L2VPN Workgroup
Internet Draft

Intended status: Standards Track

J. Rabadan
W. Henderickx
S. Palislamovic
Alcatel-Lucent

J. Drake
Juniper

F. Balus
Nuage Networks

A. Sajassi
Cisco

A. Isaac
Bloomberg

Expires: April 19, 2015

October 16, 2014

IP Prefix Advertisement in EVPN
draft-rabadan-l2vpn-evpn-prefix-advertisement-03

Abstract

EVPN provides a flexible control plane that allows intra-subnet connectivity in an IP/MPLS and/or an NVO-based network. In NVO 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 may not support their own 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.

Table of Contents

1. Terminology

GW IP: Gateway IP Address

IPL: IP address length

IRB: Integrated Routing and Bridging interface

ML: MAC address length

NVE: Network Virtualization Edge

TS: Tenant System

VA: Virtual Appliance

RT-2: EVPN route type 2, i.e. MAC/IP advertisement route

RT-5: EVPN route type 5, i.e. IP Prefix route

Overlay next-hop: object used in the IP Prefix route, as described in this document. It can be an IP address in the tenant space or an ESI, and identifies the next-hop yielded by the IP route lookup at the routing context importing the route. An overlay next-hop always needs a recursive route resolution on the NVE receiving the IP Prefix route, so that the NVE knows to which egress NVE to forward the packets.

Underlay next-hop: IP address sent by BGP along with any EVPN route, i.e. BGP next-hop. It identifies the NVE sending the route and it is used at the receiving NVE as the VXLAN destination VTEP or NVGRE destination end-point.

2. Introduction and problem statement

Inter-subnet connectivity is required for certain tenants within the Data Center. [EVPN-INTERSUBNET] defines some fairly common inter-subnet forwarding scenarios where TSes can exchange packets with TSes located in remote subnets. In order to meet this requirement, [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 program individual host routes. This document complements the scenarios described in [EVPN-INTERSUBNET] and defines how EVPN may be used to advertise IP Prefixes.

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. Once the need for a new EVPN route type is justified, sections 3, 4 and 5 will describe this route type and how it is used in some specific use cases.

2.1 Inter-subnet connectivity requirements in Data Centers

[EVPN] is used as the control plane for a Network Virtualization Overlay (NVO3) solution in Data Centers (DC), where Network Virtualization Edge (NVE) devices can be located in Hypervisors or TORs, as described in [EVPN-OVERLAYS].

If we use the term Tenant System (TS) to designate a physical or virtual system identified by MAC and IP addresses, and connected to an EVPN instance, the following considerations apply:

- 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 seating behind them.

    - These VAs can be firewalls, load balancers, NAT devices, other appliances or virtual gateways with virtual routing instances.

    - These VAs do not have their own 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 the Data Center using its own source MAC but the source IP will be the one associated to the End Device seating behind or a translated IP address (part of a public NAT pool) if the VA is performing NAT.

    - Note that the same IP address could exist behind two of these TS. 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). VRRP 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.

The following figure illustrates some of the examples described above.

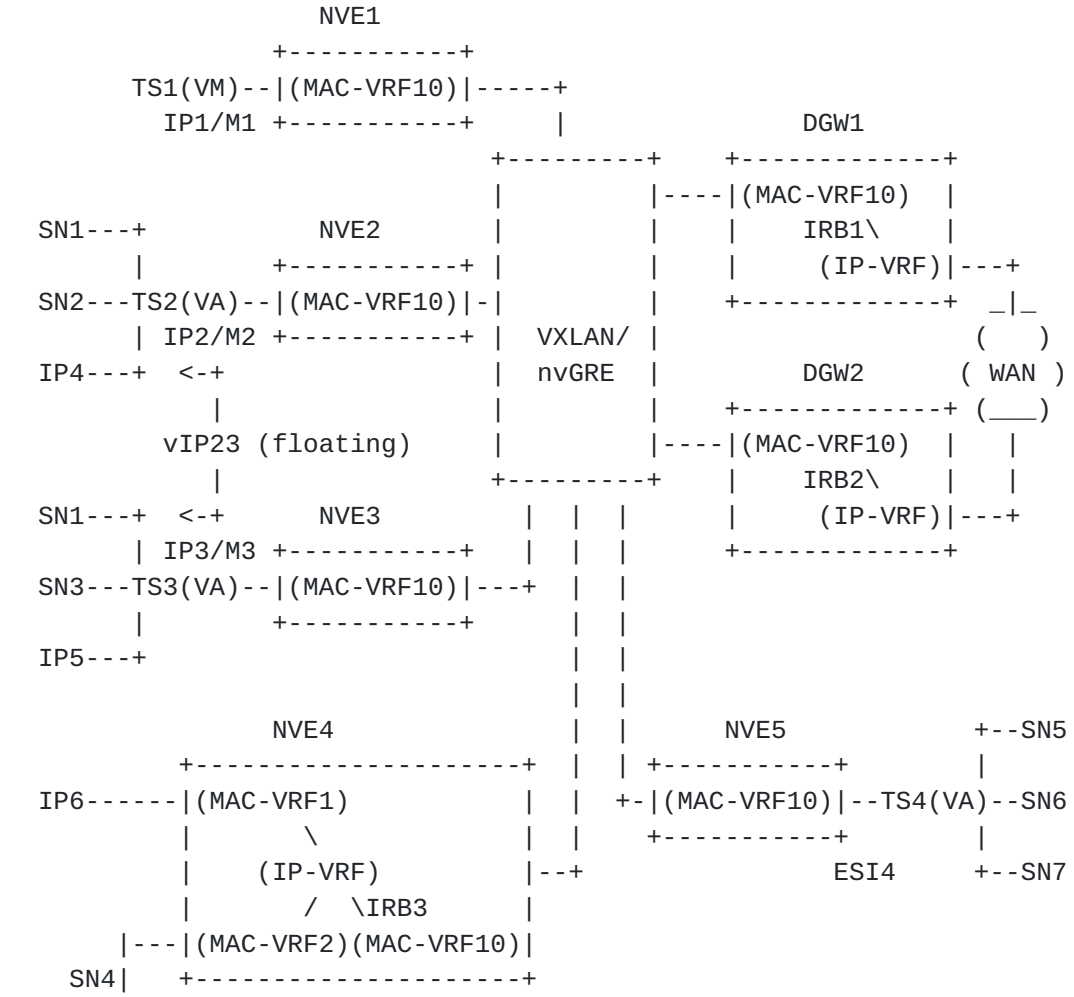


Figure 1 DC inter-subnet use-cases

Where:

NVE1, NVE2, NVE3, NVE4, NVE5, DGW1 and DGW2 share the same EVI for a particular tenant. EVI-10 is comprised of the collection of MAC-VRF10 instances defined in all the NVEs. All the hosts connected to EVI-10 belong to the same IP subnet. The hosts connected to EVI-10 are listed below:

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

- TS2 and TS3 are Virtual Appliances (VA) that generate/receive traffic from/to the subnets and hosts seating behind them (SN1, SN2, SN3, IP4 and IP5). Their IP addresses (IP2 and IP3) belong to the EVI-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 who the master is. 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 EVI-10 subnet too. These IRB interfaces connect the EVI-10 subnet to Virtual Routing and Forwarding (VRF) instances that can route the traffic to other connected 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 EVI-10. TS4 is connected to a physical port on NVE5 assigned to Ethernet Segment Identifier 4.

All the above DC use cases 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 run routing protocols) and
b) MAY be associated to an overlay next-hop that can be a VA IP address, a floating IP address or an ESI.

2.2 The requirement for a new EVPN route type

[EVPN] defines a MAC/IP route (also referred as RT-2) where a MAC address can be advertised together with an IP address length (IPL) and IP address (IP). While a variable IPL 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 we need to decouple the advertisement of the prefixes from the advertisement of the floating IP (vIP23 in figure 1) and MAC associated to it, otherwise the solution gets highly inefficient and does not scale.

E.g.: if we are advertising 1k prefixes from M2 (using RT-2) and the

floating IP owner changes from M2 to M3, we would need to withdraw 1k routes from M2 and re-advertise 1k routes from M3. However if we use a separate route type, we can advertise the 1k routes 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 1k prefixes associated to vIP23, but will only update the ARP resolution entry for vIP23 (now pointing at M3).

Other reasons to decouple the IP Prefix advertisement from the MAC/IP route are listed below:

- Clean identification, operation of troubleshooting of IP Prefixes, not subject to interpretation and independent of the IPL and the IP value. E.g.: a default IP route 0.0.0.0/0 must always be easily and clearly distinguished from the absence of IP information.

- MAC address information must not be compared by BGP when selecting two IP Prefix routes. If IP Prefixes were to be advertised using MAC/IP routes, the MAC information would always be present and part of the route key.

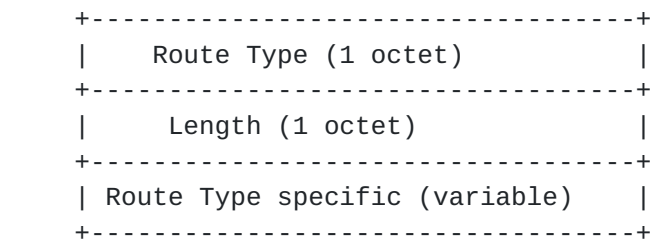
- IP Prefix routes must not be subject to MAC/IP route procedures such as MAC mobility or aliasing. Prefixes advertised from two different ESIs do not mean mobility; MACs advertised from two different ESIs do mean mobility. Similarly load balancing for IP prefixes is achieved through IP mechanisms such as ECMP, and not through MAC route mechanisms such as aliasing.

- NVEs that do not require processing IP Prefixes must have an easy way to identify an update with an IP Prefix and ignore it, rather than processing the MAC/IP route to find out only later that it carries a Prefix that must be ignored.

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

3. The BGP EVPN IP Prefix route

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



Where the route type field can contain one of the following specific values:

+ 1 - Ethernet Auto-Discovery (A-D) route

+ 2 - MAC/IP advertisement route

+ 3 - Inclusive Multicast Route

+ 4 - Ethernet Segment Route

This document defines an additional route type that will be used for the advertisement of IP Prefixes:

+ 5 - IP Prefix Route

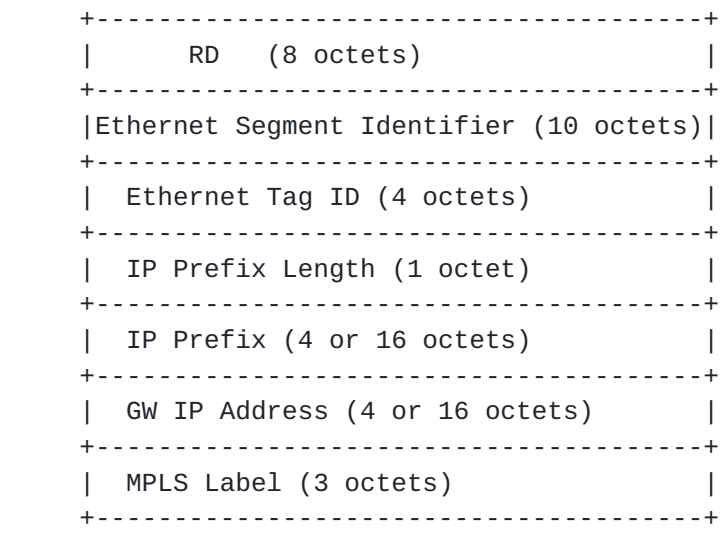
The support for this new route type is OPTIONAL.

Since this new route type is OPTIONAL, an implementation not supporting it MUST ignore the route, based on the unknown route type value.

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

3.1 IP Prefix Route encoding

An IP Prefix advertisement route NLRI consists of the following fields:



Where:

- RD, Ethernet Tag ID and MPLS Label fields will be used as defined in [EVPN] and [EVPN-OVERLAYS].

- The Ethernet Segment Identifier will be a non-zero 10-byte identifier if the ESI is used as an overlay next-hop. It will be zero otherwise.

- The IP Prefix Length can be set to a value between 0 and 32 (bits) for ipv4 and between 0 and 128 for ipv6.

- The IP Prefix will be a 32 or 128-bit field (ipv4 or ipv6).

- The GW IP (Gateway IP Address) will be a 32 or 128-bit field (ipv4 or ipv6), and will encode the overlay IP next-hop for the IP Prefixes. The GW IP field can be zero if it is not used as an overlay next-hop.

- The total route length will indicate the type of prefix (ipv4 or ipv6) and the type of GW IP address (ipv4 or ipv6). Note that the IP Prefix + the GW IP should have a length of either 64 or 256 bits, but never 160 bits (ipv4 and ipv6 mixed values are not allowed).

The Eth-Tag ID, IP Prefix Length and IP Prefix will be part of the route key used by BGP to compare routes. The rest of the fields will not be part of the route key.

The route will contain a single overlay next-hop at most, i.e. if the ESI field is different from zero, the GW IP field will be zero, and vice versa. The following table shows the different inter-subnet use-

cases described in this document and the corresponding coding of the overlay next-hop in the route type 5 (RT-5). The IP-VRF-to-IP-VRF or IRB forwarding on NVEs case is a special use-case, where there is no need for overlay next-hop, since the actual next-hop is given by the BGP next-hop. When an overlay next-hop is present in the RT-5, the receiving NVE will need to perform a recursive route resolution to find out to which egress NVE to forward the packets.

Use-case	Next-hop in the RT-5 BGP update
TS IP address	GW IP Address
Floating IP address	GW IP Address
"Bump in the wire"	ESI
IP-VRF-to-IP-VRF	BGP next-hop

4. Benefits of using the EVPN IP Prefix route

This section clarifies the different functions accomplished by the EVPN RT-2 and RT-5 routes, and provides a list of benefits derived from using a separate route type for the advertisement of IP Prefixes in EVPN.

[EVPN] describes the content of the BGP EVPN RT-2 specific NLRI, i.e. MAC/IP Advertisement Route, where the IP address length (IPL) and IP address (IP) of a specific advertised MAC are encoded. The subject of the MAC advertisement route is the MAC address (M) and MAC address length (ML) encoded in the route. The MAC mobility and other complex procedures are defined around that MAC address. The IP address information carries the host IP address required for the ARP resolution of the MAC according to [EVPN] and the host route to be programmed in the IP-VRF [EVPN-INTERSUBNET].

The BGP EVPN route type 5 defined in this document, i.e. IP Prefix Advertisement route, decouples the advertisement of IP prefixes from the advertisement of any MAC address related to it. This brings some major benefits to NVO-based networks where certain inter-subnet forwarding scenarios are required. Some of those benefits are:

a) Upon receiving a route type 2 or type 5, an egress NVE can easily distinguish MACs and IPs from IP Prefixes. E.g. an IP prefix with IPL=32 being advertised from two different ingress NVEs (as RT-5) can be identified as such and be imported in the designated routing context as two ECMP routes, as opposed to two MACs competing for the same IP.

b) Similarly, upon receiving a route, an ingress NVE not supporting

processing of IP Prefixes can easily ignore the update, based on the route type.

c) A MAC route includes the ML, M, IPL and IP in the route key that is used by BGP to compare routes, whereas for IP Prefix routes, only IPL and IP (as well as Ethernet Tag ID) are part of the route key. Advertised IP Prefixes are imported into the designated routing context, where there is no MAC information associated to IP routes. In the example illustrated in figure 1, subnet SN1 should be advertised by NVE2 and NVE3 and interpreted by DGW1 as the same route coming from two different next-hops, regardless of the MAC address associated to TS2 or TS3. This is easily accomplished in the RT-5 by including only the IP information in the route key.

d) By decoupling the MAC from the IP Prefix advertisement procedures, we can leave the IP Prefix advertisements out of the MAC mobility procedures defined in [EVPN] for MACs. In addition, this allows us to have an indirection mechanism for IP Prefixes advertised from a MAC/IP that can move between hypervisors. E.g. if there are 1,000 prefixes seating behind TS2 (figure 1), NVE2 will advertise all those prefixes in RT-5 routes associated to the next-hop IP2. Should TS2 move to a different NVE, a single MAC advertisement route withdraw for the M2/IP2 route from NVE2 will invalidate the 1,000 prefixes, as opposed to have to wait for each individual prefix to be withdrawn. This may be easily accomplished by using IP Prefix routes that are not tied to a MAC address, and use a different MAC/IP route to advertise the location and resolution of the overlay next-hop to a MAC address.

5. IP Prefix next-hop use-cases

The IP Prefix route can use a GW IP or an ESI as an overlay next-hop as well as no overlay next-hop whatsoever. This section describes some use-cases for these next-hop types.

5.1 TS IP address next-hop use-case

The following figure illustrates an example of inter-subnet forwarding for subnets seating behind Virtual Appliances (on TS2 and TS3).

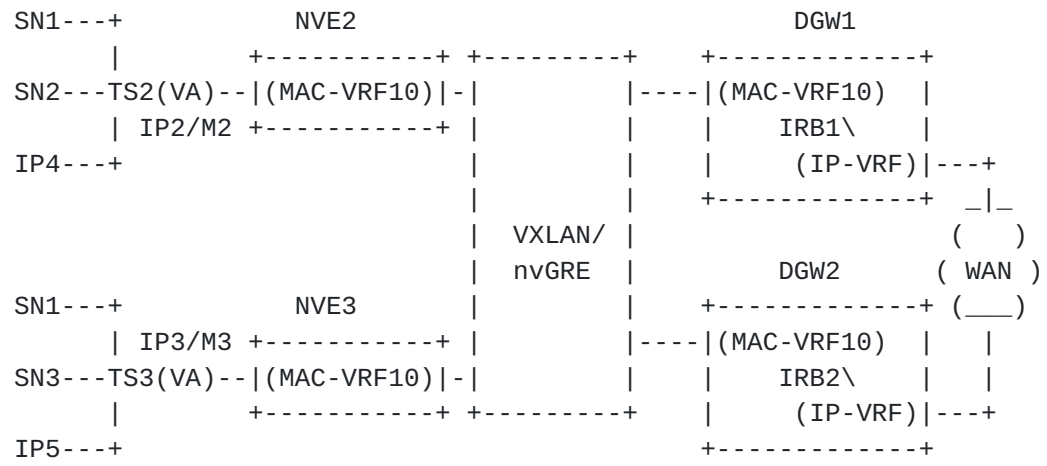


Figure 2 TS IP address use-case

An example of inter-subnet forwarding between subnet SN1/24 and a subnet seating in the WAN is described below. NVE2, NVE3, DGW1 and DGW2 are running BGP EVPN. TS2 and TS3 do not support routing protocols, only a static route to forward the traffic to the WAN.

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

- Route type 2 (MAC/IP route) containing: ML=48, M=M2, IPL=32, IP=IP2 and [RFC5512] BGP Encapsulation Extended Community with Tunnel-type= VXLAN or NVGRE.

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP2 (and BGP Encapsulation Extended Community).

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

- Route type 2 (MAC/IP route) containing: ML=48, M=M3, IPL=32, IP=IP3 (and BGP Encapsulation Extended Community).

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP3 (and BGP Encapsulation Extended Community).

(3) DGW1 and DGW2 import both received routes based on the route-targets:

- Based on the MAC-VRF10 route-target in DGW1 and DGW2, the MAC/IP route is imported and M2 is added to the MAC-VRF10 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 (underlay next-hop) and VNI from the

Ethernet Tag or MPLS fields. IP2 - M2 is added to the ARP table.

- Based on the MAC-VRF10 route-target in DGW1 and DGW2, the IP Prefix route is also imported and SN1/24 is added to the designated routing context with next-hop IP2 pointing at the local MAC-VRF10. Should ECMP be enabled in the routing context, SN1/24 would also be added to the routing table with next-hop IP3.

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

- A destination IP lookup is performed on the DGW1 IP-VRF routing table and next-hop=IP2 is found. Since IP2 is an overlay next-hop a recursive route resolution is required for IP2.

- IP2 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC FIB (remote VTEP and VNI for the VXLAN case).

- The IP packet destined to IPx is encapsulated with:

  . Source inner MAC = IRB1 MAC

  . Destination inner MAC = M2

  . Tunnel information provided by the MAC-VRF (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 MAC-VRF10 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 [EVPN]. Route type 5 prefixes are not subject to MAC mobility procedures, hence no changes in the DGW VRF routing table will occur for TS2 mobility, i.e. all the prefixes will still be pointing at IP2 as next-hop. There is an indirection for e.g. SN1/24, which still points at next-hop 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.

5.2 Floating IP next-hop 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 next-hop to get to some subnets behind. This redundancy mode, already introduced in section 2.1 and 2.2, is illustrated in Figure 3.

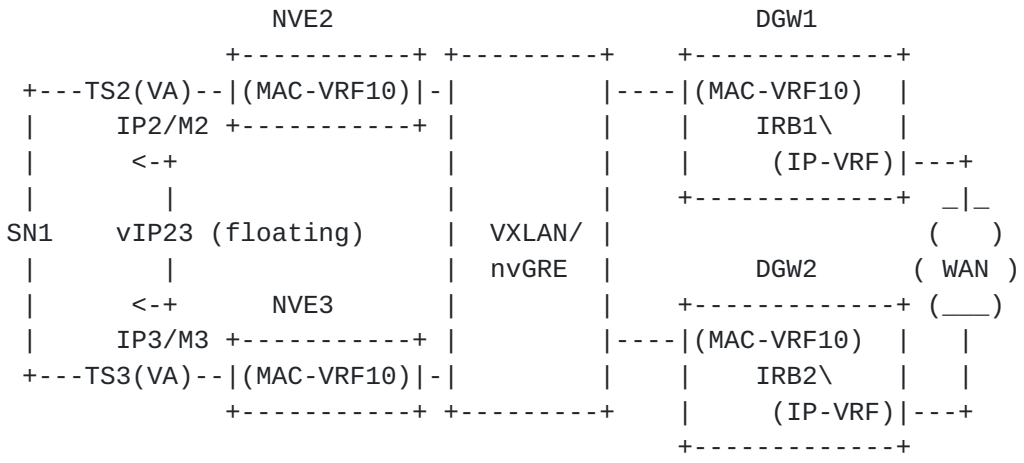


Figure 3 Floating IP next-hop for redundant TS

In this example, assuming TS2 is the active TS and owns IP23:

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

- Route type 2 (MAC/IP route) containing: ML=48, M=M2, IPL=32, IP=IP23 (and BGP Encapsulation Extended Community).

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP23 (and BGP Encapsulation Extended Community).

(2) NVE3 advertises the following BGP routes for TS3:

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP23 (and BGP Encapsulation Extended Community).

(3) DGW1 and DGW2 import both received routes based on the route-target:

- o M2 is added to the MAC-VRF10 MAC 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 Ethernet Tag or MPLS fields. IP23 - M2 is added to the ARP table.

- o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop IP23 pointing at the local MAC-VRF10.

(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 next-hop=IP23 is found. Since IP23 is an overlay next-hop, a recursive route resolution for IP23 is required.

- o IP23 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC-VRF (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 MAC 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 MAC-VRF10 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 IP23 and will signal this new ownership (GARP message or similar). Upon receiving the new owner's notification, NVE3 will issue a route type 2 for M3-IP23. DGW1 and DGW2 will update their ARP tables with the new MAC resolving the floating IP. No changes are carried out in the VRF routing table.

In the DGW1/2 BGP RIB, there will be two route type 5 routes for SN1

(from NVE2 and NVE3) but only the one with the same BGP next-hop as the IP23 RT-2 BGP next-hop will be valid.

5.3 ESI next-hop ("Bump in the wire") use-case

The following figure illustrates and example of inter-subnet forwarding for a subnet route that uses an ESI as an overlay next-hop. In this use-case, TS2 and TS3 are layer-2 VA devices without any IP address that can be included as an overlay next-hop in the GW IP field of the IP Prefix route.

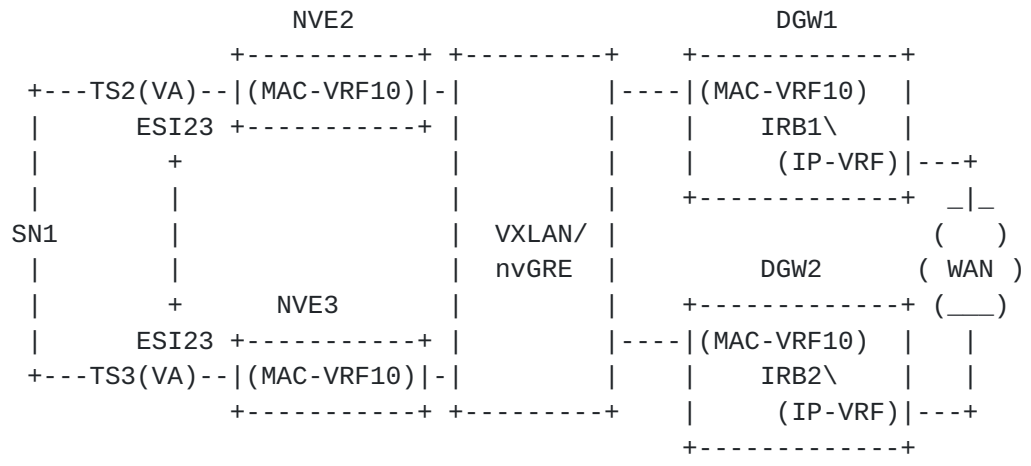


Figure 5 ESI next-hop use-case

Since neither TS2 nor TS3 can run any routing protocol and have no IP address assigned, an ESI, i.e. ESI23, will be provisioned on the attachment ports of NVE2 and NVE3. This model supports VA redundancy in a similar way as the one described in section 5.2 for the floating IP next-hop use-case, only using the EVPN Ethernet A-D route instead of the MAC advertisement route to advertise the location of the overlay next-hop. The procedure is explained below:

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

- Route type 1 (Ethernet A-D route for EVI-10) containing: ESI=ESI23 and the corresponding tunnel information (Ethernet Tag and/or MPLS label), as well as the BGP Encapsulation Extended Community. Assuming the ESI is active on NVE2, NVE2 will advertise this route.

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=ESI23, GW IP address=0 (and BGP Encapsulation Extended Community).

(2) NVE3 advertises the following BGP routes for TS3:

- Route type 1 (Ethernet A-D route for EVI-10) containing: ESI=ESI23 and the corresponding tunnel information (Ethernet Tag and/or MPLS label), as well as the BGP Encapsulation Extended Community. NVE3 will advertise this route assuming the ESI is active on NVE2. Note that if the resiliency mechanism for TS2 and TS3 is in active-active mode, both NVE2 and NVE3 will send the A-D route. Otherwise, that is, the resiliency is active-standby, only the NVE owning the active ESI will advertise the Ethernet A-D route for ESI23.

- Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=23, GW IP address=0 (and BGP Encapsulation Extended Community).

(3) DGW1 and DGW2 import the received routes based on the route-target:

- 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 Ethernet Tag or MPLS fields (see [EVPN-OVERLAYS]).

- SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop ESI23 pointing at the local MAC-VRF10.

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

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

- The IP packet destined to IPx is encapsulated with:

  . Source inner MAC = IRB1 MAC

  . Destination inner MAC = M2 (this MAC will be obtained after a lookup in the IP-VRF ARP table or in the MAC-VRF10 FDB table associated to ESI23).

  . Tunnel information provided by the Ethernet A-D route for ESI23 (VNI, VTEP IP 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 MAC-VRF10 context is identified for a MAC lookup (assuming MAC disposition model).

- 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 forwarded.

(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 will 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. No changes are carried out in the IP-VRF routing table.

In the DGW1/2 BGP RIB, there will be two route type 5 routes for SN1 (from NVE2 and NVE3) but only the one with the same BGP next-hop as the ESI23 route type 1 BGP next-hop will be valid.

5.4 IRB forwarding on NVEs for Subnets (IP-VRF-to-IP-VRF)

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 previous examples, the MAC-VRF instance can connect IRB interfaces and any other Tenant Systems connected to it. EVPN provides connectivity for:

a) Traffic destined to the IRB IP interfaces as well as

b) Traffic destined to IP subnets seating behind the TS, e.g. SN1 or SN2.

In order to provide connectivity for (a) we need MAC/IP routes (RT-2) distributing IRB MACs and IPs. Connectivity type (b) is accomplished by the exchange of IP Prefix routes (RT-5) for IPs and subnets seating behind certain overlay next-hops.

In some cases, subnets may be advertised in IP Prefix routes without any overlay next-hop since the RT-5 itself provides all the forwarding information required to send the packets to the egress NVE and no recursive route resolution is needed. This use case is depicted in the diagram below and we refer to it as the "IRB forwarding on NVEs for Subnets" or "IP-VRF-to-IP-VRF" use-case:

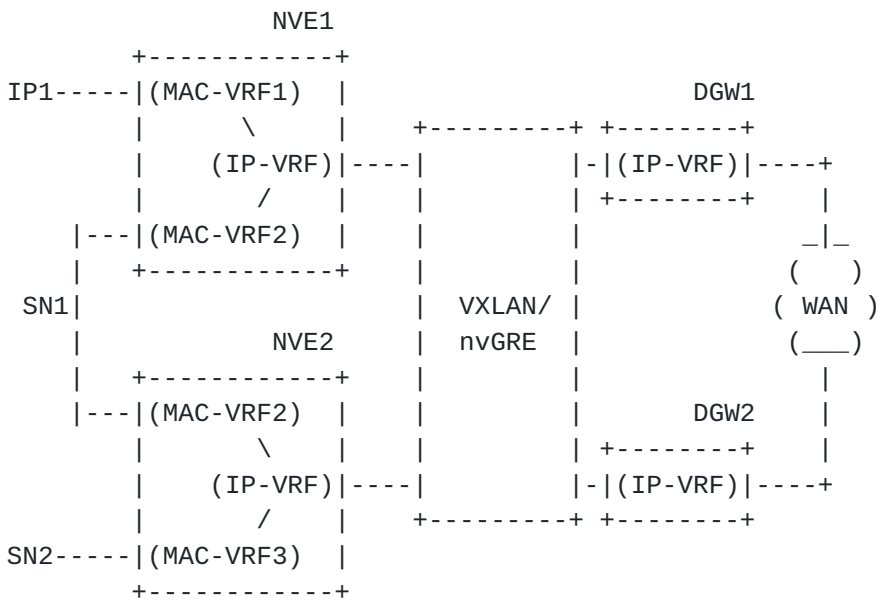


Figure 6 Inter-subnet forwarding on NVEs for Subnets

In this case, we need to provide connectivity from/to IP hosts in SN1, SN2, IP1 and hosts seating at the other end of the WAN. There is no need to define IRB interfaces to interconnect the IP-VRF instances among the NVEs for the same tenant. This is the reason why we refer to this solution as "IP-VRF-to-IP-VRF" solution.

In this case, 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]. Each NVE/DGW will advertise an RT-5 for each of its subnet prefixes with the following fields:

- RD as per [EVPN].
- Eth-Tag ID = 0 assuming VLAN-based service.
- IP address length and IP address, as explained in the previous sections.
- GW IP address=0 and ESI=0, that is, no overlay next-hop is required in this use-case, since the BGP next-hop is enough to find the egress NVE to forward the packets to.
- MPLS label or VNID corresponding to the IP-VRF.

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

- The first one is the BGP Encapsulation Extended Community, as per [RFC5512], identifying the tunnel type.

- 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 re-used for all the IP-VRFs in the node. The ingress NVE will use this MAC address as the inner MAC destination address in the packets forwarded to the owner of the RT-5.

Example of prefix advertisement for the ipv4 prefix SN1/24 advertised from NVE1:

(1) NVE1 advertises the following BGP route for SN1:

- Route type 5 (IP Prefix route) containing: Eth-Tag=0, IPL=24, IP=SN1, MPLS Label=10. An [RFC5512] BGP Encapsulation Extended Community will be sent, where Tunnel-type= VXLAN or NVGRE. A Router's MAC Extended Community will also be sent along with the RT-5, where the Router's MAC address value will contain the NVE1 MAC.

(2) DGW1 imports the received route from NVE1 and SN1/24 is added to the designated IP-VRF. The next-hop for SN1/24 will be given by the route type 5 BGP next-hop (NVE1), which is resolved to a tunnel. For instance: if the tunnel is VXLAN based, the BGP next-hop will be resolved to a VXLAN tunnel where: destination-VTEP= NVE1 IP, VNI=10, inner destination MAC = NVE1 MAC (derived from the Router's MAC Extended Community value).

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

- A destination IP lookup is performed on the DGW1 IP-VRF routing table and next-hop= "NVE1 IP" is found. The tunnel information to encapsulate the packet will be derived from the route type 5 received for SN1.

- The IP packet destined to IPx is encapsulated with: Source inner MAC = DGW1 MAC, Destination inner MAC = NVE1 MAC, Source outer IP (source VTEP) = DGW1 IP, Destination outer IP (destination VTEP) = NVE1 IP.

(4) When the packet arrives at NVE1:

- Based on the tunnel information (VNI for the VXLAN case), the routing context is identified for an IP lookup.

- o An IP lookup is performed in the routing context, where SN1 turns out to be a local subnet associated to MAC-VRF2. A subsequent lookup in the ARP table and the MAC-VRF FIB will return the forwarding information for the packet in EVI-2.

6. Conclusions

A new 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 all the Data Center (or NVO-based networks in general) inter-subnet connectivity scenarios in which an IP Prefix advertisement is required. Using this new RT-5, an IP Prefix may be advertised along with an overlay next-hop that can be a GW IP address or an ESI, or without an overlay next-hop, in which case the BGP next-hop will point at the egress NVE 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 existing EVPN RT-2 does not meet the requirements for all the DC use cases, therefore a new EVPN route type is required.

This new EVPN route type 5 decouples the IP Prefix advertisements from the MAC route advertisements in EVPN, hence:

a) Allows the clean and clear advertisements of ipv4 or ipv6 prefixes in an NLRI with no MAC addresses in the route key, so that only IP information is used in BGP route comparisons.

b) Since the route type is different from the MAC/IP advertisement route, the advertisement of prefixes will be excluded from all the procedures defined for the advertisement of VM MACs, e.g. MAC Mobility or aliasing. As a result of that, the current EVPN procedures do not need to be modified.

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

d) An EVPN implementation not requiring IP Prefixes can simply discard them by looking at the route type value. An unknown route type MUST be ignored by the receiving NVE/PE.

7. Conventions used in this document

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

8. Security Considerations

9. IANA Considerations

10. References

10.1 Normative References

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

10.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-11.txt, work in progress, October, 2014

[EVPN-OVERLAYS] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", draft-sd-l2vpn-evpn-overlay-03.txt, work in progress, June, 2014

[EVPN-INTERSUBNET] Sajassi et al., "IP Inter-Subnet Forwarding in EVPN", draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-05.txt, work in progress, October, 2014

11. Acknowledgments

The authors would like to thank Mukul Katiyar and Senthil Sathappan for their valuable feedback and contributions. The following people also helped improving this document with their feedback: Antoni Przygienda and Thomas Morin.

12. Authors' Addresses

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Florin Balus

Nuage Networks
Email: florin@nuagenetworks.net

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Senad Palislamovic
Alcatel-Lucent
Email: senad.palislamovic@alcatel-lucent.com

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

Ali Sajassi
Cisco
Email: sajassi@cisco.com