                    IP Prefix Advertisement in EVPN
              draft-ietf-bess-evpn-prefix-advertisement-08

Abstract

   EVPN provides a flexible control plane that allows intra-subnet
   connectivity in an MPLS and/or NVO-based 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.

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

This Internet-Draft will expire on April 23, 2018.

Copyright Notice

Table of Contents

[1](#). Terminology

GW IP: Gateway IP Address.

IPL: IP address length.

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.

AC: Attachment Circuit.

ARP: Address Resolution Protocol.

ND: Neighbor Discovery Protocol.

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

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

EVI: EVPN Instance spanning the NVE/PE devices that are participating
   on that EVPN.

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

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.

BT: Bridge Table. The instantiation of a BD in a MAC-VRF.

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.

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

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.).

2. **Introduction and Problem Statement**

Inter-subnet connectivity is used 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 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 program 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. 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**

[RFC7432] 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-OVERLAY].

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

o The Tenant Systems may be Virtual Machines (VMs) that generate
  traffic from their own MAC and IP.

o The Tenant Systems may be Virtual Appliance entities (VAs) that
  forward traffic to/from IP addresses of different End Devices

sitting behind them.

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

   o These VAs do not necessarily participate in dynamic routing
     protocols and hence rely on the EVPN NVEs to advertise the
     routes on their behalf.

   o 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.

   o 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). 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.

   o 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.

```
                        NVE1
                 +-----------+
        TS1(VM)--|  (BD-10)  |-----+
          IP1/M1 +-----------+     |                   DGW1
                                +---------+    +-------------+
                                |         |----|  (BD-10)    |
    SN1---+          NVE2       |         |    |    IRB1\     |
      |         +-----------+   |         |    |    (IP-VRF)|---+
    SN2---TS2(VA)--|  (BD-10)  |-|         |    +-------------+  _|_
      | IP2/M2 +-----------+  | VXLAN/ |              (   )
    IP4---+   <-+             |  nvGRE  |       DGW2        ( WAN )
          |                   |         |    +-------------+ (___)
        vIP23 (floating)      |         |----|  (BD-10)    |   |
          |                   +---------+    |    IRB2\     |   |
    SN1---+  <-+       NVE3           |  |  |    |    (IP-VRF)|---+
      | IP3/M3 +-----------+    |  |  |    +-------------+
    SN3---TS3(VA)--|  (BD-10)  |---+  |  |
      |         +-----------+     |  |
    IP5---+                       |  |
                                  |  |
              NVE4                |  |        NVE5        +--SN5
        +-------------------+     |  | +-----------+        |
    IP6------|  (BD-1)       |     |  | +-|  (BD-10)  |--TS4(VA)--SN6
        |        \          |     |  |   +-----------+        |
        |     (IP-VRF)      |--+            ESI4       +--SN7
        |     /  \IRB3      |
      |---|  (BD-2)  (BD-10)   |
    SN4|   +-------------------+
```
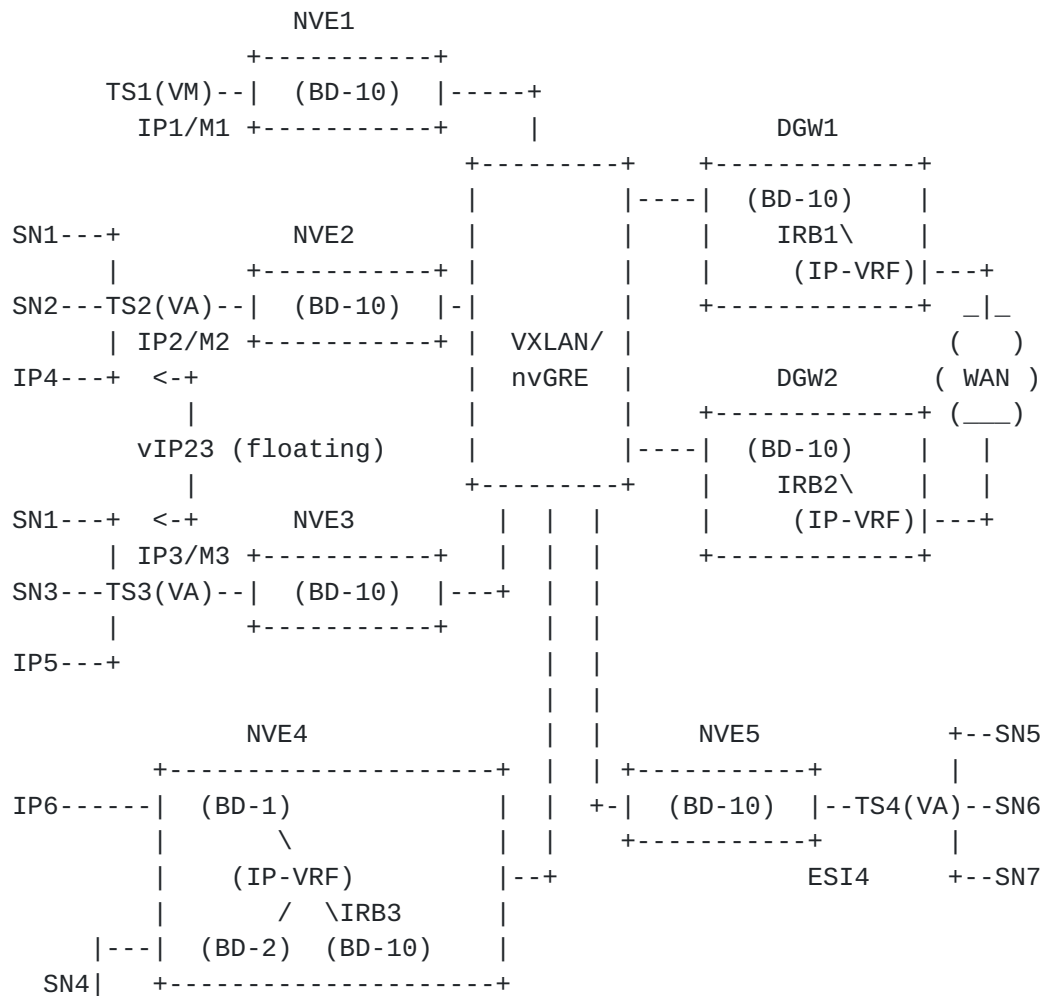
                  Figure 1 DC inter-subnet use-cases

   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 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 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 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
   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. An Overlay Index
   is a next-hop that requires a recursive resolution and it is
   described in section 3.2.

2.2 The Requirement for a New EVPN Route Type

[RFC7432] 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 MAC address of either M2 or
M3", 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:

o Clean identification, operation and troubleshooting of IP Prefixes,
  independent of and not subject to the interpretation 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.

o In MAC/IP routes, the MAC information is part of the NLRI, so if IP
  Prefixes were to be advertised using MAC/IP routes, the MAC
  information would always be present and part of the route key.

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 [RFC7432] is shown below:

```
+----------------------------------+
|      Route Type (1 octet)        |
+----------------------------------+
|       Length (1 octet)           |
+----------------------------------+
| Route Type specific (variable)   |
+----------------------------------+
```

Where the route type field can contain one of the following specific
values (refer to the IANA "EVPN Route Types registry):

+ 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 IANA has added to the registry, and 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, as specified by Section 5.4 in [RFC7606].

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:

```
+---------------------------------------+
|      RD   (8 octets)                  |
+---------------------------------------+
|Ethernet Segment Identifier (10 octets)|
+---------------------------------------+
|  Ethernet Tag ID (4 octets)           |
+---------------------------------------+
|  IP Prefix Length (1 octet)           |
+---------------------------------------+
|  IP Prefix (4 or 16 octets)           |
+---------------------------------------+
|  GW IP Address (4 or 16 octets)       |
+---------------------------------------+
|  MPLS Label (3 octets)                |
+---------------------------------------+
```

Where:

o RD, Ethernet Tag ID and MPLS Label fields will be used as defined in [RFC7432] and [EVPN-OVERLAY].

o The Ethernet Segment Identifier will be a non-zero 10-byte identifier if the ESI is used as an Overlay Index (see the definition of Overlay Index in section 3.2). It will be zero

   otherwise.

   o 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.

   o The IP Prefix will be a 32 or 128-bit field (ipv4 or ipv6). The
     size of this field does not depend on the value of the IP Prefix
     Length field.

   o The GW IP (Gateway IP Address) will be a 32 or 128-bit field (ipv4
     or ipv6), and will encode an IP address as an overlay index for the
     IP Prefixes. The GW IP field SHOULD be zero if it is not used as an
     Overlay Index. Refer to section 3.2 for the definition and use of
     the Overlay Index.

   o The MPLS Label field is encoded as 3 octets, where the high-order
     20 bits contain the label value. 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 value
     is non-zero, the route will not be used for recursive resolution
     unless a local policy says so.

   o 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 RD, 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.

   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.

3.2 Overlay Indexes and Recursive Lookup Resolution

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

   o 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.

o 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).

o 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.

o Irrespective of the recursive resolution, if there is no IGP or BGP
  route to the BGP next-hop of an RT-5, BGP SHOULD fail to install
  the RT-5 even if the Overlay Index can be resolved.

o The ESI and GW IP fields MAY both be zero, 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) will be treated 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. When the Overlay Index is "None" in Table 1, 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. Examples of
     invalid MAC addresses are broadcast or multicast MAC addresses.
     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.

*** 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.

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). Any other use-
   case using a given Overlay Index, SHOULD follow the procedures
   described in this document for the same Overlay Index.

4. Overlay Index Use-Cases

   This section describes some use-cases for the Overlay Index types
   used with the IP Prefix route.

4.1 TS IP Address Overlay Index Use-Case

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

```
IP4---+            NVE2                            DGW1
      |        +-----------+ +---------+     +-------------+
SN2---TS2(VA)--|  (BD-10)  |-|         |----|  (BD-10)     |
      | IP2/M2 +-----------+ |         |    |    IRB1\     |
  -+---+                     |         |    |     (IP-VRF)|---+
   |                         |         |    +-------------+  _|_
  SN1                        | VXLAN/  |                    (   )
   |                         | nvGRE   |         DGW2      ( WAN )
  -+---+            NVE3      |         |    +-------------+ (___)
      | IP3/M3 +-----------+ |         |----|  (BD-10)     |   |
SN3---TS3(VA)--|  (BD-10)  |-|         |    |    IRB2\     |   |
      |        +-----------+ +---------+    |     (IP-VRF)|---+
IP5---+                                     +-------------+
```
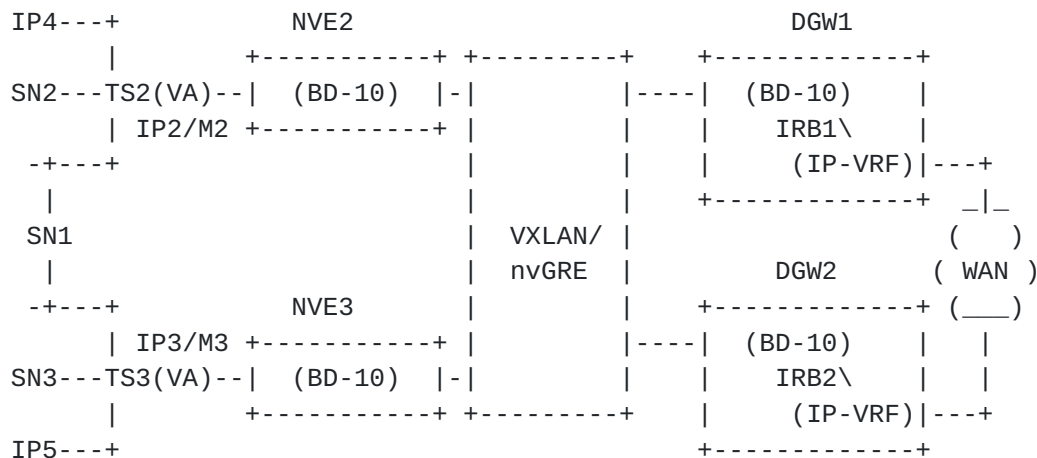
                     Figure 2 TS IP address use-case

   An example of inter-subnet forwarding between subnet SN1/24 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. We assume 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:

   o Route type 2 (MAC/IP route) containing: ML=48, M=M2, IPL=32,
     IP=IP2 and [RFC5512] BGP Encapsulation Extended Community with
     the corresponding Tunnel-type. The MAC and IP addresses may be
     learned via ARP-snooping (ND-snooping if IPv6).

   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, we assume 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:

    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) 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
3.

```
              NVE2                          DGW1
         +-----------+ +---------+    +-------------+
  +---TS2(VA)--|  (BD-10)  |-|         |----|  (BD-10)    |
  |     IP2/M2 +-----------+ |         |    |    IRB1\    |
  |       <-+                |         |    |    (IP-VRF)|---+
  |        |                 |         |    +-------------+  _|_
SN1    vIP23 (floating)      | VXLAN/  |                    (   )
  |        |                 | nvGRE   |         DGW2      ( WAN )
  |       <-+      NVE3       |         |    +-------------+ (___)
  |     IP3/M3 +-----------+ |         |----|  (BD-10)    |   |
  +---TS3(VA)--|  (BD-10)  |-|         |    |    IRB2\    |   |
              +-----------+ +---------+    |    (IP-VRF)|---+
                                          +-------------+
```
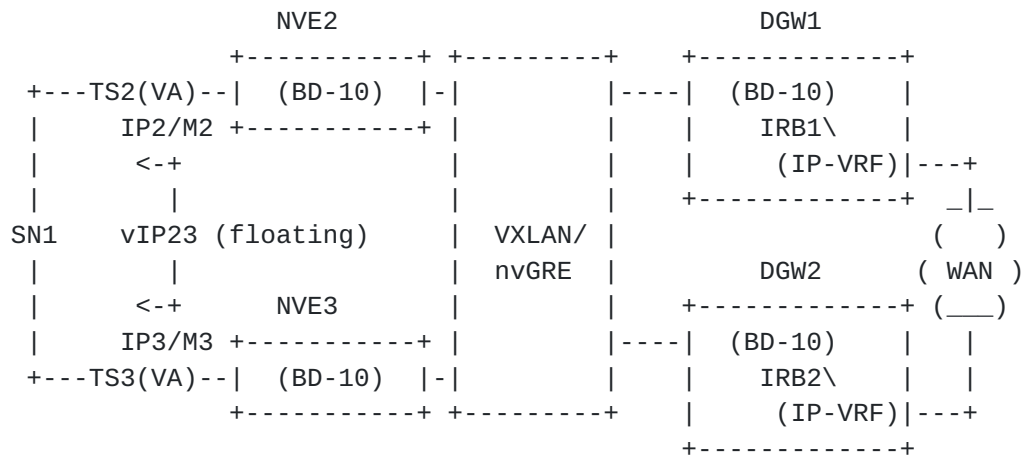
           Figure 3 Floating IP Overlay Index for redundant TS

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 IP23:

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

    o Route type 2 (MAC/IP route) containing: ML=48, M=M2, IPL=32,
      IP=IP23 (and BGP Encapsulation Extended Community). 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=IP23. 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 IP23/M3):

    o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1,
      ESI=0, GW IP address=IP23. The prefix and GW IP are learned by
      policy.

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

    o 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/VSID field. IP23 - M2 is added to the ARP table.

   o SN1/24 is added to the IP-VRF in DGW1 and DGW2 with Overlay
     index IP23 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=IP23 is found. Since IP23 is
     an Overlay Index, 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 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
    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 and NVE2 will withdraw the RT-2
    for M2-IP23. 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 5 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.

```
                   NVE2                          DGW1
          M2 +-----------+ +---------+    +-------------+
   +---TS2(VA)--|  (BD-10)  |-|         |----|  (BD-10)    |
   |      ESI23 +-----------+ |         |    |   IRB1\     |
   |        +                 |         |    |    (IP-VRF)|---+
   |        |                 |         |    +------------+  _|_
  SN1       |                 | VXLAN/  |                   (   )
   |        |                 | nvGRE   |       DGW2        ( WAN )
   |        +     NVE3        |         |    +-------------+ (___)
   |      ESI23 +-----------+ |         |----|  (BD-10)    |   |
   +---TS3(VA)--|  (BD-10)  |-|         |    |   IRB2\     |   |
          M3 +-----------+ +---------+    |    (IP-VRF)|---+
                                           +------------+
```
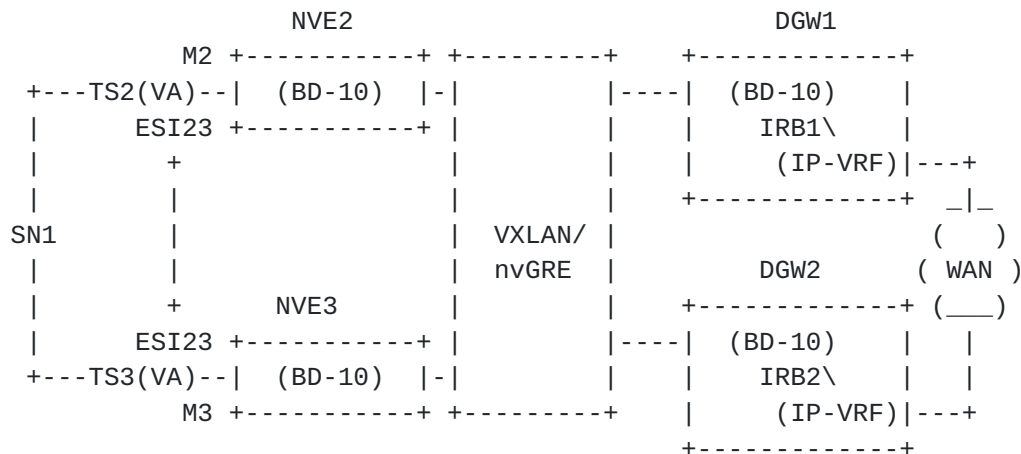
Figure 5 Bump-in-the-wire use-case

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 5.
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 AD
per-EVI route) and all the EVPN multi-homing redundancy mechanisms
can be re-used. 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 as 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/VSID
     field), as well as the BGP Encapsulation Extended Community as
     per [EVPN-OVERLAY].

   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.

(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.

(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 [EVPN-OVERLAY]).

   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 disposition model) or the VNI MAY directly
          identify the egress interface (for a label or VNI 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. We refer to this use-case 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 "Supplementary Broadcast Domain" (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 a IRB interfaces that
connect 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 6 will be used for the description of this model.

```
                      NVE1(M1)
                +------------+
    IP1+----|  (BD-1)    |                DGW1(M3)
            |      \     |    +---------+ +--------+
            |    (IP-VRF)|----|         |-|(IP-VRF)|----+
            |      /     |    |         | +--------+    |
      +---|  (BD-2)    |    |         |             _+_
      |   +------------+    |         |            (   )
     SN1|                   | VXLAN/ |          ( WAN )--H1
      |           NVE2(M2) | nvGRE/ |            (___)
      |   +------------+    | MPLS   |              +
      +---|  (BD-2)    |    |         | DGW2(M4)     |
          |      \     |    |         | +--------+   |
          |    (IP-VRF)|----|         |-|(IP-VRF)|----+
          |      /     |    +---------+ +--------+
    SN2+----|  (BD-3)    |
          +------------+
```
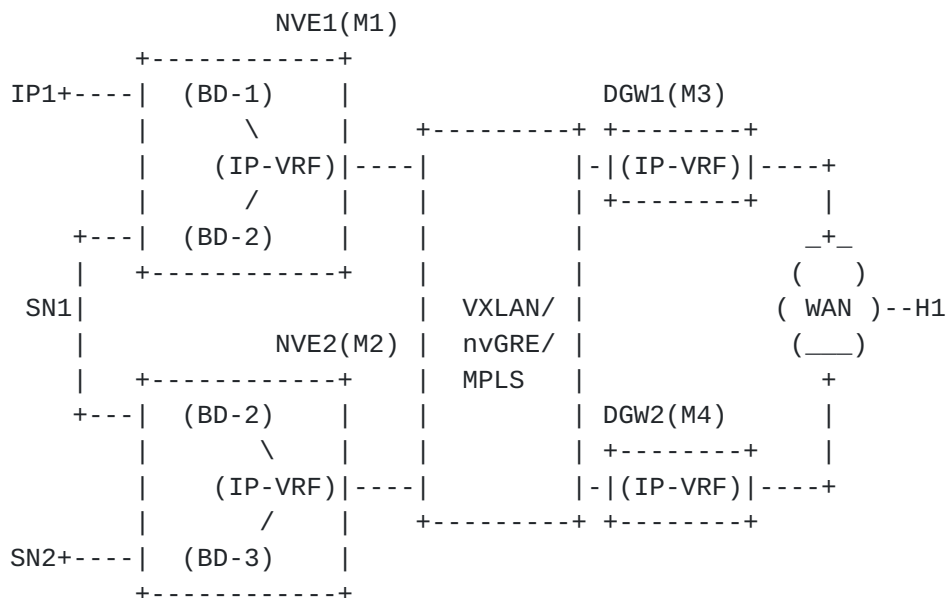
             Figure 6 Interface-less IP-VRF-to-IP-VRF model

   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. We assume 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 nvGRE.

   d) The solution may provide layer-3 connectivity among the IP-VRFs
      for IP NVO tunnels, for example, VXLAN GPE (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:

        o RD as per [RFC7432].

     o Eth-Tag ID=0.

     o IP address 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 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 re-used 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
       VXLAN GPE 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= SHOULD be 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, like in sections
4.4.2 or 4.4.3. An EVPN IP-VRF-to-IP-VRF implementation is REQUIRED
to support the ingress and egress procedures described in this
section.

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

Figure 7 will be used for the description of this model.

```
                      NVE1
           +------------+                    DGW1
   IP10+---+(BD-1)      | +--------------+ +------------+
        |  \         | |               | |            |
        |(IP-VRF)-(SBD)|               |(SBD)-(IP-VRF)|-----+
        |  /    IRB(IP1/M1)       IRB(IP3/M3)     |     |
     +---+(BD-2)      | |               | +------------+   _+_
     |   +------------+ |               |                 (   )
    SN1|                |     VXLAN/    |              ( WAN )--H1
     |            NVE2  |     nvGRE/    |                 (___)
     |   +------------+ |     MPLS      |    DGW2          +
     +---+(BD-2)      | |               | +------------+   |
        |  \         | |               | |            |   |
        |(IP-VRF)-(SBD)|               |(SBD)-(IP-VRF)|-----+
        |  /    IRB(IP2/M2)       IRB(IP4/M4)     |
    SN2+----+(BD-3)      | +--------------+ +------------+
           +------------+
```
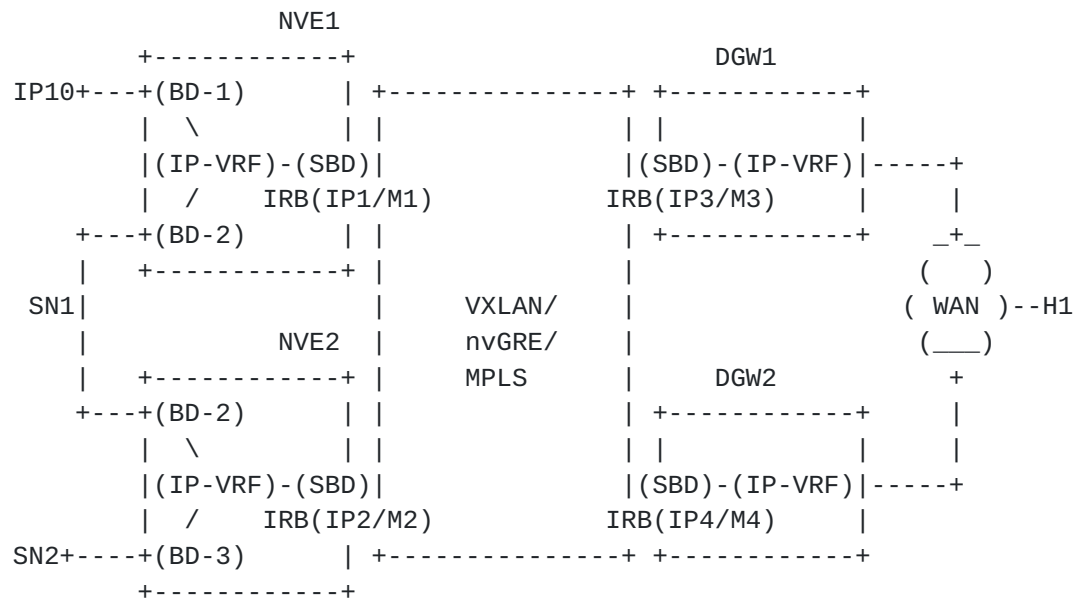
          Figure 7 Interface-ful with SBD IRB model

   In this model:

   a) As in [section 4.4.1](#), the NVEs and DGWs must provide connectivity
      between hosts in SN1, SN2, IP1 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 nvGRE.

   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:

       o RD as per [RFC7432].

       o Eth-Tag ID=0.

      o IP address length and IP address, as explained in the previous
        sections.

      o GW IP address=IRB-IP (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) = NVE1 IP.

   (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' since 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. An EVPN IP-VRF-to-IP-VRF implementation is REQUIRED
   to support the ingress and egress procedures described in this
   section.

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

   Figure 8 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.

```
                    NVE1
            +------------+                    DGW1
    IP1+----+(BD-1)      | +--------------+ +------------+
         |   \         | |                | |            |
         |(IP-VRF)-(SBD)|              (SBD)-(IP-VRF) |-----+
         |   /    IRB(M1)|              | IRB(M3)     |     |
      +---+(BD-2)        | |              | +------------+    _+_
      |   +------------+ |                |              (    )
     SN1|                |    VXLAN/      |            ( WAN )--H1
      |             NVE2 |    nvGRE/      |              (___)
      |   +------------+ |    MPLS        |    DGW2          +
      +---+(BD-2)        | |              | +------------+    |
         |   \         | |                | |            |    |
         |(IP-VRF)-(SBD)|              (SBD)-(IP-VRF) |-----+
         |   /    IRB(M2)|              | IRB(M4)     |
    SN2+----+(BD-3)      | +--------------+ +------------+
            +------------+
```
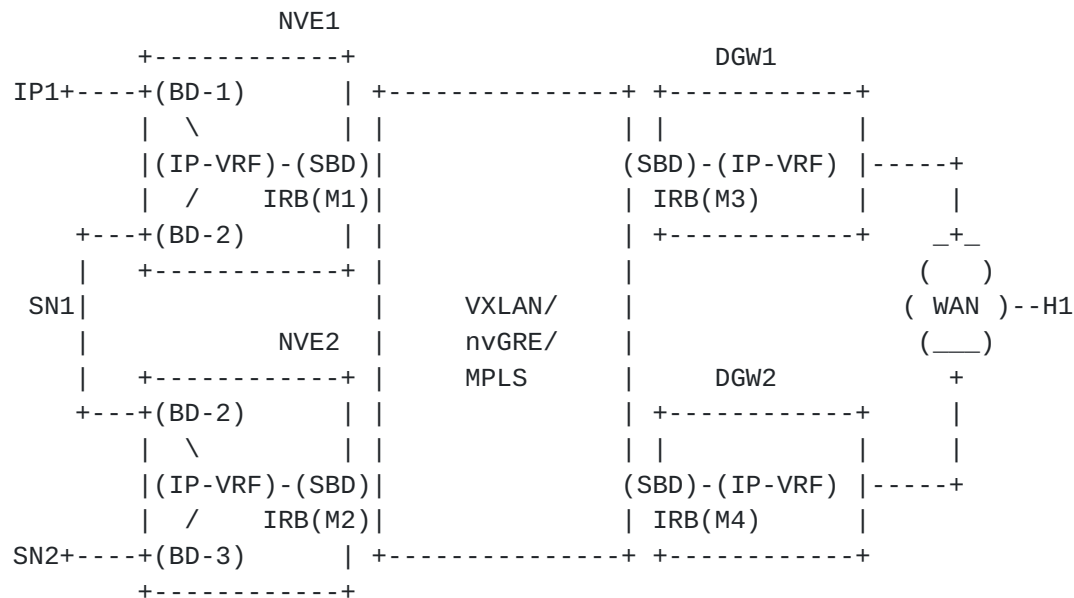
Figure 8 Interface-ful with unnumbered SBD IRB model

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 save IP
   addresses on those interfaces.

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
   nvGRE.

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= SHOULD be 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 re-
used 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:

    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
(as in section 4.4.2), only this time the SBD IRB does not have an IP
address. This model is OPTIONAL for an EVPN IP-VRF-to-IP-VRF
implementation.

5. Conclusions

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 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.

6. 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].

7. Security Considerations

The security considerations discussed in [RFC7432] apply to this
document.

8. IANA Considerations

This document requests the allocation of value 5 in the "EVPN Route
Types" registry defined by [RFC7432]:

```
Value      Description           Reference
5          IP Prefix route       [this document]
```


9. References

9.1 Normative 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>.

[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>.

[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>.

9.2 Informative References

[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

[EVPN-OVERLAY] Sajassi-Drake et al., "A Network Virtualization
Overlay Solution using EVPN", draft-ietf-bess-evpn-overlay-08.txt,
work in progress, March, 2017

10. 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.

11. 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

12. 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