAC route to find the corresponding local MAC-IP route, and sends a unicast ARP request to the host and when receiving an ARP response, it follows the procedure outlined in section 7.1. In the prior case, where MAC-only routes are also advertised, this procedure of triggering a unicast ARP probe at the target PE MAY also be used in addition to the source PE broadcast ARP probing procedure described earlier for better convergence.

7.3. Silent Host

In this scenario when a TS moves from a source NVE to a target NVE, the TS is silent and it neither initiates an ARP request nor it sends any data traffic. Therefore, neither the target nor the source NVEs are aware of the MAC move.

On the source NVE, an age-out timer (for the silent host that has moved) is used to trigger an ARP probe. This age-out timer can be either ARP timer or MAC age-out timer and this is an implementation choice. The ARP request gets sent both locally to all the attached TSes on that subnet as well as it gets sent to all the remote NVEs (including the target NVE) participating in that subnet. The source NVE also withdraw the EVPN MAC/IP Advertisement route with only the MAC address (if it has previously advertised such a route).

The target NVE passes the ARP request to its locally attached TSes and when it receives the ARP response, it updates its MAC-VRF, IP-VRF, and ARP table with the host (MAC, IP) and local adjacency information (e.g., local interface). It also sends an EVPN MAC/IP advertisement route with both the MAC and IP address fields filled in along with MAC Mobility Extended Community with the sequence number incremented by one.

When the source NVE receives the EVPN MAC/IP Advertisement route, it updates its IP-VRF table with the new adjacency information (pointing to the target NVE). In the case of the asymmetric IRB, the source NVE also updates its ARP table with the received adjacency information and in the case of the symmetric IRB, the source NVE removes the entry associated with the received (MAC, IP) from its local ARP table. Furthermore, it withdraws its previously advertised EVPN MAC/IP route with both the MAC and IP address fields filled in.

All other remote NVE devices upon receiving the MAC/IP Advertisement route 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 the TS in their corresponding MAC-VRF, IP-VRF, and ARP (in the case of asymmetric IRB) tables to point to the new NVE. Furthermore, upon receiving the MAC/IP withdraw for the TS from the old NVE, these remote PEs perform the cleanups for their BGP tables.

8. BGP Encoding

This document defines one new BGP Extended Community for EVPN.

8.1. 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 Encapsulation Extended Community defined in section 4.1 of [I-D.ietf-idr-tunnel-encaps].

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

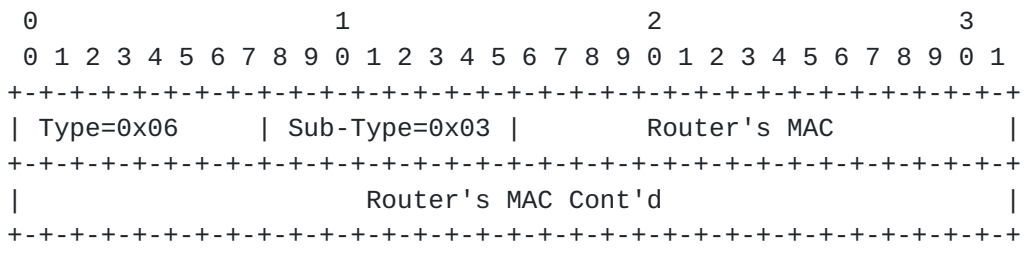


Figure 5: Router's MAC Extended Community

This extended community is used to carry the PE's MAC address for symmetric IRB scenarios and it is sent with EVPN RT-2. The advertising PE SHALL only attach a single Router's MAC Extended Community to a route. In case the receiving PE receives more than one Router's MAC Extended Community with a route, it SHALL process the first one in the list and not store and propagate the others.

9. Operational Models for Symmetric Inter-Subnet Forwarding

The following sections describe two main symmetric IRB forwarding scenarios (within a DC -- i.e., intra-DC) along with the corresponding procedures. In the following scenarios, without loss of generality, it is assumed that a given tenant is represented by a single IP-VPN instance. Therefore, on a given PE, a tenant is represented by a single IP-VRF table and one or more MAC-VRF tables.

9.1. 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, without loss of generality, it is assumed that NVEs operate in VLAN-based service interface mode with one Bridge Table (BT) per MAC-VRF. Thus, for a given tenant, an NVE has one MAC-VRF for each tenant subnet (e.g., each VLAN) that is configured for extension via VxLAN or NVGRE encapsulation. In the case of VLAN-aware bundling, then each MAC-VRF consists of multiple Bridge Tables (e.g., one BT 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.

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 8.1).

Figure 6 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 TSes that are associated with these three MAC-VRFs -- i.e., TS1, TS4, and TS5 are on the same subnet (e.g., same MAC-VRF/VLAN). TS1 and TS5 are associated with MAC-VRF1 on NVE1, while 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 the L2 forwarding (bridging) part of the IRB solution is exercised because all these TSes belong to the same subnet. However, when TS1 wants to exchange traffic with TS2 or TS3 which belong to different subnets, 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.

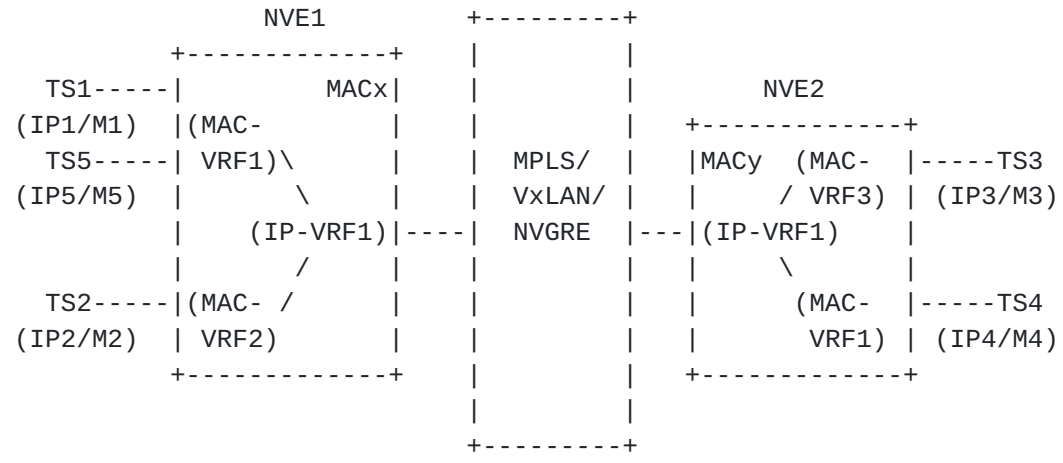


Figure 6: IRB forwarding on NVEs for Tenant Systems

9.1.1. Control Plane Operation

Each NVE advertises a MAC/IP Advertisement route (i.e., Route Type 2) for each of its TSes with the following field set:

- RD and ESI per [RFC7432]
- 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
- Label1 = MPLS Label or VNI corresponding to MAC-VRF
- Label2 = MPLS Label or VNI corresponding to IP-VRF

Each NVE advertises an EVPN RT-2 route with two Route Targets (one corresponding to its MAC-VRF and the other corresponding to its IP-VRF. Furthermore, the EVPN RT-2 is advertised with two BGP Extended Communities. The first BGP Extended Community identifies the tunnel type and it is called Encapsulation Extended Community as defined in [I-D.ietf-idr-tunnel-encaps] 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 8.1. The Router's MAC Extended community MUST be added when Ethernet NVO tunnel is used. If IP NVO tunnel type is used, then there is no need to send this second Extended Community. It should be noted that IP NVO tunnel type is only applicable to symmetric IRB procedures.

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 the MAC and IP addresses into them respectively.
- It imports the MAC address from MAC/IP Advertisement route into the MAC-VRF with BGP Next Hop address as the underlay tunnel destination address (e.g., VTEP DA for VxLAN encapsulation) and Label1 as VNI 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 uses 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 the underlay tunnel destination address,

VTEP DA for VxLAN encapsulation and Label2 as IP-VPN VNI for VxLAN encapsulation.

- If the receiving NVE uses MPLS encapsulation, then the receiving NVE imports the IP address into IP-VRF with BGP Next Hop address as the underlay tunnel destination address, and Label2 as IP-VPN label for MPLS encapsulation.

If the receiving NVE receives an EVPN RT-2 with only Label1 and only a single Route Target corresponding to IP-VRF, or if it receives an EVPN RT-2 with only a single Route Target corresponding to MAC-VRF but with both Label1 and Label2, or if it receives an EVPN RT-2 with MAC Address Length of zero, then it MUST use the treat-as-withdraw approach [RFC7606] and SHOULD log an error message.

9.1.2. 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 TS1 in subnet-1 (MAC-VRF1) on NVE1 wants to send traffic to TS3 in subnet-3 (MAC-VRF3) on NVE2.

- NVE1 receives 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 an IP lookup is performed on the destination IP address. NVE1 also decrements the TTL/hop limit for that packet by one and if it reaches zero, NVE1 discards the packet. This lookup yields the outgoing NVO tunnel and the required encapsulation. If the encapsulation is for Ethernet NVO tunnel, then it includes the egress NVE's MAC address as inner MAC DA, the egress NVE's IP address (e.g., BGP Next Hop address) as the VTEP DA, and the VPN-ID as the VNI. The inner MAC SA and VTEP SA are set to NVE's MAC and IP addresses respectively. If it is a MPLS encapsulation, then corresponding EVPN and LSP labels are added to the packet. The packet is then forwarded to the egress NVE.

- On the egress NVE, if the packet arrives on Ethernet NVO tunnel (e.g., it is VxLAN encapsulated), then the NVO tunnel 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 the VNI to identify the IP-VRF table. If the packet is MPLS encapsulated, then the EVPN label lookup identifies the IP-VRF table. Next, an IP lookup is performed for the destination TS (TS3) which results in an 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. NVE2 also decrements the TTL/hop limit for that packet by one and if it reaches zero, NVE2 discards the packet.

- The IP packet is encapsulated with an Ethernet header with MAC SA set to that of IRB interface MAC address (i.e, IRB interface between MAC-VRF3 and IP-VRF1 on NVE2) and MAC DA set to that of destination TS (TS3) MAC address. The packet is sent to the corresponding MAC-VRF (i.e., 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 VNI/MPLS label. For instance, traffic from TS2 to TS4 will be encapsulated by NVE1 using NVE2's IP-VRF VNI/MPLS label, as long as TS4's host IP is present in NVE1's IP-VRF.

9.2. IRB forwarding on NVEs for Subnets behind Tenant Systems

This section covers the symmetric IRB procedures for the scenario where some Tenant Systems (TSes) support one or more subnets and these TSes are associated with one or more NVEs. Therefore, besides the advertisement of MAC/IP addresses for each TS which can be multi-homed with All-Active redundancy mode, the associated NVE needs to also advertise the subnets statically configured on each TS.

The main difference between this solution and the previous one is the additional advertisement corresponding to each subnet. These subnet advertisements are accomplished using the EVPN IP Prefix route defined in [I-D.ietf-bess-evpn-prefix-advertisement]. 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 advertising 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 7 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 TSes associated with these three MAC-VRFs -- i.e., TS1 is 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.

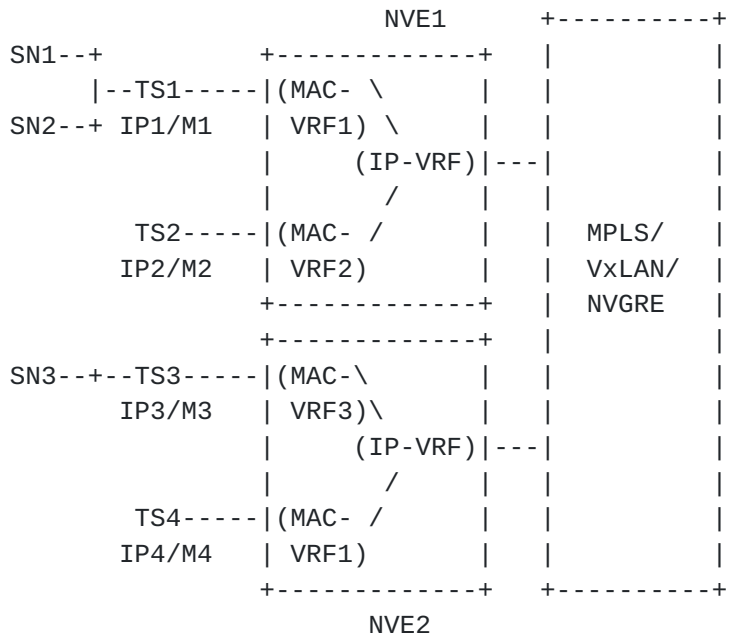


Figure 7: IRB forwarding on NVEs for subnets behind TSes

9.2.1. Control Plane Operation

Each NVE advertises a Route Type-5 (EVPN RT-5, IP Prefix Route defined in [I-D.ietf-bess-evpn-prefix-advertisement]) for each of its

subnet prefixes with the IP address of its TS as the next hop (gateway address field) as follows:

- RD associated with the IP-VRF
- ESI = 0
- Ethernet Tag = 0;
- IP Prefix Length = 0 to 32 or 0 to 128
- IP Prefix = SNi
- Gateway Address = IPi; IP address of TS
- MPLS Label = 0

This EVPN RT-5 is advertised with one or more Route Targets associated with the IP-VRF from which the route is originated.

Each NVE also advertises an EVPN RT-2 (MAC/IP Advertisement Route) along with their associated Route Targets and Extended Communities for each of its TSes exactly as described in section 9.1.1.

Upon receiving the EVPN 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 that is configured with an import RT that is one of the RTs being carried by the EVPN RT-5 route along with the IP address of the associated TS as its next hop.

When receiving the EVPN RT-2 advertisement, the receiving NVE imports MAC/IP addresses of the TS into the corresponding MAC-VRF and IP-VRF per section 9.1.1. When both routes exist, recursive route resolution is performed to resolve the IP prefix (received in EVPN RT-5) to its corresponding NVE's IP address (e.g., its BGP next hop). BGP next hop will be used as the underlay tunnel destination address (e.g., VTEP DA for VxLAN encapsulation) and Router's MAC will be used as inner MAC for VxLAN encapsulation.

9.2.2. 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 9.1.1.

- 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 VNI. MAC SA is set to NVE1's MAC address and VTEP SA is set to NVE1's IP address. NVE1 also decrements the TTL/hop limit for that packet by one and if it reaches zero, NVE1 discards the packet.

- 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 VNI.

- 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. NVE2 also decrements the TTL/hop limit for that packet by one and if it reaches zero, NVE2 discards the packet.

- 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.

10. Acknowledgements

The authors would like to thank Sami Boutros, Jeffrey Zhang, Krzysztof Szarkowicz, Lukas Krattiger and Neeraj Malhotra for their valuable comments. The authors would also like to thank Linda

Dunbar, Florin Balus, Yakov Rekhter, Wim Henderickx, Lucy Yong, and Dennis Cai for their feedback and contributions.

11. Security Considerations

The security considerations for layer-2 forwarding in this document follow that of [RFC7432] for MPLS encapsulation and it follows that of [RFC8365] for VxLAN or NVGRE encapsulations. This section describes additional considerations.

This document describes a set of procedures for Inter-Subnet Forwarding of tenant traffic across PEs (or NVEs). These procedures include both layer-2 forwarding and layer-3 routing on a packet by packet basis. The security consideration for layer-3 routing in this document follows that of [RFC4365] with the exception for the application of routing protocols between CEs and PEs. Contrary to [RFC4364], this document does not describe route distribution techniques between CEs and PEs, but rather considers the CEs as TSes or VAs that do not run dynamic routing protocols. This can be considered a security advantage, since dynamic routing protocols can be blocked on the NVE/PE ACs, not allowing the tenant to interact with the infrastructure's dynamic routing protocols.

The VPN scheme described in this document does not provide the quartet of security properties mentioned in [RFC4365] (confidentiality protection, source authentication, integrity protection, replay protection). If these are desired, they must be provided by mechanisms that are outside the scope of the VPN mechanisms.

In this document, the EVPN RT-5 is used for certain scenarios. This route uses an Overlay Index that requires a recursive resolution to a different EVPN route (an EVPN RT-2). Because of this, it is worth noting that any action that ends up filtering or modifying the EVPN RT-2 route used to convey the Overlay Indexes, will modify the resolution of the EVPN RT-5 and therefore the forwarding of packets to the remote subnet.

12. 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.

13. References

13.1. Normative References

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-22 (work in progress), January 2021.

[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>.

[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>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[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>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation", RFC 7637, DOI 10.17487/RFC7637, September 2015, <https://www.rfc-editor.org/info/rfc7637>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

13.2. Informative References

[I-D.ietf-bess-evpn-irb-extended-mobility]
Malhotra, N., Sajassi, A., Pattekar, A., Lingala, A., Rabadan, J., and J. Drake, "Extended Mobility Procedures for EVPN-IRB", draft-ietf-bess-evpn-irb-extended-mobility-03 (work in progress), May 2020.

[I-D.ietf-nvo3-vxlan-gpe]
Maino, F., Kreeger, L., and U. Elzur, "Generic Protocol Extension for VXLAN (VXLAN-GPE)", draft-ietf-nvo3-vxlan-gpe-10 (work in progress), July 2020.

[RFC4365] Rosen, E., "Applicability Statement for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4365, DOI 10.17487/RFC4365, February 2006, <https://www.rfc-editor.org/info/rfc4365>.

[RFC5798] Nadas, S., Ed., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, DOI 10.17487/RFC5798, March 2010, <https://www.rfc-editor.org/info/rfc5798>.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <https://www.rfc-editor.org/info/rfc7365>.

Authors' Addresses

Ali Sajassi
Cisco Systems
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sajassi@cisco.com

Samer Salam
Cisco Systems

Email: ssalam@cisco.com

Samir Thoria
Cisco Systems

Email: sthoria@cisco.com

John E Drake
Juniper

Email: jdrake@juniper.net

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com