                    EVPN Interworking with IPVPN
           draft-rabadan-sajassi-bess-evpn-ipvpn-interworking-02

Abstract

   EVPN is used as a unified control plane for tenant network intra and
   inter-subnet forwarding. When a tenant network spans not only EVPN
   domains but also domains where IPVPN provides inter-subnet
   forwarding, there is a need to specify the interworking aspects
   between both EVPN and IPVPN domains, so that the end to end tenant
   connectivity can be accomplished. This document specifies how EVPN
   should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families
   for inter-subnet forwarding.

Status of this Memo

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

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF), its areas, and its working groups. Note that
   other groups may also distribute working documents as Internet-
   Drafts.

   Internet-Drafts are draft documents valid for a maximum of six months

and may be updated, replaced, or obsoleted by other documents at any
time. It is inappropriate to use Internet-Drafts as reference
material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on July 11, 2019.

Copyright Notice

Table of Contents

1. Introduction and Problem Statement

   EVPN is used as a unified control plane for tenant network intra and
   inter-subnet forwarding. When a tenant network spans not only EVPN
   domains but also domains where IPVPN provides inter-subnet
   forwarding, there is a need to specify the interworking aspects
   between both EVPN and IPVPN domains, so that the end to end tenant
   connectivity can be accomplished. This document specifies how EVPN
   should interwork with VPN-IPv4/VPN-IPv6 and IPv4/IPv6 BGP families
   for inter-subnet forwarding.

   EVPN supports the advertisement of IPv4 or IPv6 prefixes in two
   different route types:

   o Route Type 2 - MAC/IP route (only for /32 and /128 host routes), as
     described by [INTER-SUBNET].

   o Route Type 5 - IP Prefix route, as described by [IP-PREFIX].

   When interworking with other BGP address families (AFIs/SAFIs) for
   inter-subnet forwarding, the IP prefixes in those two EVPN route
   types must be propagated to other domains using different SAFIs. Some
   aspects of that propagation must be clarified. Examples of these
   aspects or procedures across BGP families are: route selection, loop
   prevention or BGP Path attribute propagation. The Interworking PE
   concepts are defined in section 2, and the rest of the document
   describes the interaction between Interworking PEs and other PEs for
   end-to-end inter-subnet forwarding.

2. Terminology and Interworking PE Components

   This section summarizes the terminology related to the "Interworking
   PE" concept that will be used throughout the rest of the document.

```
    +----------------------------------------------------------------+
    |                                                                |
    |                  +-----------------+        Interworking PE |
    | Attachment    | +-----------------+                          |
    | Circuit(AC1) | | +----------+     |              MPLS/NVO tnl
   ---------------------*Bridge    |     |                  +------
    |              | | |Table(BT1)|  |    +-----------+  / \     \
  MPLS/NVO tnl +--------->|          *---------*          |<--> | Eth |
   -------+    |    | | |Eth-Tag x + |IRB1|          | \ /     /
  / Eth  / \<-+   | | +----------+   |    |          |  +------
  |     |   |     | |     ...        |    | IP-VRF1   |      |
   \      \ /<-+   | | +----------+   |    | RD2/RT2   |MPLS/NVO tnl
    -------+   |   | | |Bridge    |   |    |           |   +------
    |         +--------->|Table(BT2)|   |IRB2|          | / \     \
    |          | | |    *---------*          |<--> | IP  |
   ---------------------*Eth-Tag y |   |    +-----*-----+  \ /     /
    |   AC2     | | +----------+   |      AC3|         +------
    |           | |   MAC-VRF1     |       |         |       |
    |          +-+    RD1/RT1      |       |         |       |
    |            +-----------------+       | SAFIs        |
    |                                       | 1     +---+ |
   -----------------------------------------+  128   |BGP| |
    |                                         EVPN   +---+ |
    |                                                      |
    +----------------------------------------------------------------+
```
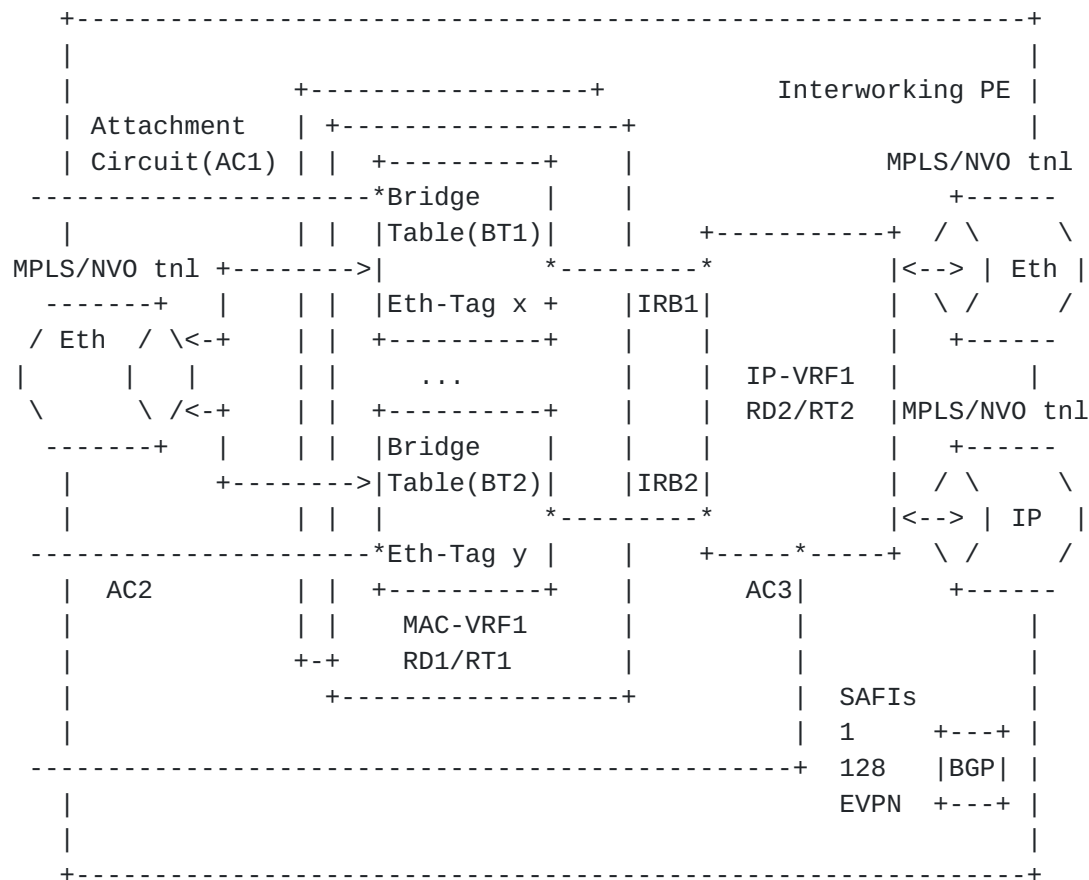
                  Figure 1 EVPN-IPVPN Interworking PE

   o ISF SAFI: Inter-Subnet Forwarding (ISF) SAFI is a MP-BGP Sub-
     Address Family that advertises reachability for IP prefixes and can
     be used for inter-subnet forwarding within a given tenant network.
     The ISF SAFIs are 1 (including IPv4 and IPv6 AFIs), 128 (including
     IPv4 and IPv6 AFIs) and 70 (EVPN, including only AFI 25).

   o ISF route: a route for a given prefix whose ISF SAFI may change as
     it transits different domains.

   o IP-VRF: an IP Virtual Routing and Forwarding table, as defined in
     [RFC4364]. It is also the instantiation of an IPVPN in a PE. Route
     Distinguisher and Route Target(s) are required properties of an IP-
     VRF.

   o MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in
     [RFC7432]. It is also the instantiation of an EVI (EVPN Instance)
     in a PE. Route Distinguisher and Route Target(s) are required
     properties and they are normally different than the ones defined in
     the associated IP-VRF.

o BT: a Bridge Table, as defined in [RFC7432]. A BT is the
  instantiation of a Broadcast Domain in a PE. When there is a single
  Broadcast Domain in a given EVI, the MAC-VRF in each PE will
  contain a single BT. When there are multiple BTs within the same
  MAC-VRF, each BT is associated to a different Ethernet Tag. The
  EVPN routes specific to a BT, will indicate which Ethernet Tag the
  route corresponds to.

  Example: In Figure 1, MAC-VRF1 has two BTs: BT1 and BT2. Ethernet
  Tag x is defined in BT1 and Ethernet Tag y in BT2.

o AC: Attachment Circuit or logical interface associated to a given
  BT or IP-VRF. To determine the AC on which a packet arrived, the PE
  will examine the combination of a physical port and VLAN tags
  (where the VLAN tags can be individual c-tags, s-tags or ranges of
  both).

  Example: In Figure 1, AC1 is associated to BT1, AC2 to BT2 and AC3
  to IP-VRF1.

o IRB: Integrated Routing and Bridging interface. It refers to the
  logical interface that connects a BT to an IP-VRF and allows to
  forward packets with destination in a different subnet.

o MPLS/NVO tnl: It refers to a tunnel that can be MPLS or NVO-based
  (Network Virtualization Overlays) and it is used by MAC-VRFs and
  IP-VRFs. Irrespective of the type, the tunnel may carry an Ethernet
  or an IP payload. MAC-VRFs can only use tunnels with Ethernet
  payloads (setup by EVPN), whereas IP-VRFs can use tunnels with
  Ethernet (setup by EVPN) or IP payloads (setup by EVPN or IPVPN).
  IPVPN-only PEs have IP-VRFs but they cannot send or receive traffic
  on tunnels with Ethernet payloads.

  Example: Figure 1 shows an MPLS/NVO tunnel that is used to
  transport Ethernet frames to/from MAC-VRF1. The PE determines the
  MAC-VRF and BT the packets belong to based on the EVPN label (MPLS
  or VNI). Figure 1 also shows two MPLS/NVO tunnels being used by IP-
  VRF1, one carrying Ethernet frames and the other one carrying IP
  packets.

o RT-2: Route Type 2 or MAC/IP route, as per [RFC7432].

o RT-5: Route Type 5 or IP Prefix route, as per [IP-PREFIX].

o Domain: Two PEs are in the same domain if they are attached to the
  same tenant and the packets between them do not require a data path
  IP lookup (in the tenant space) in any intermediate router. A
  gateway PE is always configured with multiple Domain-IDs.

Example 1: Figure 4 depicts an example where TS1 and TS2 belong to
the same tenant, and they are located in different Data Centers
that are connected by gateway PEs (see the gateway PE definition
later). These gateway PEs use IPVPN in the WAN. When TS1 sends
traffic to TS2, the intermediate routers between PE1 and PE2
require a tenant IP lookup in their IP-VRFs so that the packets can
be forwarded. In this example there are three different domains.
The gateway PEs connect the EVPN domains to the IPVPN domain.

```
                        GW1------------GW3
                       +------+        +------+
          +------------|IP-VRF|        |IP-VRF|-------------+
        PE1            +------+        +------+            PE2
         +------+   DC1    |      WAN     |    DC2    +------+
     TS1-|IP-VRF|   EVPN   |     IPVPN    |    EVPN   |IP-VRF|-TS2
         +------+        GW2            GW4           +---+--+
           |            +------+        +------+            |
           +------------|IP-VRF|        |IP-VRF|-------------+
                       +------+        +------+
                        +--------------+
           DOMAIN 1          DOMAIN 2      DOMAIN 3
         <---------------> <------------> <---------------->
```

Figure 4 Multiple domain DCI example

Example 2: Figure 5 illustrates a similar example, but PE1 and PE2
are now connected by a BGP-LU (BGP Labeled Unicast) tunnel, and
they have a BGP peer relationship for EVPN. Contrary to Example 1,
there is no need for tenant IP lookups on the intermediate routers
in order to forward packets between PE1 and PE2. Therefore, there
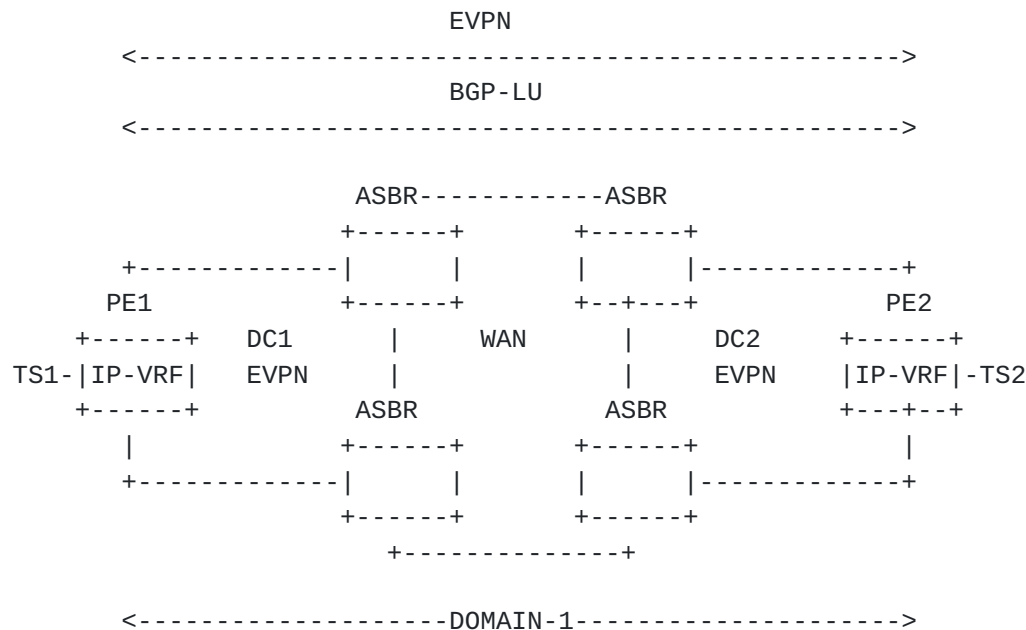is only one domain in the network and PE1/PE2 belong to it.

```
                              EVPN
           <--------------------------------------------------->
                             BGP-LU
           <--------------------------------------------------->


                    ASBR------------ASBR
                    +------+        +------+
         +-------------|      |      |      |-------------+
           PE1         +------+      +--+---+          PE2
          +------+  DC1     |    WAN    |    DC2    +------+
     TS1-|IP-VRF|  EVPN     |           |    EVPN   |IP-VRF|-TS2
          +------+     ASBR           ASBR          +---+--+
            |        +------+        +------+            |
         +-------------|      |      |      |-------------+
                    +------+        +------+
                      +-------------+


           <--------------------DOMAIN-1--------------------->
```

                  Figure 5 Single domain DCI example

   o Regular Domain: a domain in which a single control plane, IPVPN or
     EVPN, is used and which is composed of regular PEs, see below. In
     Figures 4 and 5, above, all domains are regular domains.

   o Composite Domain: a domain in which multiple control planes, IPVPN
     and EVPN, are used and which is composed of regular PEs, see below,
     and composite PEs, see below.

   o Regular PE: a PE that is attached to a domain, either regular or
     composite, and which uses one of the control plane protocols (IPVPN
     or EVPN) operating in the domain.

   o Interworking PE: a PE that may advertise a given prefix with an
     EVPN ISF route (RT-2 or RT-5) and/or an IPVPN ISF route. An
     interworking PE has one IP-VRF per tenant, and one or multiple MAC-
     VRFs per tenant. Each MAC-VRF may contain one or more BTs, where
     each BT may be attached to that IP-VRF via IRB. There are two types
     of Interworking PEs: composite PEs and gateway PEs. Both PE
     functions can be independently implemented per tenant and they may
     both be implemented for the same tenant.

     Example: Figure 1 shows an interworking PE of type gateway, where
     ISF SAFIs 1, 128 and 70 are enabled. IP-VRF1 and MAC-VRF1 are
     instantiated on the PE, and together provide inter-subnet
     forwarding for the tenant.

o Composite PE: an interworking PE that is attached to a composite
  domain and which advertises a given prefix to an IPVPN peer with an
  IPVPN ISF route, to an EVPN peer with an EVPN ISF route, and to a
  route reflector with both an IPVPN and EVPN ISF route. A composite
  PE performs the procedures of Sections 5 and 6.

  Example: Figure 2 shows an example where PE1 is a composite PE
  since PE1 has EVPN and another ISF SAFI enabled to the same route-
  reflector, and PE1 advertises a given IP prefix IPn/x twice, one
  using EVPN and another one using ISF SAFI 128. PE2 and PE3 are not
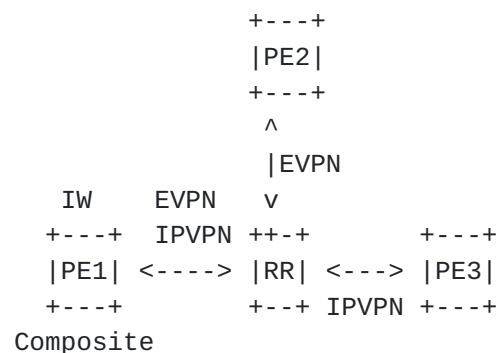 composite PEs.

```
                              +---+
                              |PE2|
                              +---+
                               ^
                               |EVPN
              IW     EVPN      v
             +---+  IPVPN ++-+          +---+
             |PE1| <----> |RR| <---> |PE3|
             +---+           +--+ IPVPN +---+
           Composite
```

Figure 2 Interworking composite PE example

o Gateway PE: an interworking PE that is attached to two domains,
  each either regular or composite, and which, based on
  configuration, does one of the following:

  - Propagates the same control plane protocol, either IPVPN or EVPN,
    between the two domains.

  - Propagates an ISF route with different ISF SAFIs between the two
    domains. E.g., propagate an EVPN ISF route in one domain as an
    IPVPN ISF route in the other domain and vice versa. A gateway PE
    performs the procedures of Sections 3, 4, 5 and 7.

    A gateway PE is always configured with multiple Domain-IDs. The
    Domain-ID is encoded in the Domain Path Attribute (D-PATH), and
    advertised along with EVPN and other ISF SAFI routes. Section 3
    describes the D-PATH attribute.

    Example: Figure 3 illustrates an example where PE1 is a gateway
    PE since the EVPN and IPVPN SAFIs are enabled on different BGP
    peers, and a given local IP prefix IPn/x is sent to both BGP

peers for the same tenant. PE2 and PE1 are in one domain and PE3 and PE1 are in another domain.
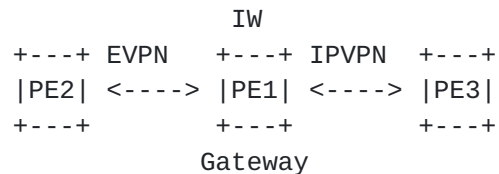
```
                          IW
            +---+ EVPN    +---+ IPVPN   +---+
            |PE2| <----> |PE1| <----> |PE3|
            +---+         +---+          +---+
                        Gateway
```

Figure 3 Interworking gateway PE example

o Composite/Gateway PE: an interworking PE that is both a composite PE and a gateway PE that is attached to two domains, one regular and one composite, and which does the following:

- Propagates an ISF route, either IPVPN or EVPN, from the regular domain into the composite domain. Within the composite domain it acts as a composite PE.

- Propagates an ISF route, either IPVPN or EVPN, from the composite domain into the regular domain. Within the regular domain it is propagated as an ISF route using the ISF SAFI for that domain.

  This is particularly useful when a tenant network is attached to both IPVPN and EVPN domains, any-to-any connectivity is required, and end-to-end control plane consistency, when possible, is desired.

  It would be instantiated by attaching the disparate, regular IPVPN and EVPN domains via these PEs to a central composite domain.

3. Domain Path Attribute (D-PATH)

The BGP Domain Path (D-PATH) attribute is an optional and transitive BGP path attribute.

Similar to AS_PATH, D-PATH is composed of a length field followed by a sequence of Domain segments, where each domain segment is represented by <DOMAIN-ID:ISF_SAFI_TYPE>.

o The length field is a 1-octet field, containing the number of domain segments.

   o DOMAIN-ID is a 6-octet field that represents a domain. It is
     composed of a 4-octet Global Administrator sub-field and a 2-octet
     Local Administrator sub-field. The Global Administrator sub-field
     MAY be filled with an Autonomous System Number (ASN), an IPv4
     address, or any value that guarantees the uniqueness of the DOMAIN-
     ID when the tenant network is connected to multiple Operators.

   o ISF_SAFI_TYPE is a 1-octet field that indicates the Inter-Subnet
     Forwarding SAFI type in which a route was advertised in the DOMAIN.
     The following types are valid in this document:

     Value   Type

     1   SAFI 1
     70   EVPN
     128   SAFI 128

   About the BGP D-PATH attribute:

   a) Identifies the sequence of domains, each identified by a <DOMAIN-
      ID:ISF_SAFI_TYPE> through which a given ISF route has passed.

      - This attribute list may contain zero, one or more entries.

      - The first entry in the list (leftmost) is the <DOMAIN-
        ID:ISF_SAFI_TYPE> from which a gateway PE is propagating an ISF
        route. The last entry in the list (rightmost) is the <DOMAIN-
        ID:ISF_SAFI_TYPE> from which a gateway PE received an ISF route
        without a D-PATH attribute. Intermediate entries in the list are
        domains that the ISF route has transited.

      - As an example, an ISF route received with a D-PATH attribute of
        {<6500:2:IPVPN>,<6500:1:EVPN>} indicates that the ISF route was
        originated in EVPN domain 6500:1, and propagated into IPVPN
        domain 6500:2.

   b) It is added/modified by a gateway PE when propagating an update to
      a different domain:

      - A gateway PE's IP-VRF, that connects two domains, belongs to two
        DOMAIN-IDs, e.g. 6500:1 for EVPN and 6500:2 for IPVPN.

      - Whenever a prefix arrives at a gateway PE in a particular ISF
        SAFI route, if the gateway PE needs to export that prefix to a
        BGP peer, the gateway PE will prepend a <DOMAIN-
        ID:ISF_SAFI_TYPE> segment to the list of segments in the

        received D-PATH.

    - For instance, in an IP-VRF configured with DOMAIN-IDs 6500:1 for
      EVPN and 6500:2 for IPVPN, if an EVPN route for prefix P is
      received and P installed in the IP-VRF, the IPVPN route for P
      that is exported to an IPVPN peer will prepend the segment
      <6500:1:EVPN> to the previously received D-PATH attribute.
      Likewise, IP-VRF prefixes that are received from IP-VPN, will be
      exported to EVPN peers with the additional segment
      <6500:2:IPVPN>.

    - In the above example, if the EVPN route is received without D-
      PATH, the gateway PE will add the D-PATH attribute with segment
      <6500:1:EVPN> when re-advertising to domain 6500:2.

    - Within the originating domain, the update does not contain a D-
      PATH attribute because the update has not passed through a
      gateway PE yet.

  c) The gateway PE MUST NOT add the D-PATH attribute to ISF routes
     generated for IP-VRF prefixes that are not learned via any ISF
     SAFI, for instance, local prefixes.

  d) An ISF route received by a gateway PE with a D-PATH attribute that
     contains one or more of its locally configured domains for the IP-
     VRF is considered to be a looped ISF route and MUST be dropped.

  e) The number of domain segments in the D-PATH attribute indicates
     the number of gateway PEs that the ISF route update has transited.

[3.1]. D-PATH and Loop Prevention

   The D-PATH attribute is used to prevent loops in interworking PE
   networks. For instance, in the example of Figure 4, gateway GW1
   receives TS1 prefix in two different ISF routes:

   o In an EVPN RT-5 with next-hop PE1 and no D-PATH attribute.

   o In a SAFI 128 route with next-hop GW2 and D-PATH = (6500:1:EVPN),
     assuming that DOMAIN-ID for domain 1 is 6500:1.

   Gateway GW1 flags the SAFI 128 route as a loop, and does not re-
   advertise it to the EVPN neighbors since the route includes the GW1's
   local domain.

   In general, any interworking PE that imports an ISF route MUST flag
   the route as "looped" if its D-PATH contains a <DOMAIN-

   ID:ISF_SAFI_TYPE> segment, where DOMAIN-ID matches a local DOMAIN-ID
   in the tenant IP-VRF.

[4](). **BGP Path Attribute Propagation across ISF SAFIs**

   Based on configurations a gateway PE is required to propagate an ISF
   route with different ISF SAFIs between two domains. This requires a
   definition of what a gateway PE is to do with Path attributes
   attached to the ISF route that it is propagating.

[4.1](). **No-Propagation-Mode**

   This is the default mode of operation. In this mode, the gateway PE
   will simply re-initialize the Path Attributes when propagating an ISF
   route, as though it would for direct or local IP prefixes. This model
   may be enough in those use-cases where the EVPN domain is considered
   an "abstracted" CE and remote IPVPN/IP PEs don't need to consider the
   original EVPN Attributes for path calculations.

   Since this mode of operation does not propagate the D-PATH attribute
   either, redundant gateway PEs are exposed to routing loops. Those
   loops may be resolved by policies and the use of other attributes,
   such as the Route Origin extended community [[RFC4360]()], however not
   all the loop situations may be solved.

[4.2](). **Uniform-Propagation-Mode**

   In this mode, the gateway PE simply keeps accumulating or mapping
   certain key commonly used Path Attributes when propagating an ISF
   route. This mode is typically used in networks where EVPN and IPVPN
   SAFIs are used seamlessly to distribute IP prefixes.

   The following rules MUST be observed by the gateway PE when
   propagating Path Attributes:

   o The gateway PE imports an ISF route in the IP-VRF and stores the
     original Path Attributes. The following set of Path Attributes
     SHOULD be propagated by the gateway PE to other ISF SAFIs (other
     Path Attributes SHOULD NOT be propagated):

     - AS_PATH
     - D-PATH
     - IBGP-only Path Attributes: LOCAL_PREF, ORIGINATOR_ID, CLUSTER_ID
     - MED
     - AIGP
     - Communities, (non-EVPN) Extended Communities and Large
       Communities

   o When propagating an ISF route to a different ISF SAFI and IBGP
     peer, the gateway PE SHOULD copy the AS_PATH of the originating
     family and add it to the destination family without any
     modification. When re-advertising to a different ISF SAFI and EBGP
     peer, the gateway PE SHOULD copy the AS_PATH of the originating
     family and prepend the IP-VRF's AS before sending the route.

   o When propagating an ISF route to IBGP peers, the gateway PE SHOULD
     copy the IBGP-only Path Attributes from the originating SAFI to the
     re-advertised route.

   o Communities, non-EVPN Extended Communities and Large Communities
     SHOULD be copied by the gateway PE from the originating SAFI route.

4.3. Aggregation of Routes and Path Attribute Propagation

   Instead of propagating a high number of (host) ISF routes between ISF
   SAFIs, a gateway PE that receives multiple ISF routes of one ISF SAFI
   MAY choose to propagate a single ISF aggregate route with a different
   ISF SAFI. In this document, aggregation is used to combine the
   characteristics of multiple ISF routes of the same ISF SAFI in such
   way that a single aggregate ISF route of a different ISF SAFI can be
   propagated. Aggregation of multiple ISF routes of one ISF SAFI into
   an aggregate ISF route of a different ISF SAFI is only done by a
   gateway PE.

   Aggregation on gateway PEs may use either the No-Propagation-Mode or
   the Uniform-Propagation-Mode explained in Sections 4.1. and 4.2,
   respectively.

   When using Uniform-Propagation-Mode, Path Attributes of the same type
   code MAY be aggregated according to the following rules:

   o AS_PATH is aggregated based on the rules in [RFC4271]. The gateway
     PEs SHOULD NOT receive AS_PATH attributes with path segments of
     type AS_SET [RFC6472]. Routes received with AS_PATH attributes
     including AS_SET path segments MUST NOT be aggregated.

   o ISF routes that have different attributes of the following type
     codes MUST NOT be aggregated: D-PATH, LOCAL_PREF, ORIGINATOR_ID,
     CLUSTER_ID, MED or AIGP.

   o The Community, Extended Community and Large Community attributes of
     the aggregate ISF route MUST contain all the Communities/Extended
     Communities/Large Communities from all of the aggregated ISF
     routes.

   Assuming the aggregation can be performed (the above rules are
   applied), the operator should consider aggregation to deal with
   scaled tenant networks where a significant number of host routes
   exists. For a example, large Data Centers.

5. **Route Selection Process between EVPN and other ISF SAFIs**

   A PE may receive an IP prefix in ISF routes with different ISF SAFIs,
   from the same or different BGP peer. It may also receive the same IP
   prefix (host route) in an EVPN RT-2 and RT-5. A route selection
   algorithm across all ISF SAFIs is needed so that:

   o Different gateway and composite PEs have a consistent and
     deterministic view on how to reach a given prefix.

   o Prefixes advertised in EVPN and other ISF SAFIs can be compared
     based on path attributes commonly used by operators across
     networks.

   o Equal Cost Multi-Path (ECMP) is allowed across EVPN and other ISF
     SAFI routes.

   For a given prefix advertised in one or more non-EVPN ISF routes, the
   BGP best path selection procedure will produce a set of "non-EVPN
   best paths". For a given prefix advertised in one or more EVPN ISF
   routes, the BGP best path selection procedure will produce a set of
   "EVPN best paths". To support IP/EVPN interworking, it is then
   necessary to run a tie-breaking selection algorithm on the union of
   these two sets. This tie-breaking algorithm begins by considering all
   EVPN and other ISF SAFI routes, equally preferable routes to the same
   destination, and then selects routes to be removed from
   consideration. The process terminates as soon as only one route
   remains in consideration.

   The route selection algorithm must remove from consideration the
   routes following the rules and the order defined in [RFC4271], with
   the following exceptions and in the following order:

   1- Immediately after removing from consideration all routes that are
      not tied for having the highest Local Preference, any routes that
      do not have the shortest D-PATH are also removed from
      consideration. Routes with no D-PATH are considered to have a
      zero-length D-PATH.

   2- Then regular [RFC4271] selection criteria is followed.

   3- At the end of the selection algorithm, if at least one route still

under consideration is an RT-2 route, remove from consideration
any RT-5 routes.

4- Steps 1-3 could possibly leave Equal Cost Multi-Path (ECMP)
between IP and EVPN paths. By default, the EVPN path is considered
(and the IP path removed from consideration). However, if ECMP
across ISF SAFIs is enabled by policy, and an "IP path" and an
"EVPN path" remain at the end of step 3, both path types will be
used.

Example 1 - PE1 receives the following routes for IP1/32, that are
candidate to be imported in IP-VRF-1:

   {SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=(100,200)}
   {SAFI=EVPN, RT-5, Local-Pref=100, AS-Path=(100,200)}
   {SAFI=128, Local-Pref=100, AS-Path=(100,200)}

   Selected route: {SAFI=EVPN, RT-2, Local-Pref=100, AS-Path=100,200]
   (due to step 3, and no ECMP)

Example 2 - PE1 receives the following routes for IP2/24, that are
candidate to be imported in IP-VRF-1:

   {SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-Path=(100,200),
   MED=10}
   {SAFI=128, D-PATH=(6500:1:EVPN,6500:2:IPVPN), AS-Path=(200),
   MED=200}

   Selected route: {SAFI=EVPN, RT-5, D-PATH=(6500:3:IPVPN), AS-
   Path=(100,200), MED=10} (due to step 1)

6. Composite PE Procedures

As described in Section 2, composite PEs are typically used in tenant
networks where EVPN and IPVPN are both used to provide inter-subnet
forwarding within the same composite domain.

Figure 6 depicts an example of a composite domain, where PE1/PE2/PE4
are composite PEs (they support EVPN and IPVPN ISF SAFIs on their
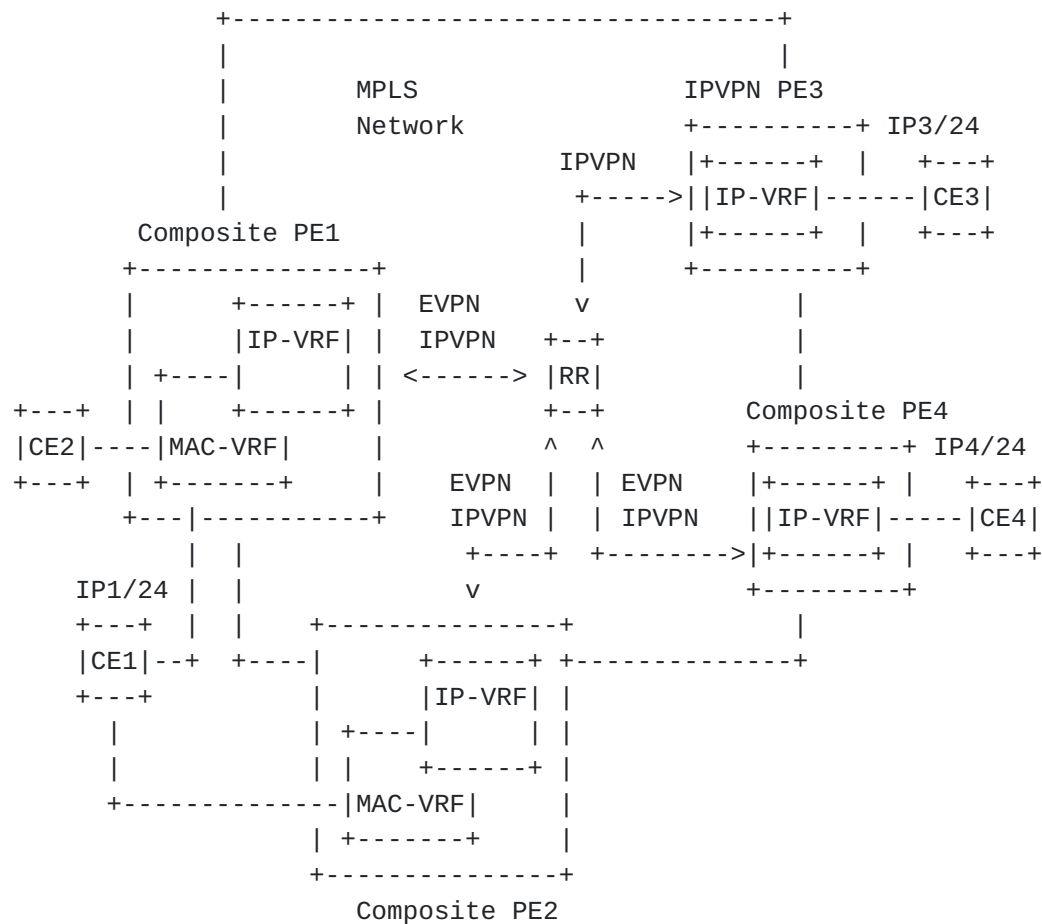peering to the Route Reflector), and PE3 is a regular IPVPN PE.

```
                +-----------------------------------+
                |                                   |
                |       MPLS            IPVPN PE3
                |       Network         +----------+ IP3/24
                |                IPVPN  |+------+  |    +---+
                |              +----->||IP-VRF|------|CE3|
             Composite PE1     |       |+------+  |    +---+
          +---------------+    |       +----------+
          |      +------+ |  EVPN    v              |
          |      |IP-VRF| |  IPVPN  +--+            |
          | +----|      | | <------> |RR|           |
   +---+  | |    +------+ |         +--+   Composite PE4
   |CE2|----|MAC-VRF|     |        ^  ^    +---------+ IP4/24
   +---+  | +-------+     |  EVPN  |  | EVPN |+------+ |    +---+
          +---|-----------+  IPVPN |  | IPVPN ||IP-VRF|-----|CE4|
              | |               +----+  +-------->|+------+ |    +---+
     IP1/24 | |                 v                +---------+
     +---+  | |     +---------------+            |
     |CE1|--+  +----|      +------+ +--------------+
     +---+          |      |IP-VRF| |
       |            | +----|      | | |
       |            | | |    +------+ |
     +-------------|MAC-VRF|       |
                   | +-------+     |
                   +---------------+
                     Composite PE2
```

                    Figure 6 Composite PE example

   In a composite domain with composite and regular PEs:

   o The composite PEs advertise the same IP prefixes in each ISF SAFI
     to the RR. For example, in Figure 6, the prefix IP1/24 is
     advertised by PE1 and PE2 to the RR in two separate NLRIs, one for
     AFI/SAFI 1/128 and another one for EVPN.

   o The RR does not forward EVPN routes to PE3 (since the RR does not
     have the EVPN SAFI enabled on its BGP session to PE3), whereas the
     IPVPN routes are forwarded to all the PEs.

   o PE3 receives only the IPVPN route for IP1/24 and resolves the BGP
     next-hop to an MPLS tunnel (with IP payload) to PE1 and/or PE2.

   o Composite PE4 receives IP1/24 encoded in EVPN and another ISF SAFI
     route (EVPN RT-5 and IPVPN). The route selection follows the
     procedures in Section 5. Assuming an EVPN route is selected, PE4

resolves the BGP next-hop to an MPLS tunnel (with Ethernet or IP
payload) to PE1 and/or PE2. As described in Section 2, two EVPN PEs
may use tunnels with Ethernet or IP payloads to connect their IP-
VRFs, depending on the [IP-PREFIX] model implemented. If some
attributes are modified so that the route selection process
(Section 5) results in PE4 selecting the IPVPN path instead of the
EVPN path, the operator should be aware that the EVPN advanced
forwarding features, e.g. recursive resolution to overlay indexes,
will be lost for PE4.

o The other composite PEs (PE1 and PE2) receive also the same IP
  prefix via EVPN and IPVPN SAFIs and they also follow the route
  selection in Section 5.

o When a given route has been selected as the route for a particular
  packet, the transmission of the packet is done according to the
  rules for that route's AFI/SAFI.

o It is important to note that in composite domains, such as the one
  in Figure 6, the EVPN advanced forwarding features will only be
  available to composite and EVPN PEs (assuming they select an RT-5
  to forward packets for a given IP prefix), and not to IPVPN PEs.
  For example, assuming PE1 sends IP1/24 in an EVPN and an IPVPN
  route and the EVPN route is the best one in the selection, the
  recursive resolution of the EVPN RT-5s can only be used in PE2 and
  PE4 (composite PEs), and not in PE3 (IPVPN PE). As a consequence of
  this, the indirection provided by the RT5's recursive resolution
  and its benefits in a scaled network, will not be available in all
  the PEs in the network.

7. Gateway PE Procedures

Section 2 defines a gateway PE as an Interworking PE that advertises
IP prefixes to different BGP peers, using EVPN to one BGP peer and
another ISF SAFI to another BGP peer. Examples of gateway PEs are
Data Center gateways connecting domains that make use of EVPN and
other ISF SAFIs for a given tenant. Figure 7 illustrates this use-
case, in which PE1 and PE2 (and PE3/PE4) are gateway PEs
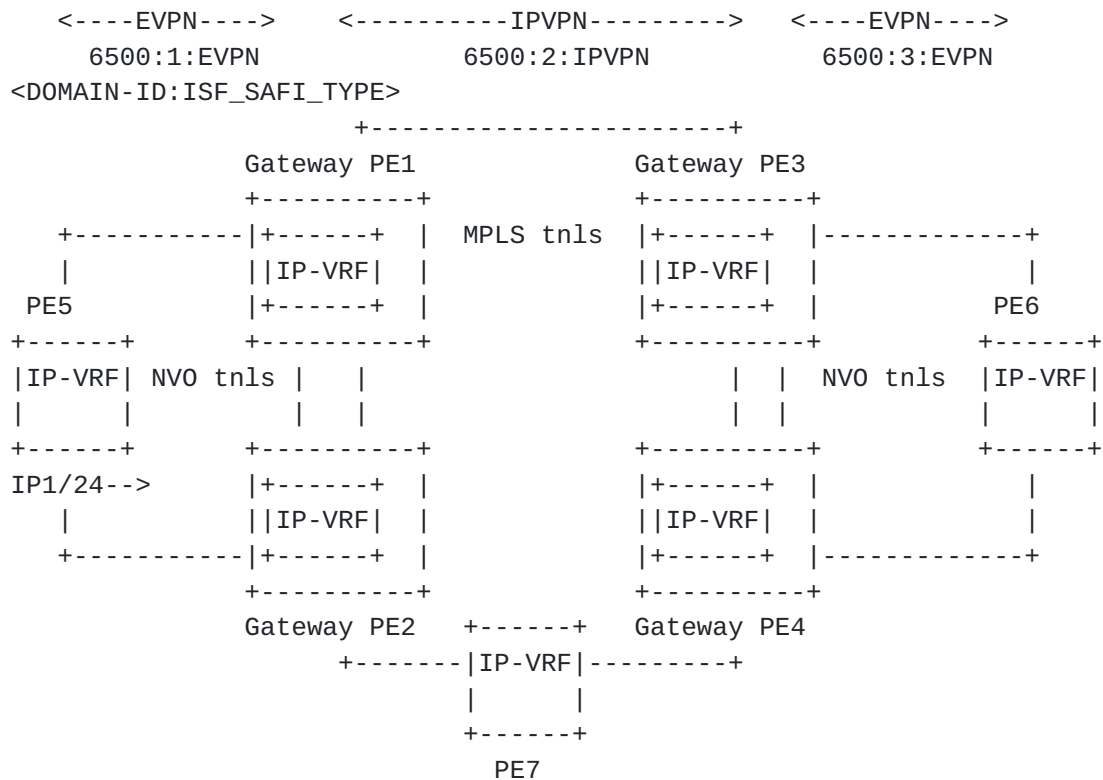interconnecting domains for the same tenant.

```
      <----EVPN---->    <----------IPVPN---------->   <----EVPN---->
        6500:1:EVPN             6500:2:IPVPN            6500:3:EVPN
     <DOMAIN-ID:ISF_SAFI_TYPE>
                       +-----------------------+
                 Gateway PE1             Gateway PE3
                 +----------+            +----------+
     +-----------|+------+  |  MPLS tnls |+------+  |-------------+
     |           ||IP-VRF|  |            ||IP-VRF|  |             |
      PE5        |+------+  |            |+------+  |            PE6
    +------+     +----------+            +----------+      +------+
    |IP-VRF| NVO tnls |    |                 |  | NVO tnls |IP-VRF|
    |      |   |      |    |                 |  |          |      |
    +------+     +----------+            +----------+      +------+
    IP1/24-->    |+------+  |            |+------+  |             |
       |         ||IP-VRF|  |            ||IP-VRF|  |             |
     +-----------|+------+  |            |+------+  |-------------+
                 +----------+            +----------+
                 Gateway PE2  +------+  Gateway PE4
                     +-------|IP-VRF|---------+
                             |      |
                             +------+
                               PE7
```

                    Figure 7 Gateway PE example

   The gateway PE procedures are described as follows:

   o A gateway PE that imports an ISF SAFI-x route to prefix P in an IP-
     VRF, MUST export P in ISF SAFI-y if:

     1. P is installed in the IP-VRF (hence the SAFI-x route is the best
        one for P) and

     2. PE has a BGP peer for SAFI-y (enabled for the same IP-VRF) and

     3. Either x or y is EVPN.

        In the example of Figure 7, gateway PE1 and PE2 receive an EVPN
        RT-5 with IP1/24, install the prefix in the IP-VRF and re-
        advertise it using SAFI 128.

   o ISF SAFI routes advertised by a gateway PE MUST include a D-PATH
     attribute, so that loops can be detected in remote gateway PEs.
     When a gateway PE propagates an IP prefix between EVPN and another
     ISF SAFI, it MUST prepend a <DOMAIN-ID:ISF_SAFI_TYPE> to the
     received D-PATH attribute. The DOMAIN-ID and ISF_SAFI_TYPE fields
     refer to the domain over which the gateway PE received the IP
     prefix and the ISF SAFI of the route, respectively. If the received

IP prefix route did not include any D-PATH attribute, the gateway IP MUST add the D-PATH when readvertising. The D-PATH in this case will have only one segment on the list, the <DOMAIN-ID:ISF_SAFI_TYPE> of the received route.

In the example of Figure 7, gateway PE1/PE2 receive the EVPN RT-5 with no D-PATH attribute since the route is originated at PE5. Therefore PE1 and PE2 will add the D-PATH attribute including <DOMAIN-ID:ISF_SAFI_TYPE> = <6500:1:EVPN>. Gateways PE3/PE4 will propagate the route again, now prepending their <DOMAIN-ID:ISF_SAFI_TYPE> = <6500:2:IPVPN>. PE6 receives the EVPN RT-5 routes with D-PATH = {<6500:2:IPVPN>,<6500:1:EVPN>} and can use that information to make BGP path decisions.

o The gateway PE MAY use the Route Distinguisher of the IP-VRF to readvertise IP prefixes in EVPN or the other ISF SAFI.

o The label allocation used by each gateway PE is a local implementation matter. The IP-VRF advertising IP prefixes for EVPN and another ISF SAFI may use a label per-VRF, per-prefix, etc.

o The gateway PE MUST be able to use the same or different set of Route Targets per ISF SAFI on the same IP-VRF. In particular, if different domains use different set of Route Targets for the same tenant, the gateway PE MUST be able to import and export routes with the different sets.

o Even though Figure 7 only shows two domains per gateway PE, the gateway PEs may be connected to more than two domains.

o There is no limitation of gateway PEs that a given IP prefix can pass through until it reaches a given PE.

o It is worth noting that an IP prefix that was originated in an EVPN domain but traversed a different ISF SAFI domain, will lose EVPN-specific attributes that are used in advanced EVPN procedures. For example, even if PE1 advertises IP1/24 along with a given non-zero ESI (for recursive resolution to that ESI), when PE6 receives the IP prefix in an EVPN route, the ESI value will be zero. This is because the route traverses an ISF SAFI domain that is different than EVPN.

8. Interworking Use-Cases

While Interworking PE networks may well be similar to the examples described in Sections 6 and 7, in some cases a combination of both functions may be required. Figure 8 illustrates an example where the

gateway PEs are also composite PEs, since not only they need to re-
advertise IP prefixes from EVPN routes to another ISF SAFI routes,
but they also need to interwork with IPVPN-only PEs in a domain with
a mix of composite and IPVPN-only PEs.

```
                    +-----------------------------------+
                    |                                   |
                    |          MPLS            IPVPN PE3 |
                    |          Network         +---------+
                    |                   IPVPN  |+------+ |
                    |                   +----->||IP-VRF|---TS3
        (GW+composite) PE1             |       |+------+ |
          +---------------+            |       +---------+
          |       +------+ |  EVPN     v            |
          |       |IP+VRF| |  IPVPN   +-++          |
          | +----|        | | <------> |RR|          |
      +--------| |   +------+ |         +--+      Composite PE4
      |        | |IP+VRF|    |        ^   ^       +---------+
      |        | +------+    |  EVPN  |   | EVPN  |+------+ |
    +----+     +---------------+ IPVPN |   | IPVPN ||IP-VRF|---TS4
  TS1-|NVE1|            |         +----+   +-------->|+------+ |
    +----+             |         v            +---------+
      |     EVPN DC    |   +---------------+          |
      |     NVO tnls  +----|      +------+ |-------------+
      |               |    |      |IP+VRF| |
      |               |    | +----|      | |
      |               |    | |   +------+ |
      |    +----+      |    | |MAC+VRF|    |
      +-----|NVE2|---------| +-------+    |
            +----+         +---------------+
              |          (GW+composite) PE2
             TS2
```
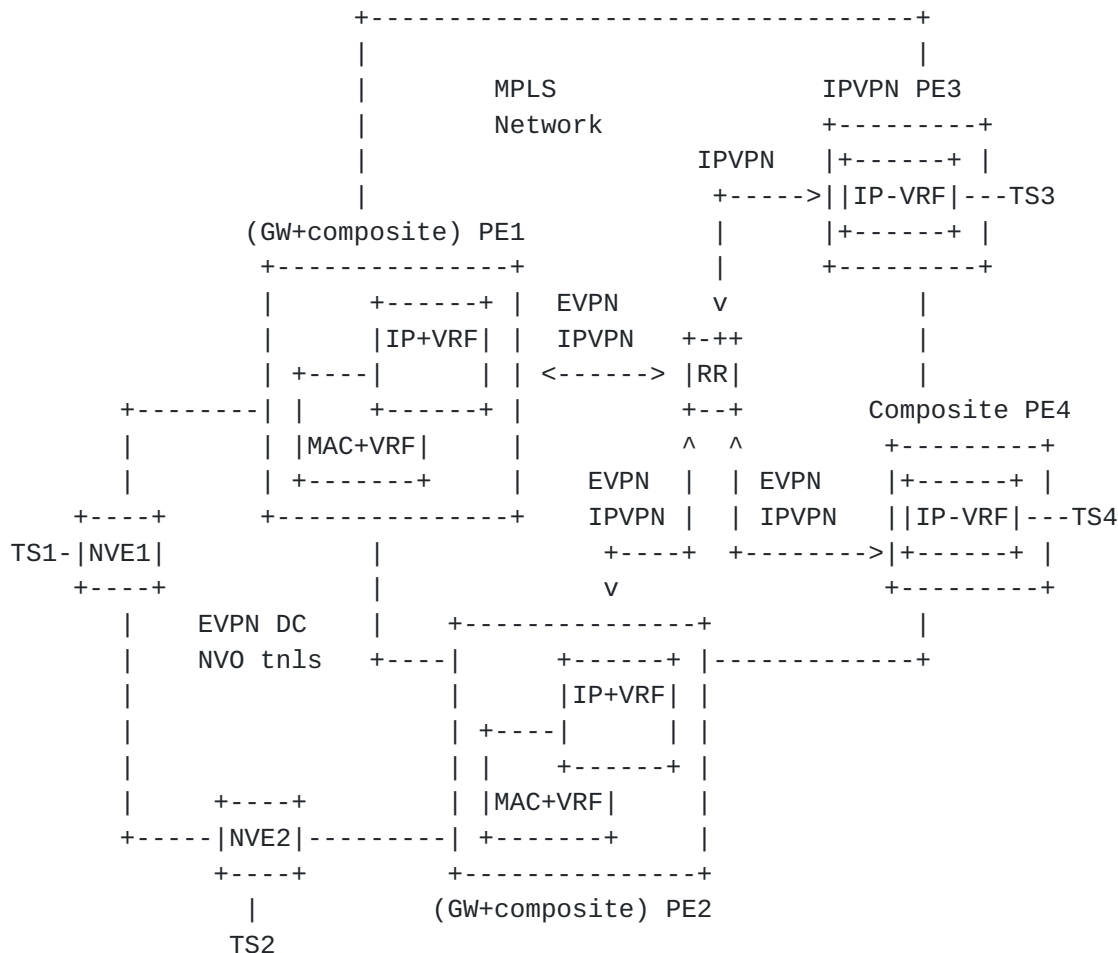
Figure 8 Gateway and composite combined functions - example

In the example above, PE1 and PE2 MUST follow the procedures
described in Sections 6 and 7. Compared to section 7, PE1 and PE2 now
need to also propagate prefixes from EVPN to EVPN, in addition to
propagating prefixes from EVPN to IPVPN.

It is worth noting that PE1 and PE2 will receive TS4's IP prefix via
IPVPN and RT-5 routes. When readvertising to NVE1 and NVE2, PE1 and
PE2 will consider the D-PATH rules and attributes of the selected
route for TS4 (Section 5 describes the Route Selection Process).

9. Conclusion

This document describes the procedures required in PEs that use EVPN
and another Inter-Subnet Forwarding SAFI to import and export IP
prefixes for a given tenant. In particular, this document defines:

o A route selection algorithm so that a PE can determine what path to
  choose between EVPN paths and other ISF SAFI paths.

o A new BGP Path attribute called D-PATH that provides loop
  protection and visibility on the domains a particular route has
  traversed.

o The way Path attributes should be propagated between EVPN and
  another ISF SAFI.

o The procedures that must be followed on Interworking PEs that
  behave as composite PEs, gateway PEs or a combination of both.

The above procedures provide an operator with the required tools to
build large tenant networks that may span multiple domains, use
different ISF SAFIs to handle IP prefixes, in a deterministic way and
with routing loop protection.

10. Conventions used in this document

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

11. Security Considerations

This section will be added in future versions.

12. IANA Considerations

This document defines a new BGP path attribute known as the BGP
Domain Path (D-PATH) attribute and requests IANA to assign a new
attribute code type from the "BGP Path Attributes" subregistry under
the "Border Gateway Protocol (BGP) Parameters" registry.

13. References

13.1. Normative References

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

   [RFC4271]   Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A
   Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271,
   January 2006, <http://www.rfc-editor.org/info/rfc4271>.

13.2. Informative References

   [RFC4360]   Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
   Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February
   2006, <http://www.rfc-editor.org/info/rfc4360>.

   [IP-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN",
   draft-ietf-bess-evpn-prefix-advertisement-11, May, 2018.

   [INTER-SUBNET] Sajassi et al., "IP Inter-Subnet Forwarding in EVPN",
   draft-ietf-bess-evpn-inter-subnet-forwarding-05.txt, work in
   progress, July, 2018

   [ENCAP-ATT] Rosen et al., "The BGP Tunnel Encapsulation Attribute",
   draft-ietf-idr-tunnel-encaps-10.txt, work in progress, August, 2018.

   [RFC6472] Kumari, W. and K. Sriram, "Recommendation for Not Using
   AS_SET and AS_CONFED_SET in BGP", BCP 172, RFC 6472, DOI
   10.17487/RFC6472, December 2011, <https://www.rfc-
   editor.org/info/rfc6472>.

14. Acknowledgments

15. Contributors

16. Authors' Addresses

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

Ali Sajassi (editor)
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
EMail: sajassi@cisco.com

Eric C. Rosen
Juniper Networks, Inc.
EMail: erosen@juniper.net

John Drake
Juniper Networks, Inc.
EMail: jdrake@juniper.net

Wen Lin
Juniper Networks, Inc.
EMail: wlin@juniper.net

Jim Uttaro
AT&T
Email: ju1738@att.com

Adam Simpson
Nokia
Email: adam.1.simpson@nokia.com