BESS Workgroup
Internet Draft
Intended status: Standards Track

J. Rabadan, Ed.
W. Henderickx
Nokia

J. Drake
W. Lin
Juniper

A. Sajassi
Cisco

Expires: November 19, 2018

May 18, 2018

IP Prefix Advertisement in EVPN
draft-ietf-bess-evpn-prefix-advertisement-11

Abstract

The BGP MPLS-based Ethernet VPN (EVPN) [RFC7432] mechanism provides a flexible control plane that allows intra-subnet connectivity in an MPLS and/or NVO (Network Virtualization Overlay) [RFC7365] network. In some networks, there is also a need for a dynamic and efficient inter-subnet connectivity across Tenant Systems and End Devices that can be physical or virtual and do not necessarily participate in dynamic routing protocols. This document defines a new EVPN route type for the advertisement of IP Prefixes and explains some use-case examples where this new route-type is used.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at

http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on November 19, 2018.

Table of Contents

1. Introduction

[RFC7365] provides a framework for Data Center (DC) Network Virtualization over Layer 3 and specifies that the Network Virtualization Edge devices (NVEs) must provide layer 2 and layer 3 virtualized network services in multi-tenant DCs. [RFC8365] discusses the use of EVPN as the technology of choice to provide layer 2 or intra-subnet services in these DCs. This document, along with [EVPN-INTERSUBNET], specifies the use of EVPN for layer 3 or inter-subnet connectivity services.

[EVPN-INTERSUBNET] defines some fairly common inter-subnet forwarding scenarios where TSes can exchange packets with TSes located in remote subnets. In order to achieve this, [EVPN-INTERSUBNET] describes how MAC/IPs encoded in TS RT-2 routes are not only used to populate MAC-VRF and overlay ARP tables, but also IP-VRF tables with the encoded TS host routes (/32 or /128). In some cases, EVPN may advertise IP Prefixes and therefore provide aggregation in the IP-VRF tables, as opposed to propagate individual host routes. This document complements the scenarios described in [EVPN-INTERSUBNET] and defines how EVPN may be used to advertise IP Prefixes. Interoperability between EVPN and L3VPN [RFC4364] IP Prefix routes is out of the scope of this document.

Section 2.1 describes the inter-subnet connectivity requirements in Data Centers. Section 2.2 explains why a new EVPN route type is required for IP Prefix advertisements. Sections 3, 4 and 5 will describe this route type and how it is used in some specific use cases.

1.1 Terminology

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

AC: Attachment Circuit.

ARP: Address Resolution Protocol.

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

BD Route Target: refers to the Broadcast Domain assigned Route Target [RFC4364]. In 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].

DGW: Data Center Gateway.

Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].

Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload. Examples of this type of tunnels are VXLAN or GENEVE.

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].

GRE: Generic Routing Encapsulation.

GW IP: Gateway IP Address.

IPL: IP Prefix Length.

IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload).

IP-VRF: A VPN 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.

ML: MAC address length.

ND: Neighbor Discovery Protocol.

NVE: Network Virtualization Edge.

GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].

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.

SBD: Supplementary Broadcast Domain. A BD that does not have any ACs, only IRB interfaces, and it is used to provide connectivity among all the IP-VRFs of the tenant. The SBD is only required in IP-VRF-to-IP-VRF use-cases (see Section 4.4.).

SN: Subnet.

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 Network Identifier in GENEVE, 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. Problem Statement

This Section describes the inter-subnet connectivity requirements in Data Centers and why a specific route type to advertise IP Prefixes is needed.

2.1 Inter-Subnet Connectivity Requirements in Data Centers

[RFC7432] is used as the control plane for a Network Virtualization Overlay (NVO) solution in Data Centers (DC), where Network Virtualization Edge (NVE) devices can be located in Hypervisors or

Top of Rack switches (ToRs), as described in [RFC8365].

The following considerations apply to Tenant Systems (TS) that are physical or virtual systems identified by MAC and maybe IP addresses and connected to BDs by Attachment Circuits:

- The Tenant Systems may be Virtual Machines (VMs) that generate traffic from their own MAC and IP.
- The Tenant Systems may be Virtual Appliance entities (VAs) that forward traffic to/from IP addresses of different End Devices sitting behind them.
    - These VAs can be firewalls, load balancers, NAT devices, other appliances or virtual gateways with virtual routing instances.
    - These VAs do not necessarily participate in dynamic routing protocols and hence rely on the EVPN NVEs to advertise the routes on their behalf.
    - In all these cases, the VA will forward traffic to other TSes using its own source MAC but the source IP will be the one associated to the End Device sitting behind or a translated IP address (part of a public NAT pool) if the VA is performing NAT.
    - Note that the same IP address and endpoint could exist behind two of these TSes. One example of this would be certain appliance resiliency mechanisms, where a virtual IP or floating IP can be owned by one of the two VAs running the resiliency protocol (the master VA). Virtual Router Redundancy Protocol (VRRP), RFC5798, is one particular example of this. Another example is multi-homed subnets, i.e., the same subnet is connected to two VAs.
    - Although these VAs provide IP connectivity to VMs and subnets behind them, they do not always have their own IP interface connected to the EVPN NVE, e.g., layer 2 firewalls are examples of VAs not supporting IP interfaces.

Figure 1 illustrates some of the examples described above.

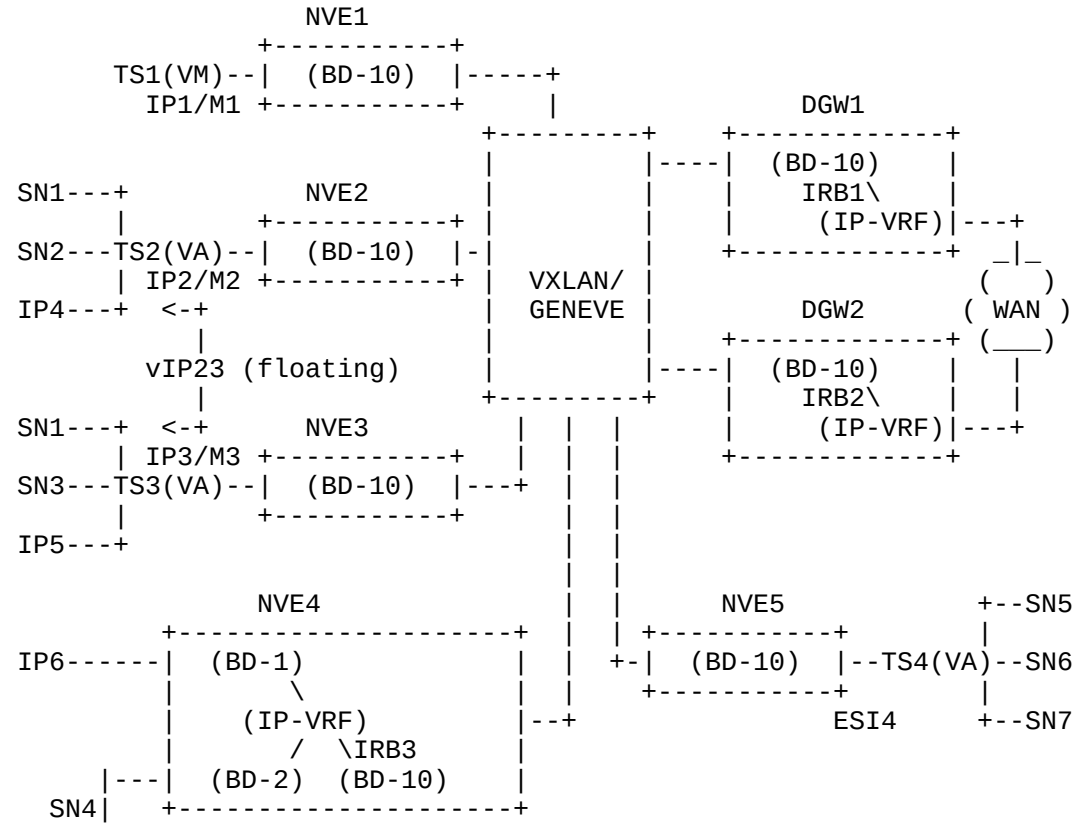

Figure 1 DC inter-subnet use-cases

Where:

NVE1, NVE2, NVE3, NVE4, NVE5, DGW1 and DGW2 share the same BD for a particular tenant. BD-10 is comprised of the collection of BD instances defined in all the NVEs. All the hosts connected to BD-10 belong to the same IP subnet. The hosts connected to BD-10 are listed below:

- o TS1 is a VM that generates/receives traffic from/to IP1, where IP1 belongs to the BD-10 subnet.

- o TS2 and TS3 are Virtual Appliances (VA) that send/receive traffic from/to the subnets and hosts sitting behind them (SN1, SN2, SN3, IP4 and IP5). Their IP addresses (IP2 and IP3) belong to the BD-10 subnet and they can also generate/receive traffic. When these VAs receive packets destined to their own MAC addresses (M2 and M3) they will route the packets to the proper subnet or host. These VAs

do not support routing protocols to advertise the subnets connected to them and can move to a different server and NVE when the Cloud Management System decides to do so. These VAs may also support redundancy mechanisms for some subnets, similar to VRRP, where a floating IP is owned by the master VA and only the master VA forwards traffic to a given subnet. E.g.,: vIP23 in Figure 1 is a floating IP that can be owned by TS2 or TS3 depending on which system is the master. Only the master will forward traffic to SN1.

- o Integrated Routing and Bridging interfaces IRB1, IRB2 and IRB3 have their own IP addresses that belong to the BD-10 subnet too. These IRB interfaces connect the BD-10 subnet to Virtual Routing and Forwarding (IP-VRF) instances that can route the traffic to other subnets for the same tenant (within the DC or at the other end of the WAN).

- o TS4 is a layer 2 VA that provides connectivity to subnets SN5, SN6 and SN7, but does not have an IP address itself in the BD-10. TS4 is connected to a port on NVE5 assigned to Ethernet Segment Identifier 4.

For a BD that an ingress NVE is attached to, "Overlay Index" is defined as an identifier that the ingress EVPN NVE requires in order to forward packets to a subnet or host in a remote subnet. As an example, vIP23 (Figure 1) is an Overlay Index that any NVE attached to BD-10 needs to know in order to forward packets to SN1. IRB3 IP address is an Overlay Index required to get to SN4, and ESI4 (Ethernet Segment Identifier 4) is an Overlay Index needed to forward traffic to SN5. In other words, the Overlay Index is a next-hop in the overlay address space that can be an IP address, a MAC address or an ESI. When advertised along with an IP Prefix, the Overlay Index requires a recursive resolution to find out to what egress NVE the EVPN packets need to be sent.

All the DC use cases in Figure 1 require inter-subnet forwarding and therefore, the individual host routes and subnets:

a) must be advertised from the NVEs (since VAs and VMs do not participate in dynamic routing protocols) and
b) may be associated to an Overlay Index that can be a VA IP address, a floating IP address, a MAC address or an ESI. The Overlay Index is further discussed in Section 3.2.

2.2 The Need for the EVPN IP Prefix Route

[RFC7432] defines a MAC/IP route (also referred as RT-2) where a MAC

address can be advertised together with an IP address length and IP address (IP). While a variable IP address length might have been used to indicate the presence of an IP prefix in a route type 2, there are several specific use cases in which using this route type to deliver IP Prefixes is not suitable.

One example of such use cases is the "floating IP" example described in Section 2.1. In this example it is needed to decouple the advertisement of the prefixes from the advertisement of MAC address of either M2 or M3, otherwise the solution gets highly inefficient and does not scale.

For example, if 1,000 prefixes are advertised from M2 (using RT-2) and the floating IP owner changes from M2 to M3, 1,000 routes would be withdrawn from M2 and readvertise 1k routes from M3. However if a separate route type is used, 1,000 routes can be advertised as associated to the floating IP address (vIP23) and only one RT-2 for advertising the ownership of the floating IP, i.e., vIP23 and M2 in the route type 2. When the floating IP owner changes from M2 to M3, a single RT-2 withdraw/update is required to indicate the change. The remote DGW will not change any of the 1,000 prefixes associated to vIP23, but will only update the ARP resolution entry for vIP23 (now pointing at M3).

An EVPN route (type 5) for the advertisement of IP Prefixes is described in this document. This new route type has a differentiated role from the RT-2 route and addresses the Data Center (or NVO-based networks in general) inter-subnet connectivity scenarios described in this document. Using this new RT-5, an IP Prefix may be advertised along with an Overlay Index that can be a GW IP address, a MAC or an ESI, or without an Overlay Index, in which case the BGP next-hop will point at the egress NVE/ASBR/ABR and the MAC in the Router’s MAC Extended Community will provide the inner MAC destination address to be used. As discussed throughout the document, the EVPN RT-2 does not meet the requirements for all the DC use cases, therefore this EVPN route type 5 is required.

The EVPN route type 5 decouples the IP Prefix advertisements from the MAC/IP route advertisements in EVPN, hence:

a) Allows the clean and clear advertisements of IPv4 or IPv6 prefixes in an NLRI (Network Layer Reachability Information message) with no MAC addresses.

b) Since the route type is different from the MAC/IP Advertisement route, the current [RFC7432] procedures do not need to be modified.

c) Allows a flexible implementation where the prefix can be linked to different types of Overlay/Underlay Indexes: overlay IP address, overlay MAC addresses, overlay ESI, underlay BGP next-hops, etc.

d) An EVPN implementation not requiring IP Prefixes can simply discard them by looking at the route type value.

The following Sections describe how EVPN is extended with a route type for the advertisement of IP prefixes and how this route is used to address the inter-subnet connectivity requirements existing in the Data Center.

3. The BGP EVPN IP Prefix Route

The BGP EVPN NLRI as defined in [RFC7432] is shown below:

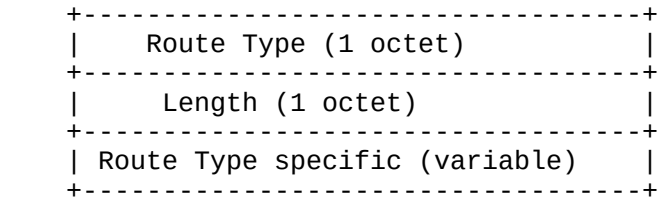

Figure 2 BGP EVPN NLRI

This document defines an additional route type (RT-5) in the IANA EVPN Route Types registry [EVPNRouteTypes], to be used for the advertisement of EVPN routes using IP Prefixes:

Value: 5
Description: IP Prefix Route

According to Section 5.4 in [RFC7606], a node that doesn't recognize the Route Type 5 (RT-5) will ignore it. Therefore an NVE following this document can still be attached to a BD where an NVE ignoring RT-5s is attached to. Regular [RFC7432] procedures would apply in that case for both NVEs. In case two or more NVEs are attached to different BDs of the same tenant, they MUST support RT-5 for the proper Inter-Subnet Forwarding operation of the tenant.

The detailed encoding of this route and associated procedures are described in the following Sections.

3.1 IP Prefix Route Encoding

An IP Prefix Route Type for IPv4 has the Length field set to 34 and consists of the following fields:

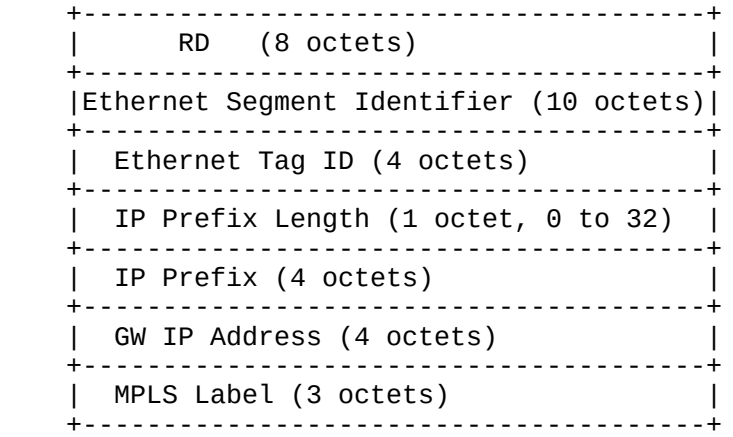

Figure 3 EVPN IP Prefix route NLRI for IPv4

An IP Prefix Route Type for IPv6 has the Length field set to 58 and consists of the following fields:

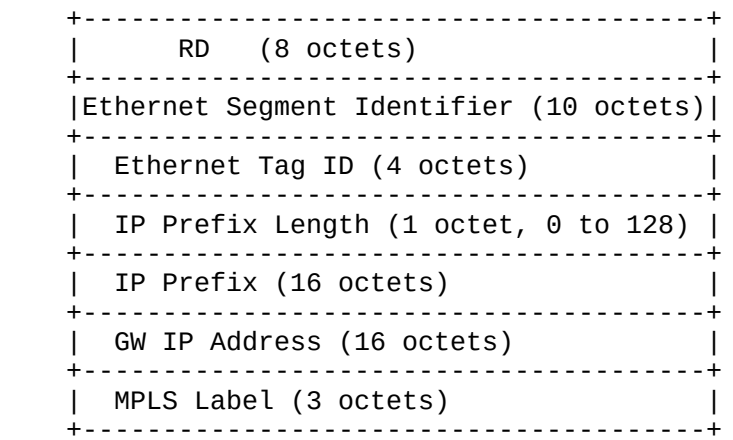

Figure 4 EVPN IP Prefix route NLRI for IPv6

Where:

- The Length field of the BGP EVPN NLRI for an EVPN IP Prefix route MUST be either 34 (if IPv4 addresses are carried) or 58 (if IPv6 addresses are carried). The IP Prefix and Gateway IP Address MUST be from the same IP address family.

- Route Distinguisher (RD) and Ethernet Tag ID MUST be used as defined in [RFC7432] and [RFC8365]. In particular, the RD is unique per MAC-VRF (or IP-VRF). The MPLS Label field is set to either an MPLS label or a VNI, as described in [RFC8365] for other EVPN route types.

- The Ethernet Segment Identifier MUST be a non-zero 10-octet identifier if the ESI is used as an Overlay Index (see the definition of Overlay Index in Section 3.2). It MUST be all bytes zero otherwise. The ESI format is described in [RFC7432].

- The IP Prefix Length can be set to a value between 0 and 32 (bits) for IPv4 and between 0 and 128 for IPv6, and specifies the number of bits in the Prefix. The value MUST NOT be greater than 128.

- The IP Prefix is a 4 or 16-octet field (IPv4 or IPv6).

- The GW (Gateway) IP Address field is a 4 or 16-octet field (IPv4 or IPv6), and will encode a valid IP address as an Overlay Index for the IP Prefixes. The GW IP field MUST be all bytes zero if it is not used as an Overlay Index. Refer to Section 3.2 for the definition and use of the Overlay Index.

- The MPLS Label field is encoded as 3 octets, where the high-order 20 bits contain the label value, as per [RFC7432]. When sending, the label value SHOULD be zero if recursive resolution based on overlay index is used. If the received MPLS Label value is zero, the route MUST contain an Overlay Index and the ingress NVE/PE MUST do recursive resolution to find the egress NVE/PE. If the received Label is zero and the route does not contain an Overlay Index, it MUST be treat-as-withdraw [RFC7606].

The RD, Ethernet Tag ID, IP Prefix Length and IP Prefix are part of the route key used by BGP to compare routes. The rest of the fields are not part of the route key.

An IP Prefix Route MAY be sent along with a Router’s MAC Extended Community (defined in [EVPN-INTERSUBNET]) to carry the MAC address that is used as the overlay index. Note that the MAC address may be that of an TS.

As described in Section 3.2, certain data combinations in a received routes would imply a "treat-as-withdraw" handling of the route

[RFC7606].

3.2 Overlay Indexes and Recursive Lookup Resolution

RT-5 routes support recursive lookup resolution through the use of Overlay Indexes as follows:

- An Overlay Index can be an ESI, IP address in the address space of the tenant or MAC address and it is used by an NVE as the next-hop for a given IP Prefix. An Overlay Index always needs a recursive route resolution on the NVE/PE that installs the RT-5 into one of its IP-VRFs, so that the NVE knows to which egress NVE/PE it needs to forward the packets. It is important to note that recursive resolution of the Overlay Index applies upon installation into an IP-VRF, and not upon BGP propagation (for instance, on an ASBR). Also, as a result of the recursive resolution, the egress NVE/PE is not necessarily the same NVE that originated the RT-5.

- The Overlay Index is indicated along with the RT-5 in the ESI field, GW IP field or Router’s MAC Extended Community, depending on whether the IP Prefix next-hop is an ESI, IP address or MAC address in the tenant space. The Overlay Index for a given IP Prefix is set by local policy at the NVE that originates an RT-5 for that IP Prefix (typically managed by the Cloud Management System).

- In order to enable the recursive lookup resolution at the ingress NVE, an NVE that is a possible egress NVE for a given Overlay Index must originate a route advertising itself as the BGP next hop on the path to the system denoted by the Overlay Index. For instance:

  . If an NVE receives an RT-5 that specifies an Overlay Index, the NVE cannot use the RT-5 in its IP-VRF unless (or until) it can recursively resolve the Overlay Index.
  . If the RT-5 specifies an ESI as the Overlay Index, recursive resolution can only be done if the NVE has received and installed an RT-1 (Auto-Discovery per-EVI) route specifying that ESI.
  . If the RT-5 specifies a GW IP address as the Overlay Index, recursive resolution can only be done if the NVE has received and installed an RT-2 (MAC/IP route) specifying that IP address in the IP address field of its NLRI.
  . If the RT-5 specifies a MAC address as the Overlay Index, recursive resolution can only be done if the NVE has received and installed an RT-2 (MAC/IP route) specifying that MAC address in the MAC address field of its NLRI.

  Note that the RT-1 or RT-2 routes needed for the recursive resolution may arrive before or after the given RT-5 route.

- Irrespective of the recursive resolution, if there is no IGP or BGP route to the BGP next-hop of an RT-5, BGP MUST NOT install the RT-5 even if the Overlay Index can be resolved.

- The ESI and GW IP fields may both be zero at the same time. However, they MUST NOT both be non-zero at the same time. A route containing a non-zero GW IP and a non-zero ESI (at the same time) SHOULD be treat-as-withdraw [RFC7606].

- If either the ESI or GW IP are non-zero, then the non-zero one is the Overlay Index, regardless of whether the Router's MAC Extended Community is present or the value of the Label. In case the GW IP is the Overlay Index (hence ESI is zero), the Router's MAC Extended Community is ignored if present.

- A route where ESI, GW IP, MAC and Label are all zero at the same time SHOULD be treat-as-withdraw.

The indirection provided by the Overlay Index and its recursive lookup resolution is required to achieve fast convergence in case of a failure of the object represented by the Overlay Index (see the example described in Section 2.2).

Table 1 shows the different RT-5 field combinations allowed by this specification and what Overlay Index must be used by the receiving NVE/PE in each case. Those cases where there is no Overlay Index, are indicated as "None" in Table 1. If there is no Overlay Index the receiving NVE/PE will not perform any recursive resolution, and the actual next-hop is given by the RT-5's BGP next-hop.

ESI	GW IP	MAC*	Label	Overlay Index
Non-Zero	Zero	Zero	Don't Care	ESI
Non-Zero	Zero	Non-Zero	Don't Care	ESI
Zero	Non-Zero	Zero	Don't Care	GW IP
Zero	Zero	Non-Zero	Zero	MAC
Zero	Zero	Non-Zero	Non-Zero	MAC or None**
Zero	Zero	Zero	Non-Zero	None***

Table 1 - RT-5 fields and Indicated Overlay Index

Table NOTES:

* MAC with Zero value means no Router's MAC extended community is present along with the RT-5. Non-Zero indicates that the extended community is present and carries a valid MAC address. The

encoding of a MAC address MUST be the 6-octet MAC address specified by [802.1Q] and [802.1D-REV]. Examples of invalid MAC addresses are broadcast or multicast MAC addresses. The route MUST be treat-as-withdraw in case of an invalid MAC address. The presence of the Router's MAC extended community alone is not enough to indicate the use of the MAC address as the Overlay Index, since the extended community can be used for other purposes.

** In this case, the Overlay Index may be the RT-5's MAC address or None, depending on the local policy of the receiving NVE/PE. Note that the advertising NVE/PE that sets the Overlay Index SHOULD advertise an RT-2 for the MAC Overlay Index if there are receiving NVE/PEs configured to use the MAC as the Overlay Index. This case in Table 1 is used in the IP-VRF-to-IP-VRF implementations described in 4.4.1 and 4.4.3. The support of a MAC Overlay Index in this model is OPTIONAL.

*** The Overlay Index is None. This is a special case used for IP-VRF-to-IP-VRF where the NVE/PEs are connected by IP NVO tunnels as opposed to Ethernet NVO tunnels.

If the combination of ESI, GW IP, MAC and Label in the receiving RT-5 is different than the combinations shown in Table 1, the router will process the route as per the rules described at the beginning of this Section (3.2).

Table 2 shows the different inter-subnet use-cases described in this document and the corresponding coding of the Overlay Index in the route type 5 (RT-5).

Section	Use-case	Overlay Index in the RT-5
4.1	TS IP address	GW IP
4.2	Floating IP address	GW IP
4.3	"Bump in the wire"	ESI or MAC
4.4	IP-VRF-to-IP-VRF	GW IP, MAC or None

Table 2 - Use-cases and Overlay Indexes for Recursive Resolution

The above use-cases are representative of the different Overlay Indexes supported by RT-5 (GW IP, ESI, MAC or None).

4. Overlay Index Use-Cases

This Section describes some use-cases for the Overlay Index types used with the IP Prefix route. Although the examples use IPv4 Prefixes and subnets, the descriptions of the RT-5 are valid for the same cases with IPv6, only replacing the IP Prefixes, IPL and GW IP by the corresponding IPv6 values.

4.1 TS IP Address Overlay Index Use-Case

Figure 5 illustrates an example of inter-subnet forwarding for subnets sitting behind Virtual Appliances (on TS2 and TS3).

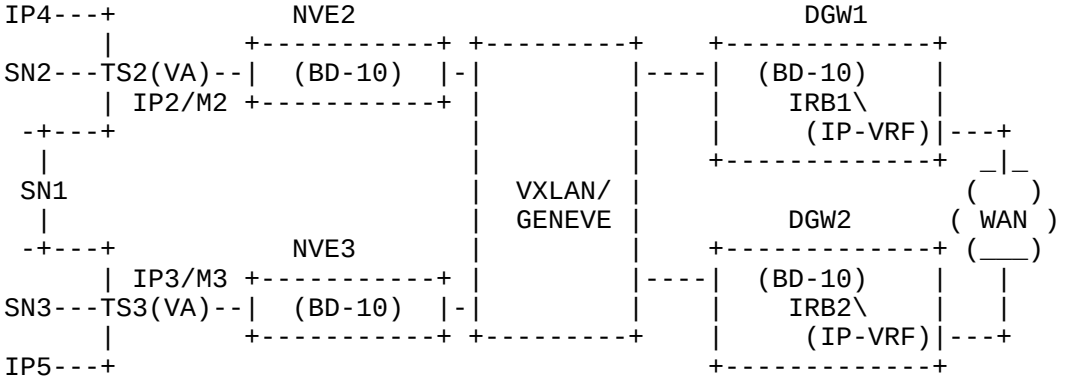

Figure 5 TS IP address use-case

An example of inter-subnet forwarding between subnet SN1, which uses a 24 bit IP prefix (written as SN1/24 in future), and a subnet sitting in the WAN is described below. NVE2, NVE3, DGW1 and DGW2 are running BGP EVPN. TS2 and TS3 do not participate in dynamic routing protocols, and they only have a static route to forward the traffic to the WAN. SN1/24 is dual-homed to NVE2 and NVE3.

In this case, a GW IP is used as an Overlay Index. Although a different Overlay Index type could have been used, this use-case assumes that the operator knows the VA's IP addresses beforehand, whereas the VA's MAC address is unknown and the VA's ESI is zero. Because of this, the GW IP is the suitable Overlay Index to be used with the RT-5s. The NVEs know the GW IP to be used for a given Prefix by policy.

(1) NVE2 advertises the following BGP routes on behalf of TS2:

- Route type 2 (MAC/IP route) containing: ML=48 (MAC Address Length), M=M2 (MAC Address), IPL=32 (IP Prefix Length), IP=IP2 and [RFC5512] BGP Encapsulation Extended Community with the corresponding Tunnel type. The MAC and IP addresses may be

learned via ARP snooping.

- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP2. The prefix and GW IP are learned by policy.

(2) Similarly, NVE3 advertises the following BGP routes on behalf of TS3:

- o Route type 2 (MAC/IP route) containing: ML=48, M=M3, IPL=32, IP=IP3 (and BGP Encapsulation Extended Community).
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP3.

(3) DGW1 and DGW2 import both received routes based on the Route Targets:

- o Based on the BD-10 Route Target in DGW1 and DGW2, the MAC/IP route is imported and M2 is added to the BD-10 along with its corresponding tunnel information. For instance, if VXLAN is used, the VTEP will be derived from the MAC/IP route BGP next-hop and VNI from the MPLS Label1 field. IP2 - M2 is added to the ARP table. Similarly, M3 is added to BD-10 and IP3 - M3 to the ARP table.
- o Based on the BD-10 Route Target in DGW1 and DGW2, the IP Prefix route is also imported and SN1/24 is added to the IP-VRF with Overlay Index IP2 pointing at the local BD-10. In this example, it is assumed that the RT-5 from NVE2 is preferred over the RT-5 from NVE3. If both routes were equally preferable and ECMP enabled, SN1/24 would also be added to the routing table with Overlay Index IP3.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 IP-VRF routing table and Overlay Index=IP2 is found. Since IP2 is an Overlay Index a recursive route resolution is required for IP2.
- o IP2 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the BD FIB (e.g., remote VTEP and VNI for the VXLAN case).
- o The IP packet destined to IPx is encapsulated with:

. Source inner MAC = IRB1 MAC.

. Destination inner MAC = M2.

. Tunnel information provided by the BD (VNI, VTEP IPs and MACs for the VXLAN case).

(5) When the packet arrives at NVE2:

- Based on the tunnel information (VNI for the VXLAN case), the BD-10 context is identified for a MAC lookup.

- Encapsulation is stripped off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(6) Should TS2 move from NVE2 to NVE3, MAC Mobility procedures will be applied to the MAC route IP2/M2, as defined in [RFC7432]. Route type 5 prefixes are not subject to MAC mobility procedures, hence no changes in the DGW IP-VRF routing table will occur for TS2 mobility, i.e., all the prefixes will still be pointing at IP2 as Overlay Index. There is an indirection for e.g., SN1/24, which still points at Overlay Index IP2 in the routing table, but IP2 will be simply resolved to a different tunnel, based on the outcome of the MAC mobility procedures for the MAC/IP route IP2/M2.

Note that in the opposite direction, TS2 will send traffic based on its static-route next-hop information (IRB1 and/or IRB2), and regular EVPN procedures will be applied.

4.2 Floating IP Overlay Index Use-Case

Sometimes Tenant Systems (TS) work in active/standby mode where an upstream floating IP - owned by the active TS - is used as the Overlay Index to get to some subnets behind. This redundancy mode, already introduced in Section 2.1 and 2.2, is illustrated in Figure 6.

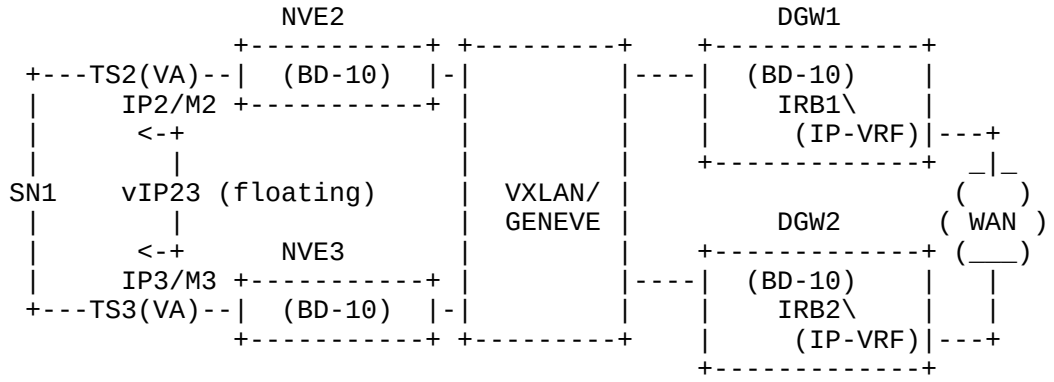

Figure 6 Floating IP Overlay Index for redundant TS

In this use-case, a GW IP is used as an Overlay Index for the same reasons as in 4.1. However, this GW IP is a floating IP that belongs to the active TS. Assuming TS2 is the active TS and owns vIP23:

(1) NVE2 advertises the following BGP routes for TS2:

- Route type 2 (MAC/IP route) containing: ML=48, M=M2, IPL=32, IP=vIP23 (and BGP Encapsulation Extended Community). The MAC and IP addresses may be learned via ARP snooping.

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=vIP23. The prefix and GW IP are learned by policy.

(2) NVE3 advertises the following BGP route for TS3 (it does not advertise an RT-2 for vIP23/M3):

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=vIP23. The prefix and GW IP are learned by policy.

(3) DGW1 and DGW2 import both received routes based on the Route Target:

- M2 is added to the BD-10 FIB along with its corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC/IP route BGP next-hop and VNI from the VNI field. vIP23 - M2 is added to the ARP table.

- SN1/24 is added to the IP-VRF in DGW1 and DGW2 with Overlay index vIP23 pointing at M2 in the local BD-10.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 IP-VRF routing table and Overlay Index=vIP23 is found. Since vIP23 is an Overlay Index, a recursive route resolution for vIP23 is required.
- o vIP23 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the BD (remote VTEP and VNI for the VXLAN case).
- o The IP packet destined to IPx is encapsulated with:
  - . Source inner MAC = IRB1 MAC.
  - . Destination inner MAC = M2.
  - . Tunnel information provided by the BD FIB (VNI, VTEP IPs and MACs for the VXLAN case).

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the BD-10 context is identified for a MAC lookup.
- o Encapsulation is stripped off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(6) When the redundancy protocol running between TS2 and TS3 appoints TS3 as the new active TS for SN1, TS3 will now own the floating vIP23 and will signal this new ownership, using a gratuitous ARP REPLY message (explained in [RFC5227]) or similar. Upon receiving the new owner’s notification, NVE3 will issue a route type 2 for M3-vIP23 and NVE2 will withdraw the RT-2 for M2-vIP23. DGW1 and DGW2 will update their ARP tables with the new MAC resolving the floating IP. No changes are made in the IP-VRF routing table.

4.3 Bump-in-the-Wire Use-Case

Figure 7 illustrates an example of inter-subnet forwarding for an IP Prefix route that carries a subnet SN1. In this use-case, TS2 and TS3 are layer 2 VA devices without any IP address that can be included as an Overlay Index in the GW IP field of the IP Prefix route. Their MAC addresses are M2 and M3 respectively and are connected to BD-10. Note that IRB1 and IRB2 (in DGW1 and DGW2 respectively) have IP addresses

in a subnet different than SN1.

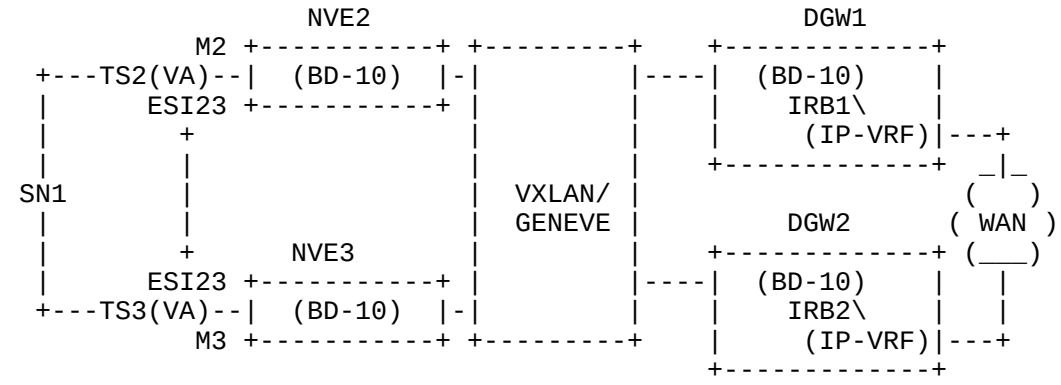

Figure 7 Bump-in-the-wire use-case

Since neither TS2 nor TS3 can participate in any dynamic routing protocol and have no IP address assigned, there are two potential Overlay Index types that can be used when advertising SN1:

a) an ESI, i.e., ESI23, that can be provisioned on the attachment ports of NVE2 and NVE3, as shown in Figure 7.
b) or the VA's MAC address, that can be added to NVE2 and NVE3 by policy.

The advantage of using an ESI as Overlay Index as opposed to the VA's MAC address, is that the forwarding to the egress NVE can be done purely based on the state of the AC in the ES (notified by the Ethernet A-D per-EVI route) and all the EVPN multi-homing redundancy mechanisms can be reused. For instance, the [RFC7432] mass-withdrawal mechanism for fast failure detection and propagation can be used. This Section assumes that an ESI Overlay Index is used in this use-case but it does not prevent the use of the VA's MAC address as an Overlay Index. If a MAC is used as Overlay Index, the control plane must follow the procedures described in Section 4.4.3.

The model supports VA redundancy in a similar way to the one described in Section 4.2 for the floating IP Overlay Index use-case, except that it uses the EVPN Ethernet A-D per-EVI route instead of the MAC advertisement route to advertise the location of the Overlay Index. The procedure is explained below:

(1) Assuming TS2 is the active TS in ESI23, NVE2 advertises the following BGP routes:

- o Route type 1 (Ethernet A-D route for BD-10) containing: ESI=ESI23 and the corresponding tunnel information (VNI field), as well as the BGP Encapsulation Extended Community as per [RFC8365].

- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=ESI23, GW IP address=0. The Router's MAC Extended Community defined in [EVPN-INTERSUBNET] is added and carries the MAC address (M2) associated to the TS behind which SN1 sits. M2 may be learned by policy, however the MAC in the Extended Community is preferred if sent with the route.

(2) NVE3 advertises the following BGP route for TS3 (no AD per-EVI route is advertised):

- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=23, GW IP address=0. The Router's MAC Extended Community is added and carries the MAC address (M3) associated to the TS behind which SN1 sits. M3 may be learned by policy, however the MAC in the Extended Community is preferred if sent with the route.

(3) DGW1 and DGW2 import the received routes based on the Route Target:

- o The tunnel information to get to ESI23 is installed in DGW1 and DGW2. For the VXLAN use case, the VTEP will be derived from the Ethernet A-D route BGP next-hop and VNI from the VNI/VSID field (see [RFC8365]).

- o The RT-5 coming from the NVE that advertised the RT-1 is selected and SN1/24 is added to the IP-VRF in DGW1 and DGW2 with Overlay Index ESI23 and MAC = M2.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 IP-VRF routing table and Overlay Index=ESI23 is found. Since ESI23 is an Overlay Index, a recursive route resolution is required to find the egress NVE where ESI23 resides.

- o The IP packet destined to IPx is encapsulated with:

  . Source inner MAC = IRB1 MAC.

  . Destination inner MAC = M2 (this MAC will be obtained from the Router's MAC Extended Community received along

with the RT-5 for SN1). Note that the Router's MAC Extended Community is used in this case to carry the TS' MAC address, as opposed to the NVE/PE's MAC address.

. Tunnel information for the NVO tunnel is provided by the Ethernet A-D route per-EVI for ESI23 (VNI and VTEP IP for the VXLAN case).

(5) When the packet arrives at NVE2:

- o Based on the tunnel demultiplexer information (VNI for the VXLAN case), the BD-10 context is identified for a MAC lookup (assuming MAC-based disposition model [RFC7432]) or the VNI may directly identify the egress interface (for a MPLS-based disposition model, which in this context is a VNI-based disposition model).

- o Encapsulation is stripped off and based on a MAC lookup (assuming MAC forwarding on the egress NVE) or a VNI lookup (in case of VNI forwarding), the packet is forwarded to TS2, where it will be forwarded to SN1.

(6) If the redundancy protocol running between TS2 and TS3 follows an active/standby model and there is a failure, appointing TS3 as the new active TS for SN1, TS3 will now own the connectivity to SN1 and will signal this new ownership. Upon receiving the new owner's notification, NVE3's AC will become active and issue a route type 1 for ESI23, whereas NVE2 will withdraw its Ethernet A-D route for ESI23. DGW1 and DGW2 will update their tunnel information to resolve ESI23. The destination inner MAC will be changed to M3.

4.4 IP-VRF-to-IP-VRF Model

This use-case is similar to the scenario described in "IRB forwarding on NVEs for Tenant Systems" in [EVPN-INTERSUBNET], however the new requirement here is the advertisement of IP Prefixes as opposed to only host routes.

In the examples described in Sections 4.1, 4.2 and 4.3, the BD instance can connect IRB interfaces and any other Tenant Systems connected to it. EVPN provides connectivity for:

1. Traffic destined to the IRB or TS IP interfaces as well as

2. Traffic destined to IP subnets sitting behind the TS, e.g., SN1 or SN2.

In order to provide connectivity for (1), MAC/IP routes (RT-2) are needed so that IRB or TS MACs and IPs can be distributed. Connectivity type (2) is accomplished by the exchange of IP Prefix routes (RT-5) for IPs and subnets sitting behind certain Overlay Indexes, e.g., GW IP or ESI or TS MAC.

In some cases, IP Prefix routes may be advertised for subnets and IPs sitting behind an IRB. This use-case is referred to as the "IP-VRF-to-IP-VRF" model.

[EVPN-INTERSUBNET] defines an asymmetric IRB model and a symmetric IRB model, based on the required lookups at the ingress and egress NVE: the asymmetric model requires an IP lookup and a MAC lookup at the ingress NVE, whereas only a MAC lookup is needed at the egress NVE; the symmetric model requires IP and MAC lookups at both, ingress and egress NVE. From that perspective, the IP-VRF-to-IP-VRF use-case described in this Section is a symmetric IRB model.

Note that, in an IP-VRF-to-IP-VRF scenario, out of the many subnets that a tenant may have, it may be the case that only a few are attached to a given NVE/PE's IP-VRF. In order to provide inter-subnet connectivity among the set of NVE/PEs where the tenant is connected, a new SBD is created on all of them if recursive resolution is needed. This SBD is instantiated as a regular BD (with no ACs) in each NVE/PE and has an IRB interface that connects the SBD to the IP-VRF. The IRB interface's IP or MAC address is used as the overlay index for recursive resolution.

Depending on the existence and characteristics of the SBD and IRB interfaces for the IP-VRFs, there are three different IP-VRF-to-IP-VRF scenarios identified and described in this document:

1) Interface-less model: no SBD and no overlay indexes required.
2) Interface-ful with SBD IRB model: it requires SBD, as well as GW IP addresses as overlay indexes.
3) Interface-ful with unnumbered SBD IRB model: it requires SBD, as well as MAC addresses as overlay indexes.

Inter-subnet IP multicast is outside the scope of this document.

4.4.1 Interface-less IP-VRF-to-IP-VRF Model

Figure 8 will be used for the description of this model.

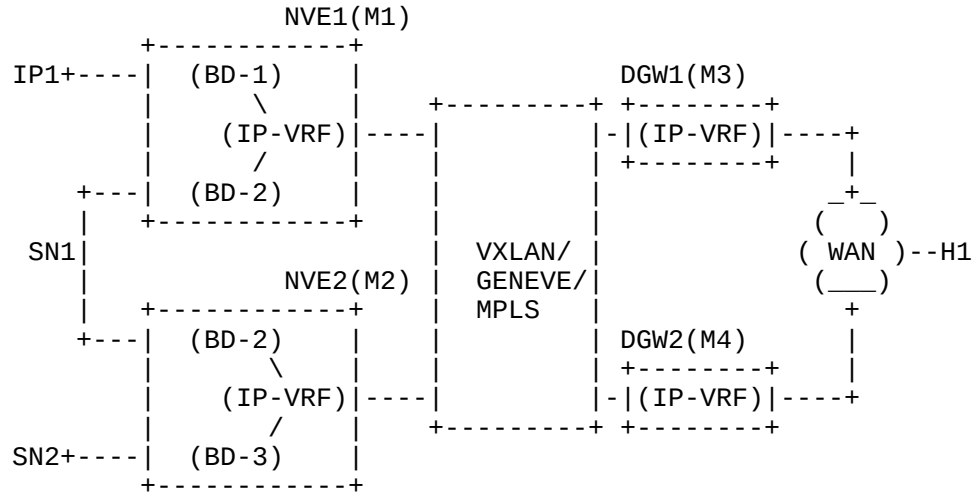

Figure 8 Interface-less IP-VRF-to-IP-VRF model

In this case:

a) The NVEs and DGWs must provide connectivity between hosts in SN1, SN2, IP1 and hosts sitting at the other end of the WAN, for example, H1. It is assumed that the DGWs import/export IP and/or VPN-IP routes from/to the WAN.

b) The IP-VRF instances in the NVE/DGWs are directly connected through NVO tunnels, and no IRBs and/or BD instances are instantiated to connect the IP-VRFs.

c) The solution must provide layer 3 connectivity among the IP-VRFs for Ethernet NVO tunnels, for instance, VXLAN or GENEVE.

d) The solution may provide layer 3 connectivity among the IP-VRFs for IP NVO tunnels, for example, GENEVE (with IP payload).

In order to meet the above requirements, the EVPN route type 5 will be used to advertise the IP Prefixes, along with the Router’s MAC Extended Community as defined in [EVPN-INTERSUBNET] if the advertising NVE/DGW uses Ethernet NVO tunnels. Each NVE/DGW will advertise an RT-5 for each of its prefixes with the following fields:

- RD as per [RFC7432].

- o Ethernet Tag ID=0.
- o IP Prefix Length and IP address, as explained in the previous Sections.
- o GW IP address=0.
- o ESI=0
- o MPLS label or VNI corresponding to the IP-VRF.

Each RT-5 will be sent with a Route Target identifying the tenant (IP-VRF) and may be sent with two BGP extended communities:

- o The first one is the BGP Encapsulation Extended Community, as per [RFC5512], identifying the tunnel type.
- o The second one is the Router's MAC Extended Community as per [EVPN-INTERSUBNET] containing the MAC address associated to the NVE advertising the route. This MAC address identifies the NVE/DGW and MAY be reused for all the IP-VRFs in the NVE. The Router's MAC Extended Community must be sent if the route is associated to an Ethernet NVO tunnel, for instance, VXLAN. If the route is associated to an IP NVO tunnel, for instance GENEVE with IP payload, the Router's MAC Extended Community should not be sent.

The following example illustrates the procedure to advertise and forward packets to SN1/24 (IPv4 prefix advertised from NVE1):

(1) NVE1 advertises the following BGP route:

- o Route type 5 (IP Prefix route) containing:
  - . IPL=24, IP=SN1, Label=10.
  - . GW IP= set to 0.
  - . [RFC5512] BGP Encapsulation Extended Community.
  - . Router's MAC Extended Community that contains M1.
  - . Route Target identifying the tenant (IP-VRF).

(2) DGW1 imports the received routes from NVE1:

- o DGW1 installs SN1/24 in the IP-VRF identified by the RT-5 Route Target.

- o Since GW IP=ESI=0, the Label is a non-zero value and the local policy indicates this interface-less model, DGW1 will use the Label and next-hop of the RT-5, as well as the MAC address conveyed in the Router's MAC Extended Community (as inner destination MAC address) to set up the forwarding state and later encapsulate the routed IP packets.

(3) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 IP-VRF routing table. The lookup yields SN1/24.
- o Since the RT-5 for SN1/24 had a GW IP=ESI=0, a non-zero Label and next-hop and the model is interface-less, DGW1 will not need a recursive lookup to resolve the route.
- o The IP packet destined to IPx is encapsulated with: Source inner MAC = DGW1 MAC, Destination inner MAC = M1, Source outer IP (tunnel source IP) = DGW1 IP, Destination outer IP (tunnel destination IP) = NVE1 IP. The Source and Destination inner MAC addresses are not needed if IP NVO tunnels are used.

(4) When the packet arrives at NVE1:

- o NVE1 will identify the IP-VRF for an IP lookup based on the Label (the Destination inner MAC is not needed to identify the IP-VRF).
- o An IP lookup is performed in the routing context, where SN1 turns out to be a local subnet associated to BD-2. A subsequent lookup in the ARP table and the BD FIB will provide the forwarding information for the packet in BD-2.

The model described above is called Interface-less model since the IP-VRFs are connected directly through tunnels and they don't require those tunnels to be terminated in SBDs instead, as in Sections 4.4.2 or 4.4.3.

4.4.2 Interface-ful IP-VRF-to-IP-VRF with SBD IRB

Figure 9 will be used for the description of this model.

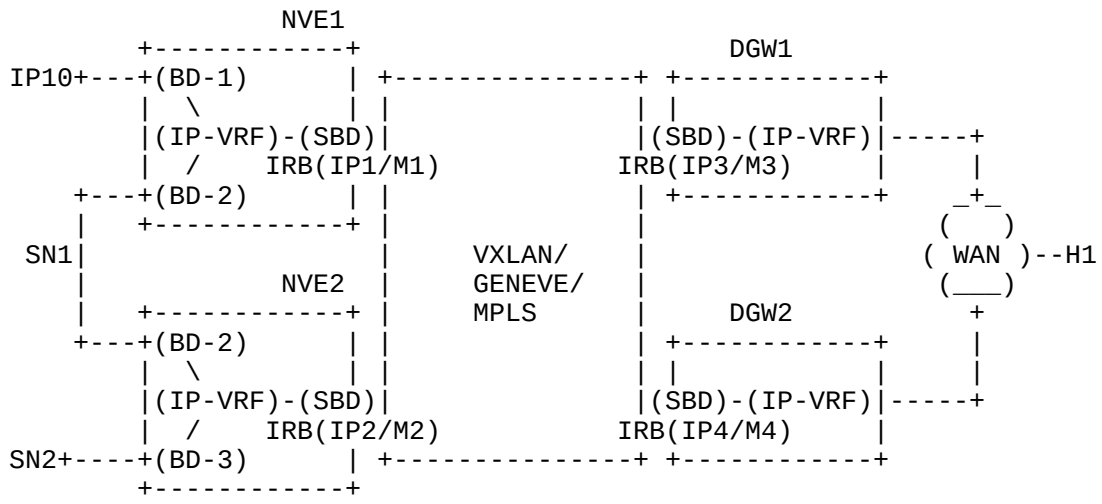

Figure 9 Interface-ful with SBD IRB model

In this model:

a) As in Section 4.4.1, the NVEs and DGWs must provide connectivity between hosts in SN1, SN2, IP10 and hosts sitting at the other end of the WAN.

b) However, the NVE/DGWs are now connected through Ethernet NVO tunnels terminated in the SBD instance. The IP-VRFs use IRB interfaces for their connectivity to the SBD.

c) Each SBD IRB has an IP and a MAC address, where the IP address must be reachable from other NVEs or DGWs.

d) The SBD is attached to all the NVE/DGWs in the tenant domain BDs.

e) The solution must provide layer 3 connectivity for Ethernet NVO tunnels, for instance, VXLAN or GENEVE (with Ethernet payload).

EVPN type 5 routes will be used to advertise the IP Prefixes, whereas EVPN RT-2 routes will advertise the MAC/IP addresses of each SBD IRB interface. Each NVE/DGW will advertise an RT-5 for each of its prefixes with the following fields:

- RD as per [RFC7432].
- Ethernet Tag ID=0.

- o IP Prefix Length and IP address, as explained in the previous Sections.
- o GW IP address=IRB-IP of the SBD (this is the Overlay Index that will be used for the recursive route resolution).
- o ESI=0
- o Label value should be zero since the RT-5 route requires a recursive lookup resolution to an RT-2 route. It is ignored on reception, and, when forwarding packets, the MPLS label or VNI from the RT-2's MPLS Label1 field is used.

Each RT-5 will be sent with a Route Target identifying the tenant (IP-VRF). The Router's MAC Extended Community should not be sent in this case.

The following example illustrates the procedure to advertise and forward packets to SN1/24 (IPv4 prefix advertised from NVE1):

(1) NVE1 advertises the following BGP routes:

- o Route type 5 (IP Prefix route) containing:
  - . IPL=24, IP=SN1, Label= SHOULD be set to 0.
  - . GW IP=IP1 (SBD IRB's IP)
  - . Route Target identifying the tenant (IP-VRF).
- o Route type 2 (MAC/IP route for the SBD IRB) containing:
  - . ML=48, M=M1, IPL=32, IP=IP1, Label=10.
  - . A [RFC5512] BGP Encapsulation Extended Community.
  - . Route Target identifying the SBD. This Route Target may be the same as the one used with the RT-5.

(2) DGW1 imports the received routes from NVE1:

- o DGW1 installs SN1/24 in the IP-VRF identified by the RT-5 Route Target.
  - . Since GW IP is different from zero, the GW IP (IP1) will be used as the Overlay Index for the recursive route resolution to the RT-2 carrying IP1.

(3) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 IP-VRF routing table. The lookup yields SN1/24, which is associated to the Overlay Index IP1. The forwarding information is derived from the RT-2 received for IP1.
- o The IP packet destined to IPx is encapsulated with: Source inner MAC = M3, Destination inner MAC = M1, Source outer IP (source VTEP) = DGW1 IP, Destination outer IP (destination VTEP) = IP1.

(4) When the packet arrives at NVE1:

- o NVE1 will identify the IP-VRF for an IP lookup based on the Label and the inner MAC DA.
- o An IP lookup is performed in the routing context, where SN1 turns out to be a local subnet associated to BD-2. A subsequent lookup in the ARP table and the BD FIB will provide the forwarding information for the packet in BD-2.

The model described above is called 'Interface-ful with SBD IRB model' because the tunnels connecting the DGWs and NVEs need to be terminated into the SBD. The SBD is connected to the IP-VRFs via SBD IRB interfaces, and that allows the recursive resolution of RT-5s to GW IP addresses.

4.4.3 Interface-ful IP-VRF-to-IP-VRF with Unnumbered SBD IRB

Figure 10 will be used for the description of this model. Note that this model is similar to the one described in Section 4.4.2, only without IP addresses on the SBD IRB interfaces.

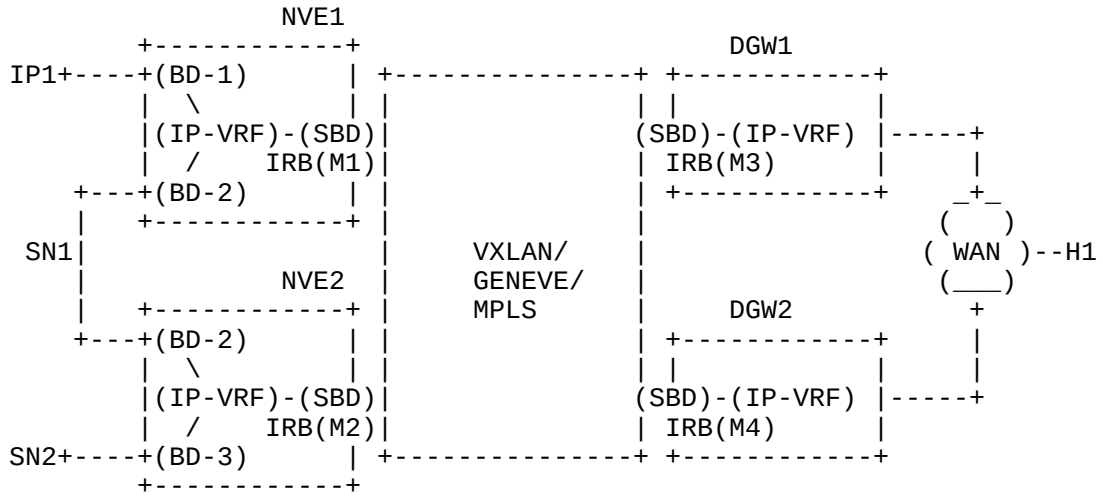

Figure 10 Interface-ful with unnumbered SBD IRB model

In this model:

a) As in Section 4.4.1 and 4.4.2, the NVEs and DGWs must provide connectivity between hosts in SN1, SN2, IP1 and hosts sitting at the other end of the WAN.

b) As in Section 4.4.2, the NVE/DGWs are connected through Ethernet NVO tunnels terminated in the SBD instance. The IP-VRFs use IRB interfaces for their connectivity to the SBD.

c) However, each SBD IRB has a MAC address only, and no IP address (that is why the model refers to an ’unnumbered’ SBD IRB). In this model, there is no need to have IP reachability to the SBD IRB interfaces themselves and there is a requirement to limit the number of IP addresses used.

d) As in Section 4.4.2, the SBD is composed of all the NVE/DGW BDs of the tenant that need inter-subnet-forwarding.

e) As in Section 4.4.2, the solution must provide layer 3 connectivity for Ethernet NVO tunnels, for instance, VXLAN or GENEVE (with Ethernet payload).

This model will also make use of the RT-5 recursive resolution. EVPN type 5 routes will advertise the IP Prefixes along with the Router’s MAC Extended Community used for the recursive lookup, whereas EVPN RT-2 routes will advertise the MAC addresses of each SBD IRB

interface (this time without an IP).

Each NVE/DGW will advertise an RT-5 for each of its prefixes with the same fields as described in 4.4.2 except for:

- o GW IP address= set to 0.

Each RT-5 will be sent with a Route Target identifying the tenant (IP-VRF) and the Router's MAC Extended Community containing the MAC address associated to SBD IRB interface. This MAC address may be reused for all the IP-VRFs in the NVE.

The example is similar to the one in Section 4.4.2:

(1) NVE1 advertises the following BGP routes:

- o Route type 5 (IP Prefix route) containing the same values as in the example in Section 4.4.2, except for:
  - . GW IP= SHOULD be set to 0.
  - . Router's MAC Extended Community containing M1 (this will be used for the recursive lookup to a RT-2).
- o Route type 2 (MAC route for the SBD IRB) with the same values as in Section 4.4.2 except for:
  - . ML=48, M=M1, IPL=0, Label=10.

(2) DGW1 imports the received routes from NVE1:

- o DGW1 installs SN1/24 in the IP-VRF identified by the RT-5 Route Target.
  - . The MAC contained in the Router's MAC Extended Community sent along with the RT-5 (M1) will be used as the Overlay Index for the recursive route resolution to the RT-2 carrying M1.

(3) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 IP-VRF routing table. The lookup yields SN1/24, which is associated to the Overlay Index M1. The forwarding information is derived from the RT-2 received for M1.
- o The IP packet destined to IPx is encapsulated with: Source

inner MAC = M3, Destination inner MAC = M1, Source outer IP (source VTEP) = DGW1 IP, Destination outer IP (destination VTEP) = NVE1 IP.

(4) When the packet arrives at NVE1:

- NVE1 will identify the IP-VRF for an IP lookup based on the Label and the inner MAC DA.
- An IP lookup is performed in the routing context, where SN1 turns out to be a local subnet associated to BD-2. A subsequent lookup in the ARP table and the BD FIB will provide the forwarding information for the packet in BD-2.

The model described above is called Interface-ful with unnumbered SBD IRB model (as in Section 4.4.2), only this time the SBD IRB does not have an IP address.

5. Security Considerations

This document provides a set of procedures to achieve Inter-Subnet Forwarding across NVEs or PEs attached to a group of BDs that belong to the same tenant (or VPN). The security considerations discussed in [RFC7432] apply to the Intra-Subnet Forwarding or communication within each of those BDs. In addition, the security considerations in [RFC4364] should also be understood, since this document and [RFC4364] may be used in similar applications.

Contrary to [RFC4364], this document does not describe PE/CE route distribution techniques, 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.

In this document, the RT-5 may use a regular BGP Next Hop for its resolution or an Overlay Index that requires a recursive resolution to a different EVPN route (an RT-2 or an RT-1). In the latter case, it is worth noting that any action that ends up filtering or modifying the RT-2/RT-1 routes used to convey the Overlay Indexes, will modify the resolution of the RT-5 and therefore the forwarding of packets to the remote subnet.

6. IANA Considerations

This document requests value 5 in the [EVPNRouteTypes] registry

defined by [RFC7432]:

Value	Description	Reference
5	IP Prefix route	[this document]

7. References

7.1 Normative References

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

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <http://www.rfc-editor.org/info/rfc5512>.

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <http://www.rfc-editor.org/info/rfc8174>.

[RFC8365] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", RFC 8365, DOI 10.17487/RFC8365, March, 2018.

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

[EVPNRouteTypes] IANA EVPN Route Type registry, https://www.iana.org/assignments/evpn

7.2 Informative References

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

[RFC7606] Chen, E., Scudder, J., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, August

2015, <http://www.rfc-editor.org/info/rfc7606>.

[802.1D-REV] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges", IEEE Std. 802.1D, June 2004.

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

[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>.

[RFC5227] Cheshire, S., "IPv4 Address Conflict Detection", RFC 5227, DOI 10.17487/RFC5227, July 2008, <https://www.rfc-editor.org/info/rfc5227>.

[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>.

[GENEVE] Gross, J., Ed., Ganga, I., Ed., and T. Sridhar, Ed., "Geneve: Generic Network Virtualization Encapsulation", Work in Progress, draft-ietf-nvo3-geneve-06, March 2018.

8. Acknowledgments

The authors would like to thank Mukul Katiyar and Jeffrey Zhang for their valuable feedback and contributions. The following people also helped improving this document with their feedback: Tony Przygienda and Thomas Morin. Special THANK YOU to Eric Rosen for his detailed review, it really helped improve the readability and clarify the concepts. Thank you to Alvaro Retana for his thorough review.

9. Contributors

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

Senthil Sathappan
Florin Balus
Aldrin Isaac
Senad Palislamovic

Samir Thoria

10. Authors' Addresses

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

Wim Henderickx
Nokia
Email: wim.henderickx@nokia.com

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

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Wen Lin
Juniper
Email: wlin@juniper.net