Internet Engineering Task Force (IETF)
Request for Comments: 9135
Category: Standards Track
ISSN: 2070-1721

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

October 2021

# Integrated Routing and Bridging in Ethernet VPN (EVPN)

## Abstract

Ethernet VPN (EVPN) provides an extensible and flexible multihoming 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 multihoming capabilities of EVPN. This document describes an Integrated Routing and Bridging (IRB) solution based on EVPN to address such requirements.

## Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9135.



## Table of Contents

## 1. Introduction

EVPN [RFC7432] provides an extensible and flexible multihoming VPN solution over an MPLS/IP network for intra-subnet connectivity among Tenant Systems (TSs) 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) association 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 multihoming capabilities of EVPN. This document describes an Integrated Routing and Bridging (IRB) solution based on EVPN to address such requirements.

Inter-subnet communication is typically performed by centralized Layer 3 (L3) gateway (GW) devices, which enforce all inter-subnet communication policies and perform all inter-subnet forwarding. When two TSs belonging to two different subnets connected to the same Provider Edge (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 sent back to the PE. For today's large multi-tenant Data Center (DC), this scheme is very inefficient and sometimes impractical.

In order to overcome the drawback of the centralized L3 GW approach, IRB functionality is needed on the PEs (also referred to as EVPN Network Virtualization Edges (NVEs)) attached to TSs in order to avoid inefficient forwarding of tenant traffic (i.e., avoid backhauling and hair pinning). When a PE with IRB capability receives tenant traffic over an Attachment Circuit (AC), it cannot only locally bridge the tenant intra-subnet traffic but also 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 a centralized L3 GW approach.

Some TSs 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 provide each tenant with IP routing of its inter-subnet traffic and Ethernet bridging of its intra-subnet traffic and non-routable traffic, where non-routable traffic refers to both non-IP traffic and IP traffic whose version differs from the IP version configured in IP Virtual Routing and Forwarding (IP-VRF). For example, if an IP-VRF in an NVE is configured for IPv6 and that NVE receives IPv4 traffic on the corresponding VLAN, then the IPv4 traffic is treated as non-routable traffic.

R2: The solution must allow IP routing of inter-subnet traffic to be disabled on a per-VLAN basis on those PEs that are backhauling that traffic to another PE for routing.

## 2. Terminology

AC: Attachment Circuit

ARP: Address Resolution Protocol

ARP Table: A logical view of a forwarding table on a PE that maintains an IP to a MAC binding entry on an IP interface for both IPv4 and IPv6. These entries are learned through ARP/ND or through EVPN.

BD: Broadcast Domain. As per [RFC7432], an EVI consists of a single BD 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 a 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 a 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].

CE: Customer Edge

DA: Destination Address

Ethernet NVO Tunnel: Refers to Network Virtualization Overlay tunnels with an Ethernet payload, as specified for VXLAN in [RFC7348] and for NVGRE in [RFC7637].

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

EVPN: Ethernet VPN, as per [RFC7432].

IP NVO Tunnel: Refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload) as specified for Generic Protocol Extension (GPE) in [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: Media Access Control

MAC-VRF: A Virtual Routing and Forwarding table for 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

NVE: Network Virtualization Edge

NVGRE: Network Virtualization Using Generic Routing Encapsulation, as per [RFC7637].

NVO: Network Virtualization Overlay

PE: Provider Edge

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 [RFC9136].

SA: Source Address

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 a Virtual Subnet Identifier in NVGRE, etc., unless it is stated otherwise.

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

VXLAN: Virtual eXtensible Local Area Network, as per [RFC7348].

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

### 2.1. Requirements Language

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

## 3. EVPN PE Model for IRB Operation

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

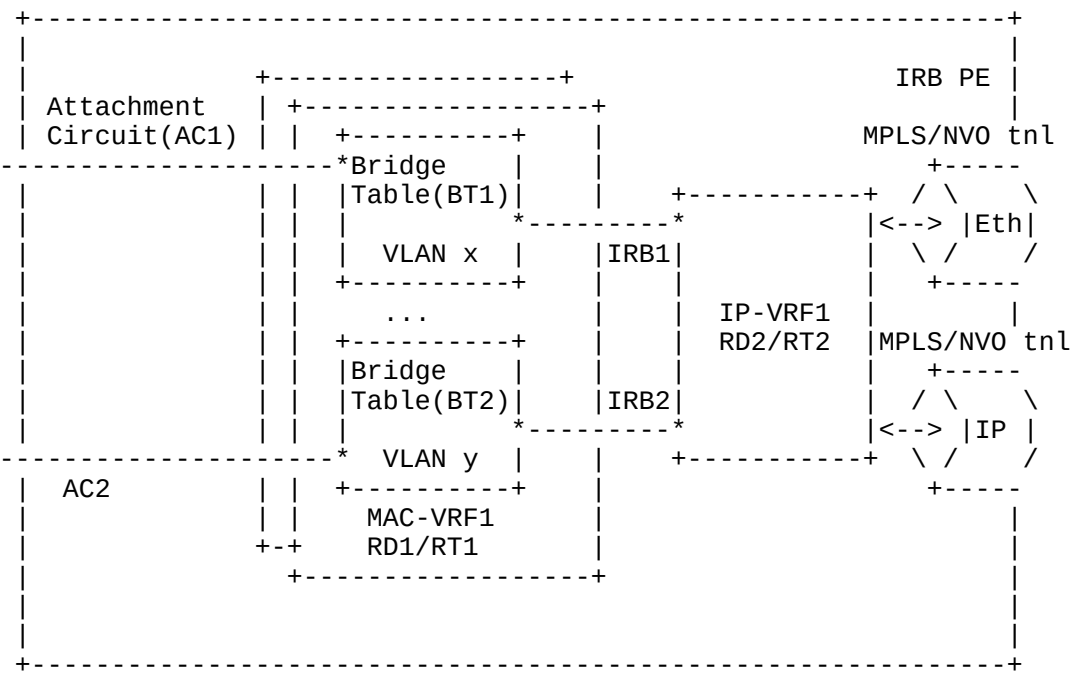



Figure 1: EVPN IRB PE Model

A tenant needing IRB services on a PE requires an IP-VRF table along with one or more MAC-VRF tables. An IP-VRF, as defined in [RFC4364], is the instantiation of an IP-VPN instance in a PE. A MAC-VRF, as defined in [RFC7432], is the instantiation of an EVI in a PE. A MAC-VRF consists of one or more bridge tables, where each bridge table corresponds to a VLAN (broadcast domain). 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 bridge table 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 bridge tables per MAC-VRF (per EVI) -- i.e., there are several tenant VLANs per EVI.

Each bridge table is connected to an IP-VRF via an L3 interface called an "IRB interface". Since a single tenant subnet is typically (and in this document) represented by a VLAN (and thus supported by a single bridge table), for a given tenant, there are as many bridge tables as there are subnets. Thus, there are also as many IRB interfaces between the tenant IP-VRF and the associated bridge tables 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 a MAC-VRF Route Target can identify the corresponding bridge table; 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 bridge table.

## 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 trade-offs. Specifically,

* References to an 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 protocols. For IP-to-MAC bindings learned 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 a 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 the 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.

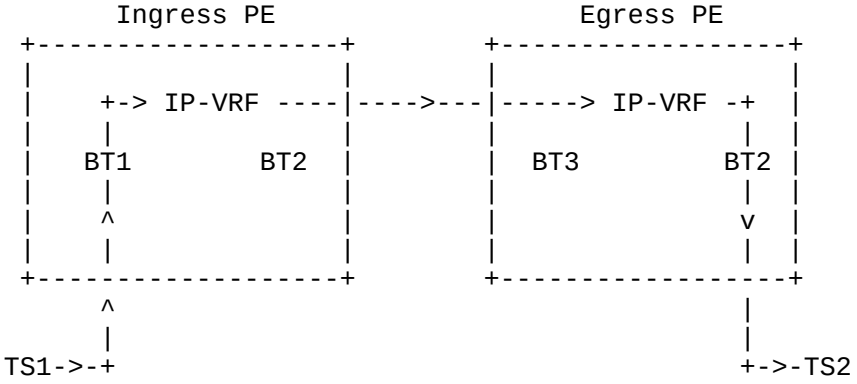
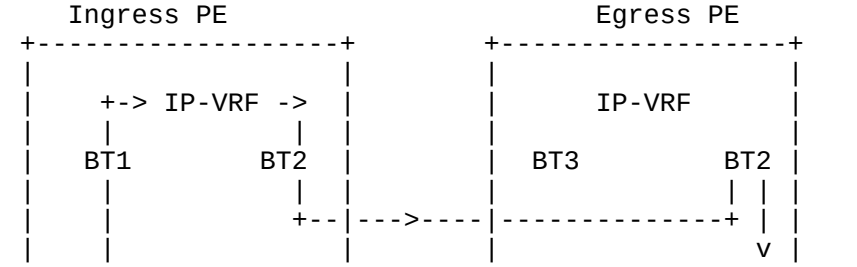



Figure 2: Symmetric IRB

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 an IP-only tunnel (e.g., in the case of MPLS or GPE encapsulation) or an Ethernet NVO tunnel (e.g., in the 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 PE MAC addresses -- i.e., there is no need for the 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 the association of the destination TS2's IP and MAC addresses in its ARP table. Each PE participating in symmetric IRB only maintains ARP entries for locally connected hosts and 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. The egress PE performs just a single MAC lookup as depicted in Figure 3 below.



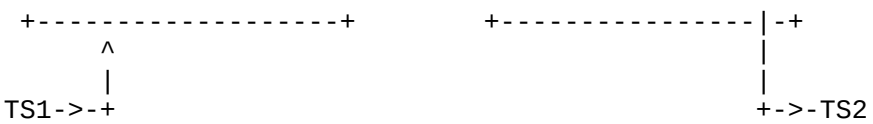



Figure 3: Asymmetric IRB

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 the 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 bridge table even though it may not have any TSs 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 the PE scale aspects for its ARP table size, its IRB interfaces, and the 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, the IPv4 Time To Live (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 sections define the control and data plane procedures for symmetric and asymmetric IRB on ingress and egress PEs. The following figure is used to describe these procedures, showing a single IP-VRF and a number of BDs on each PE for a given tenant. That is, an IP-VRF connects one or more EVIs, and each EVI contains one MAC-VRF; each MAC VRF consists of one or more bridge tables, one per BD; and a PE has an associated IRB interface for each BD.

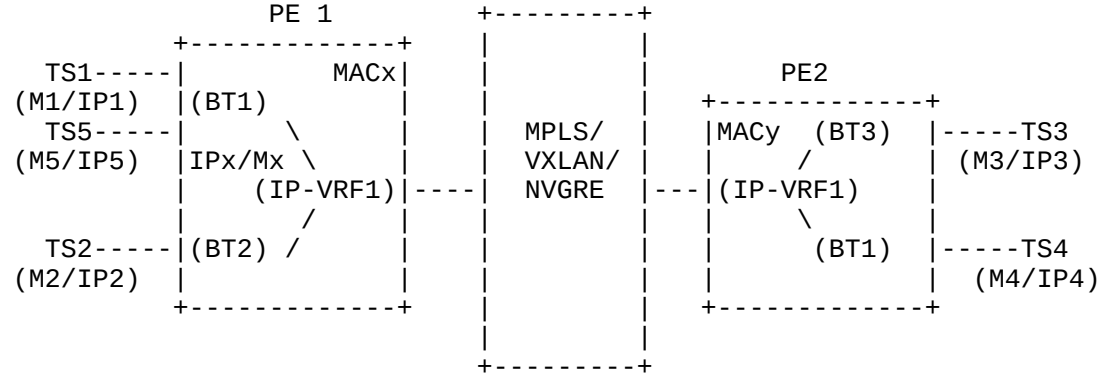



Figure 4: IRB Forwarding

## 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 fall 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, these default gateway IP and MAC addresses correspond to the IRB interface connecting the bridge table 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 bridge table 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 TSs are employing or relying on any form of MAC security, then the first option (i.e., using an anycast MAC address) should be used to ensure that the applications receive traffic from the same IRB interface MAC address to which they are sending. 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, 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}

* Anycast IPv6 IRB case: 00-00-5E-00-02-{VRID}

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 sending a 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 the EVPN route type 2 advertisement.

For option 1 -- irrespective of whether only the anycast MAC address or both anycast and non-anycast MAC addresses (where the latter one is used for the purpose of OAM) are used on the same IRB -- when a TS sends an ARP request or ND Neighbor Solicitation (NS) to the PE to

which it is attached, the request is sent for the anycast IP address of the IRB interface associated with the TS's subnet. The reply will use an anycast MAC address (in both the 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; 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 the source MAC address.

## 4.2. Operational Considerations

Symmetric and asymmetric IRB modes may coexist in the same network, and an ingress PE that supports both forwarding modes for a given tenant can interwork with egress PEs that support either IRB mode. The egress PE will indicate the desired forwarding mode for a given host based on the presence of the Label2 field and the IP-VRF Route Target in the EVPN MAC/IP Advertisement route. If the Label2 field of the received MAC/IP Advertisement route for host H1 is non-zero, and one of its Route Targets identifies the IP-VRF, the ingress PE will use symmetric IRB mode when forwarding packets destined to H1. If the Label2 field is zero and the MAC/IP Advertisement route for H1 does not carry any Route Target that identifies the IP-VRF, the ingress PE will use asymmetric mode when forwarding traffic to H1.

As an example that illustrates the previous statement, suppose PE1 and PE2 need to forward packets from TS2 to TS4 in Figure 4. Since both PEs are attached to the bridge table of the destination host, symmetric and asymmetric IRB modes are both possible as long as the ingress PE, PE1, supports both modes. The forwarding mode will depend on the mode configured in the egress PE, PE2. That is:

1. If PE2 is configured for symmetric IRB mode, PE2 will advertise TS4 MAC/IP addresses in a MAC/IP Advertisement route with a non-zero Label2 field, e.g., Label2 = Lx, and a Route Target that identifies IP-VRF1 in PE1. IP4 will be installed in PE1's IP-VRF1; TS4's ARP and MAC information will also be installed in PE1's IRB interface ARP table and BT1, respectively. When a packet from TS2 destined to TS4 is looked up in PE1's IP-VRF route table, a longest prefix match lookup will find IP4 in the IP-VRF, and PE1 will forward using the symmetric IRB mode and Label Lx.

2. However, if PE2 is configured for asymmetric IRB mode, PE2 will advertise TS4 MAC/IP information in a MAC/IP Advertisement route with a zero Label2 field and no Route Target identifying IP-VRF1. In this case, PE2 will install TS4 information in its ARP table and BT1. When a packet from TS2 to TS4 arrives at PE1, a longest prefix match on IP-VRF1's route table will yield the local IRB interface to BT1, where a subsequent ARP and bridge table lookup will provide the information for an asymmetric forwarding mode to PE2.

Refer to [EVPN] for more information about interoperability between symmetric and asymmetric forwarding modes.

The choice between symmetric or asymmetric mode is based on the operator's preference, and it is a trade-off between scale (which is better in the symmetric IRB mode) and control plane simplicity (asymmetric IRB mode simplifies the control plane). In cases where a tenant has hosts for every subnet attached to all (or most of) the PEs, the ARP and MAC entries need to be learned by all PEs anyway; therefore, the asymmetric IRB mode simplifies the forwarding model and saves space in the IP-VRF route table, since host routes are not installed in the route table. However, if the tenant does not need

to stretch subnets (broadcast domains) to multiple PEs and inter-subnet forwarding is needed, the symmetric IRB model will save ARP and bridge table space in all the PEs (in comparison with the asymmetric IRB model).

# 5. Symmetric IRB Procedures

## 5.1. Control Plane - Advertising PE

When a PE (e.g., PE1 in Figure 4 above) learns the 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 Neighbor Discovery Protocol (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 Network Layer Reachability Information (NLRI) for an EVPN MAC/IP Advertisement route MUST be either 40 (if the IPv4 address is carried) or 52 (if the IPv6 address is carried).

* The 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. EVPN Router's MAC Extended Community

This route is advertised with one or more Encapsulation Extended Communities [RFC9012], one for each encapsulation type supported by the advertising PE. If one or more encapsulation types require an Ethernet frame, a single EVPN Router's MAC Extended Community (Section 8.1) is also advertised. This extended community specifies the MAC address to be used as the inner destination MAC address in an Ethernet frame sent to the advertising PE.

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:

* The MAC-VRF Route Target and Ethernet Tag, if the latter is non-zero, are used to identify the correct MAC-VRF and bridge table, and if they are found, the MAC address is imported. The IP-VRF Route Target is used to identify the correct IP-VRF, and if it is

found, the IP address is imported.

If the MPLS Label2 field is non-zero, it means that this route is to be used for symmetric IRB, and the MPLS label2 value is to be used when sending a packet for this IP address to the advertising PE.

If the receiving PE supports asymmetric IRB mode and receives this route with both the MAC-VRF and IP-VRF Route Targets but the MAC/IP Advertisement route does not include the MPLS Label2 field, then the receiving PE installs the MAC address in the corresponding MAC-VRF and the (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 but has the corresponding MAC-VRF, then it only imports the MAC address.

If the receiving PE receives this route with both the MAC-VRF and IP-VRF Route Targets and the MAC/IP Advertisement route includes the MPLS Label2 field but the receiving PE only supports asymmetric IRB mode, then the receiving PE MUST ignore the 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 are 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 (IP, MAC) learning using gleaning procedures initiated by an inter-subnet data packet.

That is, if a given host’s (IP, MAC) association is unknown, and an ingress PE needs to send a packet to that host, then that ingress PE needs to know which egress PEs are attached to the subnet in which the host resides in order to send the packet to one of those PEs, causing the PE receiving the packet to probe for that host. For example, consider a subnet A that is locally attached to PE1 and subnet B that is locally attached to PE2 and PE3. Host A in subnet A, which is attached to PE1, initiates a data packet destined to host B in subnet B, which is attached to PE3. If host B’s (IP, MAC) has not yet been learned via either a gratuitous ARP OR a prior gleaning procedure, a new gleaning procedure MUST be triggered for host B’s (IP, MAC) to be learned 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 (IP, MAC). 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 (IP, MAC) 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 the BGP next hop of the 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 an MPLS or IP-only NVO tunnel, then the TS's IP packet is sent over the tunnel without any Ethernet header. However, if the tunnel type is that of an 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 learned via the EVPN Router's MAC Extended Community attached to the route. The MPLS VPN label is set to the received label2 in the route. In the case of the Ethernet NVO tunnel type, the VNI may be set one of two ways:

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

global mode: The VNI is set to the received label2 in the route, which is assigned domain-wide. This VNI value from the 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. If the received label2 value does not match the locally configured VNI value, the route MUST NOT be used, and an error message SHOULD be logged.

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 the 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 the 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 the local adjacency (e.g., local interface) over which the Ethernet frame is sent.

## 6. Asymmetric IRB Procedures

### 6.1. Control Plane - Advertising PE

When a PE (e.g., PE1 in Figure 4 above) learns the 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 an IPv4 address is carried) or 49 (if an IPv6 address is carried).

* The 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, the EVPN Router's MAC Extended Community 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 the 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 TSs for some of these subnets. This is because the ingress PE needs to do forwarding based on the destination TS's MAC address and perform NVO tunnel encapsulation as the property of a lookup in the MAC-VRF/BT.

* If only the 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. In this case, the MAC-VRF Route Target may be used by the receiving PE to identify the corresponding IP-VRF.

* Using the 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 the IP-VRF Route Target is included, it may be used to import the route to IP-VRF. If the IP-VRF Route Target is not included, MAC-VRF is used to derive the corresponding IP-VRF for import, as explained in the prior section. In both cases, an 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 the MPLS Label2 field but the receiving PE only supports asymmetric IRB mode, then the receiving PE MUST ignore the 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 the IRB interface identified by the MAC-VRF).

### 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/ bridge table. 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 a BGP next-hop address of the egress PE along with label1 (L2 VPN MPLS label or VNI). The ingress PE encapsulates the packet using the 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 the ingress PE's MAC-VRF/BT and the egress PE's MAC-VRF/ bridge table, 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 the 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 the MAC lookup needs to be performed.

The MAC lookup results in a 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 the 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 an EVPN label is used for an 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 the 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 be properly executed and the corresponding MAC-VRF and IP-VRF tables on all participating NVEs be updated. [RFC7432] describes the MAC Mobility procedures for L2-only services for both single-homed TS and multihomed 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 a single-homed TS with the expectation that the additional steps needed for a multihomed 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 a 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 expected TS’s behavior, an NVE needs to handle at least the first option and should be able to handle the second and third options. The following subsections describe the procedures for each scenario where it is assumed that the MAC and IP addresses of a TS have a one-to-one relationship (i.e., there is one IP address per MAC address and vice versa). The procedures for host mobility detection in the presence of a many-to-one relationship is outside the scope of this document, and it is covered in [EXTENDED-MOBILITY]. The "many-to-one relationship" refers to many host IP addresses corresponding to a single host MAC address or many host MAC addresses corresponding to a single IP address. It should be noted that in the case of IPv6, a link-local IP address does not count in a many-to-one relationship because that address is confined to a single Ethernet segment, and it is not used for host mobility (i.e., by definition, host mobility is between two different Ethernet segments). Therefore, when an IPv6 host is configured with both a Global Unicast address (or a Unique Local address) and a link-local address, for the purpose of host mobility, it is considered with a single IP address.

### 7.1. Initiating a Gratuitous 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 the 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 (IP, MAC) 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 the 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 the 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 the 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 a route previously). Furthermore, it searches for the corresponding MAC-IP entry and sends an ARP probe for this (IP, MAC) pair. The ARP request message is sent both locally to all attached TSs in that subnet as well as 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 TSs, and when it receives the ARP response, it updates its IP-VRF and ARP table with the host (IP, MAC) information. It also sends an EVPN MAC/IP Advertisement route with both the MAC and IP addresses filled in along with the MAC Mobility extended community, with the sequence number set to the same value as the one for the 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 (IP, MAC) 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 the 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 an EVPN-IRB NVE is 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. 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 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 an ARP timer or a MAC age-out timer, and this is an implementation choice. The ARP request gets sent both locally to all the attached TSs on that subnet as well as to all the remote NVEs (including the target NVE) participating in that subnet. The source NVE also withdraws 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 TSs, and when it receives the ARP response, it updates its MAC-VRF, IP-VRF, and ARP table with the host (IP, MAC) 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 the 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 (IP, MAC) 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 the 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. EVPN Router’s MAC Extended Community

A new EVPN BGP Extended Community called "EVPN Router’s MAC" is introduced here. This new extended community is a transitive extended community with a Type field of 0x06 (EVPN) and a Sub-Type field of 0x03. It may be advertised along with the Encapsulation Extended Community defined in Section 4.1 of [RFC9012].

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

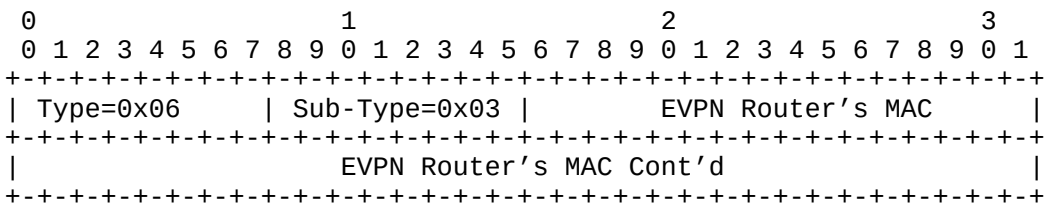



Figure 5: EVPN Router’s MAC Extended Community

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 EVPN Router’s MAC Extended Community to a route. In case the receiving PE receives more than one EVPN 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 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(s) 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, each MAC-VRF consists of multiple bridge tables (e.g., one bridge table 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 an inner Ethernet header (inner MAC SA/DA) and since a TS MAC address cannot be used for inter-subnet traffic, the ingress NVE's MAC address is used as an 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 TSs that are associated with these three MAC-VRFs -- i.e., TS1, TS4, and TS5 are on the same subnet (e.g., the 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 TSs belong to the same subnet. However, when TS1 wants to exchange traffic with TS2 or TS3, which belong to different subnets, both the bridging and routing parts of the IRB solution are exercised. The following subsections describe the control and data plane operations for this IRB scenario in detail.

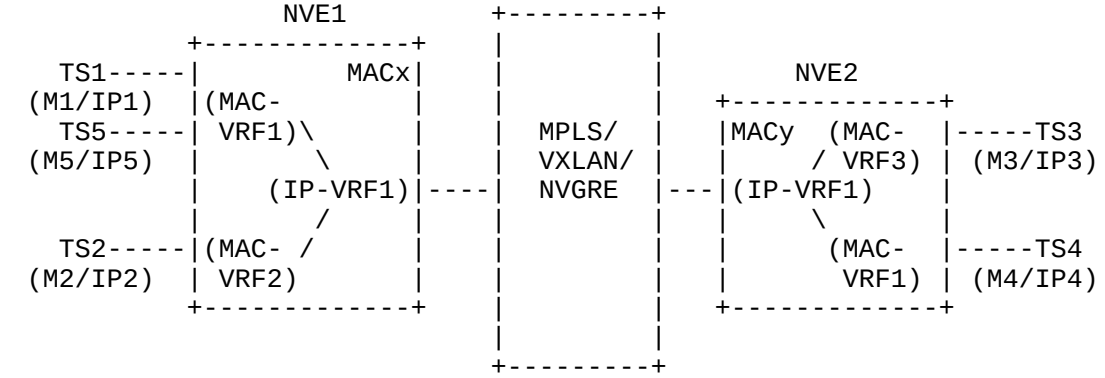



Figure 6: IRB Forwarding on NVEs for Tenant Systems

### 9.1.1. Control Plane Operation

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

* RD and Ethernet Segment Identifier (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 Figure 6, above

* IP Address Length = 32 or 128

* IP Address = IPi (where i = 1, 2, 3, 4, or 5) in Figure 6, above

* 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 [RFC9012], 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 EVPN Router's MAC Extended Community MUST be added when the Ethernet NVO tunnel is used. If the IP NVO tunnel type is used, then there is no need to send this second Extended Community. It should be noted that the 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 the MAC/IP Advertisement route into the MAC-VRF with the BGP next-hop address as the underlay tunnel destination address (e.g., VTEP DA for VXLAN encapsulation) and label1 as the VNI for VXLAN encapsulation or an EVPN label for MPLS encapsulation.

* If the route carries the new EVPN Router's MAC Extended Community and if the receiving NVE uses an Ethernet NVO tunnel, then the receiving NVE imports the IP address into IP-VRF with NVE's MAC address (from the new EVPN Router's MAC Extended Community) as the inner MAC DA, the BGP next-hop address as the underlay tunnel destination address, the VTEP DA for VXLAN encapsulation, and label2 as the 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 the BGP next-hop address as the underlay tunnel destination address and label2 as the 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; an EVPN RT-2 with only a single Route Target corresponding to MAC-VRF but with both label1 and label2; or an EVPN RT-2 with a 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. Let's consider the 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 the MAC DA corresponding to the MAC-VRF1 IRB interface on NVE1 (the interface between MAC-VRF1 and IP-VRF1) and the VLAN tag corresponding to MAC-VRF1.

* Upon receiving the packet, the NVE1 uses the 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 the Ethernet NVO tunnel, then it includes the egress NVE's MAC address as the 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 an MPLS encapsulation, then the corresponding EVPN and LSP labels are added to the packet. The packet is then forwarded to the egress NVE.

* If the egress NVE receives a packet from the Ethernet NVO tunnel (e.g., it is VXLAN encapsulated), then it removes the Ethernet header. 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 the MAC SA set to that of the IRB interface MAC address (i.e., the IRB interface between MAC-VRF3 and IP-VRF1 on NVE2) and the MAC DA set to that of the 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 TSs support one or more subnets and these TSs are associated with one or more NVEs. Therefore, besides the advertisement of MAC/IP addresses for each TS, which can be multihomed 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 [RFC9136]. These subnet prefixes are advertised with the IP address of their associated TS (which is in an 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 that is needed is to advertise the IP/MAC addresses associated with the TS itself and exercise the 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 TSs 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 the 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 the bridging and routing parts of the IRB solution are exercised (i.e., the traffic goes through the corresponding MAC-VRFs and IP-VRFs). If TS4, for example, wants to reach SN1, it uses its default route and sends the packet to the MAC address associated with the IRB interface on NVE2; NVE2 then performs an IP lookup in its IP-VRF and finds an entry for SN1. The following subsections describe the control and data plane operations for this IRB scenario in detail.

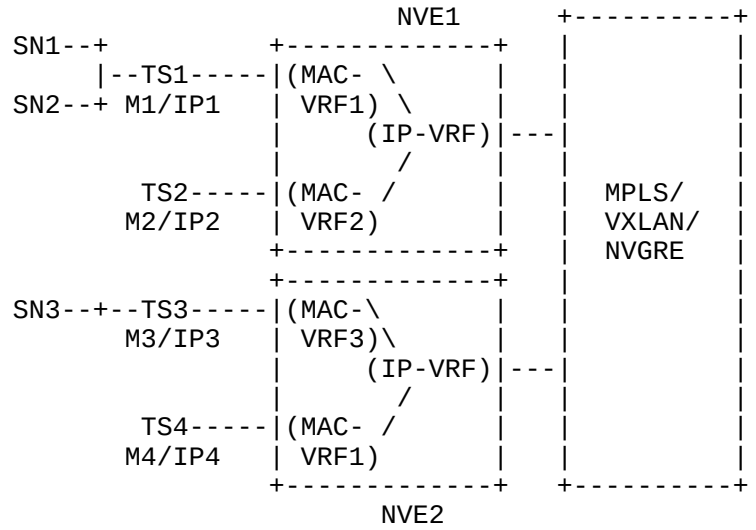



Figure 7: IRB Forwarding on NVEs for Subnets behind TSs

Note that in Figure 7, above, SN1 and SN2 are configured on NVE1, which then advertises each in an IP Prefix route. Similarly, SN3 is configured on NVE2, which then advertises it in an IP Prefix route.

### 9.2.1. Control Plane Operation

Each NVE advertises a route type 5 (EVPN RT-5, IP Prefix route defined in [RFC9136]) 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 its associated Route Targets and Extended Communities for each of its TSs 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 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 the 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). The BGP next hop will be used as the underlay tunnel destination address (e.g., VTEP DA for VXLAN encapsulation), and the EVPN Router's MAC will be used as the 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. Let's consider the data plane operation when a host in SN1 behind TS1 wants to send traffic to a host in SN3 behind TS3.

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

* Upon receiving the packet, the ingress NVE1 uses the VLAN tag to identify the MAC-VRF1. It then looks up the MAC DA and forwards the frame to its IRB interface as in Section 9.1.1.

* 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 address. This lookup yields the fields needed for VXLAN encapsulation with NVE2's MAC address as the inner MAC DA, NVE2's IP address as the VTEP DA, and the VNI. The MAC SA is set to NVE1's MAC address, and the 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 VNI.

* Next, a lookup is performed based on the 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 the 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. Security Considerations

The security considerations for Layer 2 forwarding in this document follow those of [RFC7432] for MPLS encapsulation and those 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 of 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 TSs 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, and 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.

## 11. IANA Considerations

IANA has allocated Sub-Type value 0x03 in the "EVPN Extended Community Sub-Types" registry as follows:

| Sub-Type Value | Name | Reference |
|---|---|---|
| 0x03 | EVPN Router’s MAC Extended Community | RFC 9135 |

Table 1

This document has been listed as an additional reference for the MAC/

IP Advertisement route in the "EVPN Route Types" registry.

## 12. References

### 12.1. Normative References


[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[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>.

[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>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[RFC9136] Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021, <https://www.rfc-editor.org/info/rfc9136>.


### 12.2. Informative References


[EVPN] Krattiger, L., Ed., Sajassi, A., Ed., Thoria, S., Rabadan, J., and J. Drake, "EVPN Interoperability Modes", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-modes-interop-00, 26 May 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-modes-interop-00>.

[EXTENDED-MOBILITY] Malhotra, N., Ed., Sajassi, A., Pattekar, A., Rabadan, J., Lingala, A., and J. Drake, "Extended Mobility Procedures for EVPN-IRB", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-irb-extended-mobility-07, 2 October 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-irb-extended-mobility-07>.

[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>.

[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>.

[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>.

[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>.

[VXLAN-GPE] Maino, F., Ed., Kreeger, L., Ed., and U. Elzur, Ed., "Generic Protocol Extension for VXLAN (VXLAN-GPE)", Work in Progress, Internet-Draft, draft-ietf-nvo3-vxlan-gpe-12, 22 September 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-nvo3-vxlan-gpe-12>.

## 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.

## Authors' Addresses

Ali Sajassi
Cisco Systems

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