        A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)

Abstract

   This document specifies how Ethernet VPN (EVPN) can be used as a
   Network Virtualization Overlay (NVO) solution and explores the
   various tunnel encapsulation options over IP and their impact on the
   EVPN control plane and procedures. In particular, the following
   encapsulation options are analyzed: Virtual Extensible LAN (VXLAN),
   Network Virtualization using Generic Routing Encapsulation (NVGRE),
   and MPLS over GRE. This specification is also applicable to Generic
   Network Virtualization Encapsulation (GENEVE); however, some
   incremental work is required, which will be covered in a separate
   document. This document also specifies new multihoming procedures
   for split-horizon filtering and mass withdrawal. It also specifies
   EVPN route constructions for VXLAN/NVGRE encapsulations and
   Autonomous System Border Router (ASBR) procedures for multihoming of
   Network Virtualization Edge (NVE) devices.

Status of This Memo

   This is an Internet Standards Track document.

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

   Information about the current status of this document, any errata,
   and how to provide feedback on it may be obtained at
   https://www.rfc-editor.org/info/rfc8365.

Table of Contents

1. Introduction

   This document specifies how Ethernet VPN (EVPN) [RFC7432] can be used
   as a Network Virtualization Overlay (NVO) solution and explores the
   various tunnel encapsulation options over IP and their impact on the
   EVPN control plane and procedures. In particular, the following
   encapsulation options are analyzed: Virtual Extensible LAN (VXLAN)
   [RFC7348], Network Virtualization using Generic Routing Encapsulation
   (NVGRE) [RFC7637], and MPLS over Generic Routing Encapsulation (GRE)
   [RFC4023]. This specification is also applicable to Generic Network
   Virtualization Encapsulation (GENEVE) [GENEVE]; however, some
   incremental work is required, which will be covered in a separate
   document [EVPN-GENEVE]. This document also specifies new multihoming
   procedures for split-horizon filtering and mass withdrawal. It also
   specifies EVPN route constructions for VXLAN/NVGRE encapsulations and
   Autonomous System Border Router (ASBR) procedures for multihoming of
   Network Virtualization Edge (NVE) devices.

   In the context of this document, an NVO is a solution to address the
   requirements of a multi-tenant data center, especially one with
   virtualized hosts, e.g., Virtual Machines (VMs) or virtual workloads.
   The key requirements of such a solution, as described in [RFC7364],
   are the following:

   - Isolation of network traffic per tenant

   - Support for a large number of tenants (tens or hundreds of
      thousands)

   - Extension of Layer 2 (L2) connectivity among different VMs
      belonging to a given tenant segment (subnet) across different
      Points of Delivery (PoDs) within a data center or between
      different data centers

   - Allowing a given VM to move between different physical points of
      attachment within a given L2 segment

   The underlay network for NVO solutions is assumed to provide IP
   connectivity between NVO endpoints.

This document describes how EVPN can be used as an NVO solution and
explores applicability of EVPN functions and procedures. In
particular, it describes the various tunnel encapsulation options for
EVPN over IP and their impact on the EVPN control plane as well as
procedures for two main scenarios:

(a) single-homing NVEs - when an NVE resides in the hypervisor, and

(b) multihoming NVEs - when an NVE resides in a Top-of-Rack (ToR)
     device.

The possible encapsulation options for EVPN overlays that are
analyzed in this document are:

- VXLAN and NVGRE

- MPLS over GRE

Before getting into the description of the different encapsulation
options for EVPN over IP, it is important to highlight the EVPN
solution's main features, how those features are currently supported,
and any impact that the encapsulation has on those features.

2. Requirements Notation and Conventions

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

3. Terminology

Most of the terminology used in this documents comes from [RFC7432]
and [RFC7365].

VXLAN: Virtual Extensible LAN

GRE: Generic Routing Encapsulation

NVGRE: Network Virtualization using Generic Routing Encapsulation

GENEVE: Generic Network Virtualization Encapsulation

PoD: Point of Delivery

NV: Network Virtualization

   NVO: Network Virtualization Overlay

   NVE: Network Virtualization Edge

   VNI: VXLAN Network Identifier

   VSID: Virtual Subnet Identifier (for NVGRE)

   I-SID: Service Instance Identifier

   EVPN: Ethernet VPN

   EVI: EVPN Instance. An EVPN instance spanning the Provider Edge
      (PE) devices participating in that EVPN

   MAC-VRF: A Virtual Routing and Forwarding table for Media Access
      Control (MAC) addresses on a PE

   IP-VRF: A Virtual Routing and Forwarding table for Internet Protocol
      (IP) addresses on a PE

   ES: Ethernet Segment. When a customer site (device or network) is
      connected to one or more PEs via a set of Ethernet links, then
      that set of links is referred to as an 'Ethernet segment'.

   Ethernet Segment Identifier (ESI): A unique non-zero identifier that
      identifies an Ethernet segment is called an 'Ethernet Segment
      Identifier'.

   Ethernet Tag: An Ethernet tag identifies a particular broadcast
      domain, e.g., a VLAN. An EVPN instance consists of one or more
      broadcast domains.

   PE: Provider Edge

   Single-Active Redundancy Mode: When only a single PE, among all the
      PEs attached to an ES, is allowed to forward traffic to/from that
      ES for a given VLAN, then the Ethernet segment is defined to be
      operating in Single-Active redundancy mode.

   All-Active Redundancy Mode: When all PEs attached to an Ethernet
      segment are allowed to forward known unicast traffic to/from that
      ES for a given VLAN, then the ES is defined to be operating in
      All-Active redundancy mode.

   PIM-SM: Protocol Independent Multicast - Sparse-Mode

   PIM-SSM: Protocol Independent Multicast - Source-Specific Multicast

   BIDIR-PIM: Bidirectional PIM

4. EVPN Features

   EVPN [RFC7432] was originally designed to support the requirements
   detailed in [RFC7209] and therefore has the following attributes
   which directly address control-plane scaling and ease of deployment
   issues.

   1.   Control-plane information is distributed with BGP and broadcast
         and multicast traffic is sent using a shared multicast tree or
         with ingress replication.

   2.   Control-plane learning is used for MAC (and IP) addresses
         instead of data-plane learning. The latter requires the
         flooding of unknown unicast and Address Resolution Protocol
         (ARP) frames; whereas, the former does not require any flooding.

   3.   Route Reflector (RR) is used to reduce a full mesh of BGP
         sessions among PE devices to a single BGP session between a PE
         and the RR. Furthermore, RR hierarchy can be leveraged to scale
         the number of BGP routes on the RR.

   4.   Auto-discovery via BGP is used to discover PE devices
         participating in a given VPN, PE devices participating in a
         given redundancy group, tunnel encapsulation types, multicast
         tunnel types, multicast members, etc.

   5.   All-Active multihoming is used. This allows a given Customer
         Edge (CE) device to have multiple links to multiple PEs, and
         traffic to/from that CE fully utilizes all of these links.

   6.   When a link between a CE and a PE fails, the PEs for that EVI
         are notified of the failure via the withdrawal of a single EVPN
         route. This allows those PEs to remove the withdrawing PE as a
         next hop for every MAC address associated with the failed link.
         This is termed "mass withdrawal".

   7.   BGP route filtering and constrained route distribution are
         leveraged to ensure that the control-plane traffic for a given
         EVI is only distributed to the PEs in that EVI.

   8.   When an IEEE 802.1Q [IEEE.802.1Q] interface is used between a CE
        and a PE, each of the VLAN IDs (VIDs) on that interface can be
        mapped onto a bridge table (for up to 4094 such bridge tables).
        All these bridge tables may be mapped onto a single MAC-VRF (in
        case of VLAN-aware bundle service).

   9.   VM Mobility mechanisms ensure that all PEs in a given EVI know
        the ES with which a given VM, as identified by its MAC and IP
        addresses, is currently associated.

   10. RTs are used to allow the operator (or customer) to define a
        spectrum of logical network topologies including mesh, hub and
        spoke, and extranets (e.g., a VPN whose sites are owned by
        different enterprises), without the need for proprietary
        software or the aid of other virtual or physical devices.

   Because the design goal for NVO is millions of instances per common
   physical infrastructure, the scaling properties of the control plane
   for NVO are extremely important. EVPN and the extensions described
   herein, are designed with this level of scalability in mind.

5. Encapsulation Options for EVPN Overlays

5.1. VXLAN/NVGRE Encapsulation

   Both VXLAN and NVGRE are examples of technologies that provide a data
   plane encapsulation which is used to transport a packet over the
   common physical IP infrastructure between Network Virtualization
   Edges (NVEs) - e.g., VXLAN Tunnel End Points (VTEPs) in VXLAN
   network. Both of these technologies include the identifier of the
   specific NVO instance, VNI in VXLAN and VSID in NVGRE, in each
   packet. In the remainder of this document we use VNI as the
   representation for NVO instance with the understanding that VSID can
   equally be used if the encapsulation is NVGRE unless it is stated
   otherwise.

   Note that a PE is equivalent to an NVE/VTEP.

   VXLAN encapsulation is based on UDP, with an 8-byte header following
   the UDP header. VXLAN provides a 24-bit VNI, which typically
   provides a one-to-one mapping to the tenant VID, as described in
   [RFC7348]. In this scenario, the ingress VTEP does not include an
   inner VLAN tag on the encapsulated frame, and the egress VTEP
   discards the frames with an inner VLAN tag. This mode of operation
   in [RFC7348] maps to VLAN-Based Service in [RFC7432], where a tenant
   VID gets mapped to an EVI.

VXLAN also provides an option of including an inner VLAN tag in the
encapsulated frame, if explicitly configured at the VTEP. This mode
of operation can map to VLAN Bundle Service in [RFC7432] because all
the tenant's tagged frames map to a single bridge table / MAC-VRF,
and the inner VLAN tag is not used for lookup by the disposition PE
when performing VXLAN decapsulation as described in Section 6 of
[RFC7348].

[RFC7637] encapsulation is based on GRE encapsulation, and it
mandates the inclusion of the optional GRE Key field, which carries
the VSID. There is a one-to-one mapping between the VSID and the
tenant VID, as described in [RFC7637]. The inclusion of an inner
VLAN tag is prohibited. This mode of operation in [RFC7637] maps to
VLAN Based Service in [RFC7432].

As described in the next section, there is no change to the encoding
of EVPN routes to support VXLAN or NVGRE encapsulation, except for
the use of the BGP Encapsulation Extended Community to indicate the
encapsulation type (e.g., VXLAN or NVGRE). However, there is
potential impact to the EVPN procedures depending on where the NVE is
located (i.e., in hypervisor or ToR) and whether multihoming
capabilities are required.

5.1.1. Virtual Identifiers Scope

Although VNIs are defined as 24-bit globally unique values, there are
scenarios in which it is desirable to use a locally significant value
for the VNI, especially in the context of a data-center interconnect.

5.1.1.1. Data-Center Interconnect with Gateway

In the case where NVEs in different data centers need to be
interconnected, and the NVEs need to use VNIs as globally unique
identifiers within a data center, then a Gateway (GW) needs to be
employed at the edge of the data-center network (DCN). This is
because the Gateway will provide the functionality of translating the
VNI when crossing network boundaries, which may align with operator
span-of-control boundaries. As an example, consider the network of
Figure 1. Assume there are three network operators: one for each of
the DC1, DC2, and WAN networks. The Gateways at the edge of the data
centers are responsible for translating the VNIs between the values
used in each of the DCNs and the values used in the WAN.

```
                         +--------------+
                         |              |
             +---------+ |     WAN      |     +---------+
    +----+   |         +---+ +----+     +----+ +---+    |    +----+
    |NVE1|--|         |   | |WAN |     |WAN | |   |    |--|NVE3|
    +----+   |IP       |GW |--|Edge|     |Edge|--|GW | IP |    +----+
    +----+   |Fabric   +---+ +----+     +----+  +---+ Fabric |    +----+
    |NVE2|--|         |       |           |       |    |--|NVE4|
    +----+   +---------+       +-------------+       +---------+ +----+

    |<------ DC 1 ------>                    <------ DC2  ------>|
```

                 Figure 1: Data-Center Interconnect with Gateway

5.1.1.2. Data-Center Interconnect without Gateway

   In the case where NVEs in different data centers need to be
   interconnected, and the NVEs need to use locally assigned VNIs (e.g.,
   similar to MPLS labels), there may be no need to employ Gateways at
   the edge of the DCN. More specifically, the VNI value that is used
   by the transmitting NVE is allocated by the NVE that is receiving the
   traffic (in other words, this is similar to a "downstream-assigned"
   MPLS label). This allows the VNI space to be decoupled between
   different DCNs without the need for a dedicated Gateway at the edge
   of the data centers. This topic is covered in Section 10.2.

```
                         +--------------+
                         |              |
             +---------+ |     WAN      |     +---------+
    +----+   |         | +----+     +----+ |         |    +----+
    |NVE1|--|         | |ASBR|     |ASBR| |         |--|NVE3|
    +----+   |IP Fabric|---|    |     |    |--|IP Fabric|    +----+
    +----+   |         | +----+     +----+ |         |    +----+
    |NVE2|--|         | |    |     |    | |         |--|NVE4|
    +----+   +---------+ +-------------+   +---------+ +----+

    |<------ DC 1 ----->                    <---- DC2  ------>|
```

                 Figure 2: Data-Center Interconnect with ASBR

5.1.2. Virtual Identifiers to EVI Mapping

   Just like in [RFC7432], where two options existed for mapping
   broadcast domains (represented by VLAN IDs) to an EVI, when the EVPN
   control plane is used in conjunction with VXLAN (or NVGRE
   encapsulation), there are also two options for mapping broadcast
   domains represented by VXLAN VNIs (or NVGRE VSIDs) to an EVI:

      Option 1: A Single Broadcast Domain per EVI

   In this option, a single Ethernet broadcast domain (e.g., subnet)
   represented by a VNI is mapped to a unique EVI. This corresponds to
   the VLAN-Based Service in [RFC7432], where a tenant-facing interface,
   logical interface (e.g., represented by a VID), or physical interface
   gets mapped to an EVI. As such, a BGP Route Distinguisher (RD) and
   Route Target (RT) are needed per VNI on every NVE. The advantage of
   this model is that it allows the BGP RT constraint mechanisms to be
   used in order to limit the propagation and import of routes to only
   the NVEs that are interested in a given VNI. The disadvantage of
   this model may be the provisioning overhead if the RD and RT are not
   derived automatically from the VNI.

   In this option, the MAC-VRF table is identified by the RT in the
   control plane and by the VNI in the data plane. In this option, the
   specific MAC-VRF table corresponds to only a single bridge table.

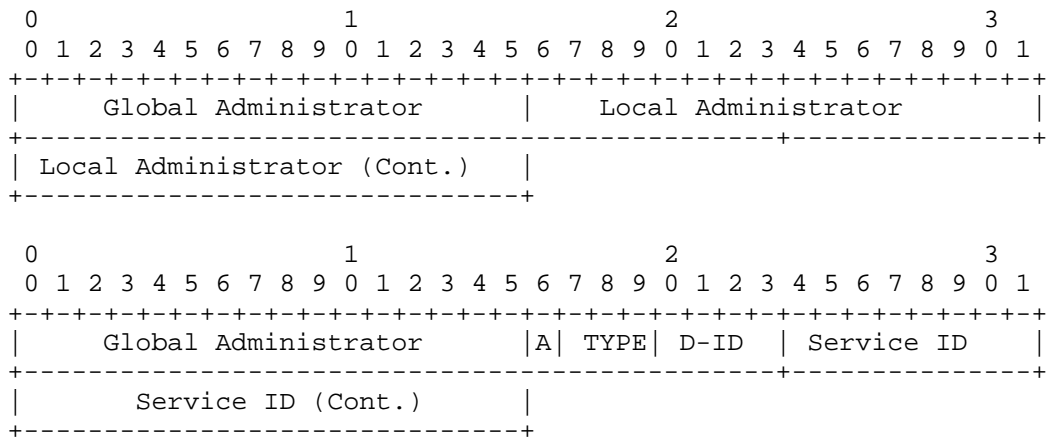   Option 2: Multiple Broadcast Domains per EVI

   In this option, multiple subnets, each represented by a unique VNI,
   are mapped to a single EVI. For example, if a tenant has multiple
   segments/subnets each represented by a VNI, then all the VNIs for
   that tenant are mapped to a single EVI; for example, the EVI in this
   case represents the tenant and not a subnet. This corresponds to the
   VLAN-aware bundle service in [RFC7432]. The advantage of this model
   is that it doesn't require the provisioning of an RD/RT per VNI.
   However, this is a moot point when compared to Option 1 where auto-
   derivation is used. The disadvantage of this model is that routes
   would be imported by NVEs that may not be interested in a given VNI.

   In this option, the MAC-VRF table is identified by the RT in the
   control plane; a specific bridge table for that MAC-VRF is identified
   by the <RT, Ethernet Tag ID> in the control plane. In this option,
   the VNI in the data plane is sufficient to identify a specific bridge
   table.

5.1.2.1. Auto-Derivation of RT

   In order to simplify configuration, when the option of a single VNI
   per EVI is used, the RT used for EVPN can be auto-derived. RD can be
   auto-generated as described in [RFC7432], and RT can be auto-derived
   as described next.

   Since a Gateway PE as depicted in Figure 1 participates in both the
   DCN and WAN BGP sessions, it is important that, when RT values are
   auto-derived from VNIs, there be no conflict in RT spaces between
   DCNs and WANs, assuming that both are operating within the same
   Autonomous System (AS). Also, there can be scenarios where both
   VXLAN and NVGRE encapsulations may be needed within the same DCN, and
   their corresponding VNIs are administered independently, which means
   VNI spaces can overlap. In order to avoid conflict in RT spaces, the
   6-byte RT values with 2-octet AS number for DCNs can be auto-derived
   as follow:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |      Global Administrator         |    Local Administrator    |
   +-----------------------------------------------+---------------+
   | Local Administrator (Cont.)   |
   +-------------------------------+


    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |      Global Administrator     |A| TYPE| D-ID  | Service ID    |
   +-----------------------------------------------+---------------+
   |       Service ID (Cont.)      |
   +-------------------------------+
```

   The 6-octet RT field consists of two sub-fields:

   - Global Administrator sub-field: 2 octets. This sub-field contains
      an AS number assigned by IANA <https://www.iana.org/assignments/
      as-numbers/>.

   - Local Administrator sub-field: 4 octets

      * A: A single-bit field indicating if this RT is auto-derived

            0: auto-derived
            1: manually derived

   * Type: A 3-bit field that identifies the space in which the
      other 3 bytes are defined. The following spaces are defined:

         0 : VID (802.1Q VLAN ID)
         1 : VXLAN
         2 : NVGRE
         3 : I-SID
         4 : EVI
         5 : dual-VID (QinQ VLAN ID)

   * D-ID: A 4-bit field that identifies domain-id. The default
      value of domain-id is zero, indicating that only a single
      numbering space exist for a given technology. However, if more
      than one number space exists for a given technology (e.g.,
      overlapping VXLAN spaces), then each of the number spaces need
      to be identified by its corresponding domain-id starting from
      1.

   * Service ID: This 3-octet field is set to VNI, VSID, I-SID, or
      VID.

   It should be noted that RT auto-derivation is applicable for 2-octet
   AS numbers. For 4-octet AS numbers, the RT needs to be manually
   configured because 3-octet VNI fields cannot be fit within the
   2-octet local administrator field.

5.1.3. Constructing EVPN BGP Routes

   In EVPN, an MPLS label, for instance, identifying the forwarding
   table is distributed by the egress PE via the EVPN control plane and
   is placed in the MPLS header of a given packet by the ingress PE.
   This label is used upon receipt of that packet by the egress PE for
   disposition of that packet. This is very similar to the use of the
   VNI by the egress NVE, with the difference being that an MPLS label
   has local significance while a VNI typically has global significance.
   Accordingly, and specifically to support the option of locally
   assigned VNIs, the MPLS Label1 field in the MAC/IP Advertisement
   route, the MPLS label field in the Ethernet A-D per EVI route, and
   the MPLS label field in the P-Multicast Service Interface (PMSI)
   Tunnel attribute of the Inclusive Multicast Ethernet Tag (IMET) route
   are used to carry the VNI. For the balance of this memo, the above
   MPLS label fields will be referred to as the VNI field. The VNI
   field is used for both local and global VNIs; for either case, the
   entire 24-bit field is used to encode the VNI value.

For the VLAN-Based Service (a single VNI per MAC-VRF), the Ethernet
Tag field in the MAC/IP Advertisement, Ethernet A-D per EVI, and IMET
route MUST be set to zero just as in the VLAN-Based Service in
[RFC7432].

For the VLAN-Aware Bundle Service (multiple VNIs per MAC-VRF with
each VNI associated with its own bridge table), the Ethernet Tag
field in the MAC Advertisement, Ethernet A-D per EVI, and IMET route
MUST identify a bridge table within a MAC-VRF; the set of Ethernet
Tags for that EVI needs to be configured consistently on all PEs
within that EVI. For locally assigned VNIs, the value advertised in
the Ethernet Tag field MUST be set to a VID just as in the VLAN-aware
bundle service in [RFC7432]. Such setting must be done consistently
on all PE devices participating in that EVI within a given domain.
For global VNIs, the value advertised in the Ethernet Tag field
SHOULD be set to a VNI as long as it matches the existing semantics
of the Ethernet Tag, i.e., it identifies a bridge table within a
MAC-VRF and the set of VNIs are configured consistently on each PE in
that EVI.

In order to indicate which type of data-plane encapsulation (i.e.,
VXLAN, NVGRE, MPLS, or MPLS in GRE) is to be used, the BGP
Encapsulation Extended Community defined in [RFC5512] is included
with all EVPN routes (i.e., MAC Advertisement, Ethernet A-D per EVI,
Ethernet A-D per ESI, IMET, and Ethernet Segment) advertised by an
egress PE. Five new values have been assigned by IANA to extend the
list of encapsulation types defined in [RFC5512]; they are listed in
Section 11.

The MPLS encapsulation tunnel type, listed in Section 11, is needed
in order to distinguish between an advertising node that only
supports non-MPLS encapsulations and one that supports MPLS and
non-MPLS encapsulations. An advertising node that only supports MPLS
encapsulation does not need to advertise any encapsulation tunnel
types; i.e., if the BGP Encapsulation Extended Community is not
present, then either MPLS encapsulation or a statically configured
encapsulation is assumed.

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST
be set to the IPv4 or IPv6 address of the NVE. The remaining fields
in each route are set as per [RFC7432].

Note that the procedure defined here -- to use the MPLS Label field
to carry the VNI in the presence of a Tunnel Encapsulation Extended
Community specifying the use of a VNI -- is aligned with the
procedures described in Section 8.2.2.2 of [TUNNEL-ENCAP] ("When a
Valid VNI has not been Signaled").

5.2. MPLS over GRE

   The EVPN data plane is modeled as an EVPN MPLS client layer sitting
   over an MPLS PSN tunnel server layer. Some of the EVPN functions
   (split-horizon, Aliasing, and Backup Path) are tied to the MPLS
   client layer. If MPLS over GRE encapsulation is used, then the EVPN
   MPLS client layer can be carried over an IP PSN tunnel transparently.
   Therefore, there is no impact to the EVPN procedures and associated
   data-plane operation.

   [RFC4023] defines the standard for using MPLS over GRE encapsulation,
   which can be used for this purpose. However, when MPLS over GRE is
   used in conjunction with EVPN, it is recommended that the GRE key
   field be present and be used to provide a 32-bit entropy value only
   if the P nodes can perform Equal-Cost Multipath (ECMP) hashing based
   on the GRE key; otherwise, the GRE header SHOULD NOT include the GRE
   key field. The Checksum and Sequence Number fields MUST NOT be
   included, and the corresponding C and S bits in the GRE header MUST
   be set to zero. A PE capable of supporting this encapsulation SHOULD
   advertise its EVPN routes along with the Tunnel Encapsulation
   Extended Community indicating MPLS over GRE encapsulation as
   described in the previous section.

6. EVPN with Multiple Data-Plane Encapsulations

   The use of the BGP Encapsulation Extended Community per [RFC5512]
   allows each NVE in a given EVI to know each of the encapsulations
   supported by each of the other NVEs in that EVI. That is, each of
   the NVEs in a given EVI may support multiple data-plane
   encapsulations. An ingress NVE can send a frame to an egress NVE
   only if the set of encapsulations advertised by the egress NVE forms
   a non-empty intersection with the set of encapsulations supported by
   the ingress NVE; it is at the discretion of the ingress NVE which
   encapsulation to choose from this intersection. (As noted in
   Section 5.1.3, if the BGP Encapsulation extended community is not
   present, then the default MPLS encapsulation or a locally configured
   encapsulation is assumed.)

   When a PE advertises multiple supported encapsulations, it MUST
   advertise encapsulations that use the same EVPN procedures including
   procedures associated with split-horizon filtering described in
   Section 8.3.1. For example, VXLAN and NVGRE (or MPLS and MPLS over
   GRE) encapsulations use the same EVPN procedures; thus, a PE can
   advertise both of them and can support either of them or both of them
   simultaneously. However, a PE MUST NOT advertise VXLAN and MPLS
   encapsulations together because (a) the MPLS field of EVPN routes is

set to either an MPLS label or a VNI, but not both and (b) some EVPN
procedures (such as split-horizon filtering) are different for VXLAN/
NVGRE and MPLS encapsulations.

An ingress node that uses shared multicast trees for sending
broadcast or multicast frames MAY maintain distinct trees for each
different encapsulation type.

It is the responsibility of the operator of a given EVI to ensure
that all of the NVEs in that EVI support at least one common
encapsulation. If this condition is violated, it could result in
service disruption or failure. The use of the BGP Encapsulation
Extended Community provides a method to detect when this condition is
violated, but the actions to be taken are at the discretion of the
operator and are outside the scope of this document.

7. Single-Homing NVEs - NVE Residing in Hypervisor

When an NVE and its hosts/VMs are co-located in the same physical
device, e.g., when they reside in a server, the links between them
are virtual and they typically share fate. That is, the subject
hosts/VMs are typically not multihomed or, if they are multihomed,
the multihoming is a purely local matter to the server hosting the VM
and the NVEs, and it need not be "visible" to any other NVEs residing
on other servers. Thus, it does not require any specific protocol
mechanisms. The most common case of this is when the NVE resides on
the hypervisor.

In the subsections that follow, we will discuss the impact on EVPN
procedures for the case when the NVE resides on the hypervisor and
the VXLAN (or NVGRE) encapsulation is used.

7.1. Impact on EVPN BGP Routes & Attributes for VXLAN/NVGRE
      Encapsulations

In scenarios where different groups of data centers are under
different administrative domains, and these data centers are
connected via one or more backbone core providers as described in
[RFC7365], the RD must be a unique value per EVI or per NVE as
described in [RFC7432]. In other words, whenever there is more than
one administrative domain for global VNI, a unique RD must be used;
or, whenever the VNI value has local significance, a unique RD must
be used. Therefore, it is recommended to use a unique RD as
described in [RFC7432] at all times.

When the NVEs reside on the hypervisor, the EVPN BGP routes and
attributes associated with multihoming are no longer required. This
reduces the required routes and attributes to the following subset of
four out of the total of eight listed in Section 7 of [RFC7432]:

- MAC/IP Advertisement Route

- Inclusive Multicast Ethernet Tag Route

- MAC Mobility Extended Community

- Default Gateway Extended Community

However, as noted in Section 8.6 of [RFC7432], in order to enable a
single-homing ingress NVE to take advantage of fast convergence,
Aliasing, and Backup Path when interacting with multihomed egress
NVEs attached to a given ES, the single-homing ingress NVE should be
able to receive and process routes that are Ethernet A-D per ES and
Ethernet A-D per EVI.

7.2. Impact on EVPN Procedures for VXLAN/NVGRE Encapsulations

When the NVEs reside on the hypervisors, the EVPN procedures
associated with multihoming are no longer required. This limits the
procedures on the NVE to the following subset.

   1. Local learning of MAC addresses received from the VMs per
       Section 10.1 of [RFC7432].

   2. Advertising locally learned MAC addresses in BGP using the MAC/IP
       Advertisement routes.

   3. Performing remote learning using BGP per Section 9.2 of
       [RFC7432].

   4. Discovering other NVEs and constructing the multicast tunnels
       using the IMET routes.

   5. Handling MAC address mobility events per the procedures of
       Section 15 in [RFC7432].

However, as noted in Section 8.6 of [RFC7432], in order to enable a
single-homing ingress NVE to take advantage of fast convergence,
Aliasing, and Backup Path when interacting with multihomed egress
NVEs attached to a given ES, a single-homing ingress NVE should
implement the ingress node processing of routes that are Ethernet A-D
per ES and Ethernet A-D per EVI as defined in Sections 8.2 ("Fast
Convergence") and 8.4 ("Aliasing and Backup Path") of [RFC7432].

8. Multihoming NVEs - NVE Residing in ToR Switch

   In this section, we discuss the scenario where the NVEs reside in the
   ToR switches AND the servers (where VMs are residing) are multihomed
   to these ToR switches. The multihoming NVE operates in All-Active or
   Single-Active redundancy mode. If the servers are single-homed to
   the ToR switches, then the scenario becomes similar to that where the
   NVE resides on the hypervisor, as discussed in Section 7, as far as
   the required EVPN functionality is concerned.

   [RFC7432] defines a set of BGP routes, attributes, and procedures to
   support multihoming. We first describe these functions and
   procedures, then discuss which of these are impacted by the VXLAN (or
   NVGRE) encapsulation and what modifications are required. As will be
   seen later in this section, the only EVPN procedure that is impacted
   by non-MPLS overlay encapsulation (e.g., VXLAN or NVGRE) where it
   provides space for one ID rather than a stack of labels, is that of
   split-horizon filtering for multihomed ESs described in
   Section 8.3.1.

8.1. EVPN Multihoming Features

   In this section, we will recap the multihoming features of EVPN to
   highlight the encapsulation dependencies. The section only describes
   the features and functions at a high level. For more details, the
   reader is to refer to [RFC7432].

8.1.1. Multihomed ES Auto-Discovery

   EVPN NVEs (or PEs) connected to the same ES (e.g., the same server
   via Link Aggregation Group (LAG)) can automatically discover each
   other with minimal to no configuration through the exchange of BGP
   routes.

8.1.2. Fast Convergence and Mass Withdrawal

   EVPN defines a mechanism to efficiently and quickly signal, to remote
   NVEs, the need to update their forwarding tables upon the occurrence
   of a failure in connectivity to an ES (e.g., a link or a port
   failure). This is done by having each NVE advertise an Ethernet A-D
   route per ES for each locally attached segment. Upon a failure in
   connectivity to the attached segment, the NVE withdraws the
   corresponding Ethernet A-D route. This triggers all NVEs that
   receive the withdrawal to update their next-hop adjacencies for all
   MAC addresses associated with the ES in question. If no other NVE
   had advertised an Ethernet A-D route for the same segment, then the

NVE that received the withdrawal simply invalidates the MAC entries
for that segment. Otherwise, the NVE updates the next-hop adjacency
list accordingly.

8.1.3. Split-Horizon

If a server is multihomed to two or more NVEs (represented by an ES
ES1) and operating in an All-Active redundancy mode, sends a BUM
(i.e., Broadcast, Unknown unicast, or Multicast) packet to one of
these NVEs, then it is important to ensure the packet is not looped
back to the server via another NVE connected to this server. The
filtering mechanism on the NVE to prevent such loop and packet
duplication is called "split-horizon filtering".

8.1.4. Aliasing and Backup Path

In the case where a station is multihomed to multiple NVEs, it is
possible that only a single NVE learns a set of the MAC addresses
associated with traffic transmitted by the station. This leads to a
situation where remote NVEs receive MAC Advertisement routes, for
these addresses, from a single NVE even though multiple NVEs are
connected to the multihomed station. As a result, the remote NVEs
are not able to effectively load-balance traffic among the NVEs
connected to the multihomed ES. For example, this could be the case
when the NVEs perform data-path learning on the access and the load-
balancing function on the station hashes traffic from a given source
MAC address to a single NVE. Another scenario where this occurs is
when the NVEs rely on control-plane learning on the access (e.g.,
using ARP), since ARP traffic will be hashed to a single link in the
LAG.

To alleviate this issue, EVPN introduces the concept of "Aliasing".
This refers to the ability of an NVE to signal that it has
reachability to a given locally attached ES, even when it has learned
no MAC addresses from that segment. The Ethernet A-D route per EVI
is used to that end. Remote NVEs that receive MAC Advertisement
routes with non-zero ESIs should consider the MAC address as
reachable via all NVEs that advertise reachability to the relevant
Segment using Ethernet A-D routes with the same ESI and with the
Single-Active flag reset.

Backup Path is a closely related function, albeit one that applies to
the case where the redundancy mode is Single-Active. In this case,
the NVE signals that it has reachability to a given locally attached
ES using the Ethernet A-D route as well. Remote NVEs that receive
the MAC Advertisement routes, with non-zero ESI, should consider the
MAC address as reachable via the advertising NVE. Furthermore, the
remote NVEs should install a Backup Path, for said MAC, to the NVE

that had advertised reachability to the relevant segment using an
Ethernet A-D route with the same ESI and with the Single-Active flag
set.

8.1.5. DF Election

   If a host is multihomed to two or more NVEs on an ES operating in
   All-Active redundancy mode, then, for a given EVI, only one of these
   NVEs, termed the "Designated Forwarder" (DF) is responsible for
   sending it broadcast, multicast, and, if configured for that EVI,
   unknown unicast frames.

   This is required in order to prevent duplicate delivery of multi-
   destination frames to a multihomed host or VM, in case of All-Active
   redundancy.

   In NVEs where frames tagged as IEEE 802.1Q [IEEE.802.1Q] are received
   from hosts, the DF election should be performed based on host VIDs
   per Section 8.5 of [RFC7432]. Furthermore, multihoming PEs of a
   given ES MAY perform DF election using configured IDs such as VNI,
   EVI, normalized VIDs, and etc., as along the IDs are configured
   consistently across the multihoming PEs.

   In GWs where VXLAN-encapsulated frames are received, the DF election
   is performed on VNIs. Again, it is assumed that, for a given
   Ethernet segment, VNIs are unique and consistent (e.g., no duplicate
   VNIs exist).

8.2. Impact on EVPN BGP Routes and Attributes

   Since multihoming is supported in this scenario, the entire set of
   BGP routes and attributes defined in [RFC7432] is used. The setting
   of the Ethernet Tag field in the MAC Advertisement, Ethernet A-D per
   EVI, and IMET) routes follows that of Section 5.1.3. Furthermore,
   the setting of the VNI field in the MAC Advertisement and Ethernet
   A-D per EVI routes follows that of Section 5.1.3.

8.3. Impact on EVPN Procedures

   Two cases need to be examined here, depending on whether the NVEs are
   operating in Single-Active or in All-Active redundancy mode.

   First, let's consider the case of Single-Active redundancy mode,
   where the hosts are multihomed to a set of NVEs; however, only a
   single NVE is active at a given point of time for a given VNI. In
   this case, the Aliasing is not required, and the split-horizon

filtering may not be required, but other functions such as multihomed
ES auto-discovery, fast convergence and mass withdrawal, Backup Path,
and DF election are required.

Second, let's consider the case of All-Active redundancy mode. In
this case, out of all the EVPN multihoming features listed in
Section 8.1, the use of the VXLAN or NVGRE encapsulation impacts the
split-horizon and Aliasing features, since those two rely on the MPLS
client layer. Given that this MPLS client layer is absent with these
types of encapsulations, alternative procedures and mechanisms are
needed to provide the required functions. Those are discussed in
detail next.

8.3.1. Split Horizon

In EVPN, an MPLS label is used for split-horizon filtering to support
All-Active multihoming where an ingress NVE adds a label
corresponding to the site of origin (aka an ESI label) when
encapsulating the packet. The egress NVE checks the ESI label when
attempting to forward a multi-destination frame out an interface, and
if the label corresponds to the same site identifier (ESI) associated
with that interface, the packet gets dropped. This prevents the
occurrence of forwarding loops.

Since VXLAN and NVGRE encapsulations do not include the ESI label,
other means of performing the split-horizon filtering function must
be devised for these encapsulations. The following approach is
recommended for split-horizon filtering when VXLAN (or NVGRE)
encapsulation is used.

Every NVE tracks the IP address(es) associated with the other NVE(s)
with which it has shared multihomed ESs. When the NVE receives a
multi-destination frame from the overlay network, it examines the
source IP address in the tunnel header (which corresponds to the
ingress NVE) and filters out the frame on all local interfaces
connected to ESs that are shared with the ingress NVE. With this
approach, it is required that the ingress NVE perform replication
locally to all directly attached Ethernet segments (regardless of the
DF election state) for all flooded traffic ingress from the access
interfaces (i.e., from the hosts). This approach is referred to as
"Local Bias", and has the advantage that only a single IP address
need be used per NVE for split-horizon filtering, as opposed to
requiring an IP address per Ethernet segment per NVE.

In order to allow proper operation of split-horizon filtering among
the same group of multihoming PE devices, a mix of PE devices with
MPLS over GRE encapsulations running the procedures from [RFC7432]

for split-horizon filtering on the one hand and VXLAN/NVGRE
encapsulation running local-bias procedures on the other on a given
Ethernet segment MUST NOT be configured.

8.3.2. Aliasing and Backup Path

The Aliasing and the Backup Path procedures for VXLAN/NVGRE
encapsulation are very similar to the ones for MPLS. In the case of
MPLS, Ethernet A-D route per EVI is used for Aliasing when the
corresponding ES operates in All-Active multihoming, and the same
route is used for Backup Path when the corresponding ES operates in
Single-Active multihoming. In the case of VXLAN/NVGRE, the same
route is used for the Aliasing and the Backup Path with the
difference that the Ethernet Tag and VNI fields in Ethernet A-D per
EVI route are set as described in Section 5.1.3.

8.3.3. Unknown Unicast Traffic Designation

In EVPN, when an ingress PE uses ingress replication to flood unknown
unicast traffic to egress PEs, the ingress PE uses a different EVPN
MPLS label (from the one used for known unicast traffic) to identify
such BUM traffic. The egress PEs use this label to identify such BUM
traffic and, thus, apply DF filtering for All-Active multihomed
sites. In absence of an unknown unicast traffic designation and in
the presence of enabling unknown unicast flooding, there can be
transient duplicate traffic to All-Active multihomed sites under the
following condition: the host MAC address is learned by the egress
PE(s) and advertised to the ingress PE; however, the MAC
Advertisement has not been received or processed by the ingress PE,
resulting in the host MAC address being unknown on the ingress PE but
known on the egress PE(s). Therefore, when a packet destined to that
host MAC address arrives on the ingress PE, it floods it via ingress
replication to all the egress PE(s), and since they are known to the
egress PE(s), multiple copies are sent to the All-Active multihomed
site. It should be noted that such transient packet duplication only
happens when a) the destination host is multihomed via All-Active
redundancy mode, b) flooding of unknown unicast is enabled in the
network, c) ingress replication is used, and d) traffic for the
destination host is arrived on the ingress PE before it learns the
host MAC address via BGP EVPN advertisement. If it is desired to
avoid occurrence of such transient packet duplication (however low
probability that may be), then VXLAN-GPE encapsulation needs to be
used between these PEs and the ingress PE needs to set the BUM
Traffic Bit (B bit) [VXLAN-GPE] to indicate that this is an ingress-
replicated BUM traffic.

9. Support for Multicast

   The EVPN IMET route is used to discover the multicast tunnels among
   the endpoints associated with a given EVI (e.g., given VNI) for VLAN-
   Based Service and a given <EVI, VLAN> for VLAN-Aware Bundle Service.
   All fields of this route are set as described in Section 5.1.3. The
   originating router's IP address field is set to the NVE's IP address.
   This route is tagged with the PMSI Tunnel attribute, which is used to
   encode the type of multicast tunnel to be used as well as the
   multicast tunnel identifier. The tunnel encapsulation is encoded by
   adding the BGP Encapsulation Extended Community as per Section 5.1.1.
   For example, the PMSI Tunnel attribute may indicate the multicast
   tunnel is of type Protocol Independent Multicast - Sparse-Mode (PIM-
   SM); whereas, the BGP Encapsulation Extended Community may indicate
   the encapsulation for that tunnel is of type VXLAN. The following
   tunnel types as defined in [RFC6514] can be used in the PMSI Tunnel
   attribute for VXLAN/NVGRE:

          + 3 - PIM-SSM Tree
          + 4 - PIM-SM Tree
          + 5 - BIDIR-PIM Tree
          + 6 - Ingress Replication

   In case of VXLAN and NVGRE encapsulations with locally assigned VNIs,
   just as in [RFC7432], each PE MUST advertise an IMET route to other
   PEs in an EVPN instance for the multicast tunnel type that it uses
   (i.e., ingress replication, PIM-SM, PIM-SSM, or BIDIR-PIM tunnel).
   However, for globally assigned VNIs, each PE MUST advertise an IMET
   route to other PEs in an EVPN instance for ingress replication or a
   PIM-SSM tunnel, and they MAY advertise an IMET route for a PIM-SM or
   BIDIR-PIM tunnel. In case of a PIM-SM or BIDIR-PIM tunnel, no
   information in the IMET route is needed by the PE to set up these
   tunnels.

   In the scenario where the multicast tunnel is a tree, both the
   Inclusive as well as the Aggregate Inclusive variants may be used.
   In the former case, a multicast tree is dedicated to a VNI. Whereas,
   in the latter, a multicast tree is shared among multiple VNIs. For
   VNI-Based Service, the Aggregate Inclusive mode is accomplished by
   having the NVEs advertise multiple IMET routes with different RTs
   (one per VNI) but with the same tunnel identifier encoded in the PMSI
   Tunnel attribute. For VNI-Aware Bundle Service, the Aggregate
   Inclusive mode is accomplished by having the NVEs advertise multiple
   IMET routes with different VNIs encoded in the Ethernet Tag field,
   but with the same tunnel identifier encoded in the PMSI Tunnel
   attribute.

10. Data-Center Interconnections (DCIs)

   For DCIs, the following two main scenarios are considered when
   connecting data centers running evpn-overlay (as described here) over
   an MPLS/IP core network:

   - Scenario 1: DCI using GWs

   - Scenario 2: DCI using ASBRs

   The following two subsections describe the operations for each of
   these scenarios.

10.1. DCI Using GWs

   This is the typical scenario for interconnecting data centers over
   WAN. In this scenario, EVPN routes are terminated and processed in
   each GW and MAC/IP route are always re-advertised from DC to WAN but
   from WAN to DC, they are not re-advertised if unknown MAC addresses
   (and default IP address) are utilized in the NVEs. In this scenario,
   each GW maintains a MAC-VRF (and/or IP-VRF) for each EVI. The main
   advantage of this approach is that NVEs do not need to maintain MAC
   and IP addresses from any remote data centers when default IP routes
   and unknown MAC routes are used; that is, they only need to maintain
   routes that are local to their own DC. When default IP routes and
   unknown MAC routes are used, any unknown IP and MAC packets from NVEs
   are forwarded to the GWs where all the VPN MAC and IP routes are
   maintained. This approach reduces the size of MAC-VRF and IP-VRF
   significantly at NVEs. Furthermore, it results in a faster
   convergence time upon a link or NVE failure in a multihomed network
   or device redundancy scenario, because the failure-related BGP routes
   (such as mass withdrawal message) do not need to get propagated all
   the way to the remote NVEs in the remote DCs. This approach is
   described in detail in Section 3.4 of [DCI-EVPN-OVERLAY].

10.2. DCI Using ASBRs

   This approach can be considered as the opposite of the first
   approach. It favors simplification at DCI devices over NVEs such
   that larger MAC-VRF (and IP-VRF) tables need to be maintained on
   NVEs; whereas DCI devices don't need to maintain any MAC (and IP)
   forwarding tables. Furthermore, DCI devices do not need to terminate
   and process routes related to multihoming but rather to relay these
   messages for the establishment of an end-to-end Label Switched Path
   (LSP). In other words, DCI devices in this approach operate similar
   to ASBRs for inter-AS Option B (see Section 10 of [RFC4364]). This
   requires locally assigned VNIs to be used just like downstream-
   assigned MPLS VPN labels where, for all practical purposes, the VNIs

function like 24-bit VPN labels. This approach is equally applicable
to data centers (or Carrier Ethernet networks) with MPLS
encapsulation.

In inter-AS Option B, when ASBR receives an EVPN route from its DC
over internal BGP (iBGP) and re-advertises it to other ASBRs, it
re-advertises the EVPN route by re-writing the BGP next hops to
itself, thus losing the identity of the PE that originated the
advertisement. This rewrite of BGP next hop impacts the EVPN mass
withdrawal route (Ethernet A-D per ES) and its procedure adversely.
However, it does not impact the EVPN Aliasing mechanism/procedure
because when the Aliasing routes (Ethernet A-D per EVI) are
advertised, the receiving PE first resolves a MAC address for a given
EVI into its corresponding <ES, EVI>, and, subsequently, it resolves
the <ES, EVI> into multiple paths (and their associated next hops)
via which the <ES, EVI> is reachable. Since Aliasing and MAC routes
are both advertised on a per-EVI-basis and they use the same RD and
RT (per EVI), the receiving PE can associate them together on a
per-BGP-path basis (e.g., per originating PE). Thus, it can perform
recursive route resolution, e.g., a MAC is reachable via an <ES, EVI>
which in turn, is reachable via a set of BGP paths; thus, the MAC is
reachable via the set of BGP paths. Due to the per-EVI basis, the
association of MAC routes and the corresponding Aliasing route is
fixed and determined by the same RD and RT; there is no ambiguity
when the BGP next hop for these routes is rewritten as these routes
pass through ASBRs. That is, the receiving PE may receive multiple
Aliasing routes for the same EVI from a single next hop (a single
ASBR), and it can still create multiple paths toward that <ES, EVI>.

However, when the BGP next-hop address corresponding to the
originating PE is rewritten, the association between the mass
withdrawal route (Ethernet A-D per ES) and its corresponding MAC
routes cannot be made based on their RDs and RTs because the RD for
the mass Withdrawal route is different than the one for the MAC
routes. Therefore, the functionality needed at the ASBRs and the
receiving PEs depends on whether the Mass Withdrawal route is
originated and whether there is a need to handle route resolution
ambiguity for this route. The following two subsections describe the
functionality needed by the ASBRs and the receiving PEs depending on
whether the NVEs reside in a hypervisors or in ToR switches.

10.2.1. ASBR Functionality with Single-Homing NVEs

When NVEs reside in hypervisors as described in Section 7.1, there is
no multihoming; thus, there is no need for the originating NVE to
send Ethernet A-D per ES or Ethernet A-D per EVI routes. However, as
noted in Section 7, in order to enable a single-homing ingress NVE to
take advantage of fast convergence, Aliasing, and Backup Path when

interacting with multihoming egress NVEs attached to a given ES, the
single-homing NVE should be able to receive and process Ethernet A-D
per ES and Ethernet A-D per EVI routes. The handling of these routes
is described in the next section.

10.2.2. ASBR Functionality with Multihoming NVEs

When NVEs reside in ToR switches and operate in multihoming
redundancy mode, there is a need, as described in Section 8, for the
originating multihoming NVE to send Ethernet A-D per ES route(s)
(used for mass withdrawal) and Ethernet A-D per EVI routes (used for
Aliasing). As described above, the rewrite of BGP next hop by ASBRs
creates ambiguities when Ethernet A-D per ES routes are received by
the remote NVE in a different ASBR because the receiving NVE cannot
associate that route with the MAC/IP routes of that ES advertised by
the same originating NVE. This ambiguity inhibits the function of
mass withdrawal per ES by the receiving NVE in a different AS.

As an example, consider a scenario where a CE is multihomed to PE1
and PE2, where these PEs are connected via ASBR1 and then ASBR2 to
the remote PE3. Furthermore, consider that PE1 receives M1 from CE1
but not PE2. Therefore, PE1 advertises Ethernet A-D per ES1,
Ethernet A-D per EVI1, and M1; whereas, PE2 only advertises Ethernet
A-D per ES1 and Ethernet A-D per EVI1. ASBR1 receives all these five
advertisements and passes them to ASBR2 (with itself as the BGP next
hop). ASBR2, in turn, passes them to the remote PE3, with itself as
the BGP next hop. PE3 receives these five routes where all of them
have the same BGP next hop (i.e., ASBR2). Furthermore, the two
Ethernet A-D per ES routes received by PE3 have the same information,
i.e., same ESI and the same BGP next hop. Although both of these
routes are maintained by the BGP process in PE3 (because they have
different RDs and, thus, are treated as different BGP routes),
information from only one of them is used in the L2 routing table (L2
RIB).

```
                     PE1
                    /   \
                 CE      ASBR1---ASBR2---PE3
                    \   /
                     PE2
```
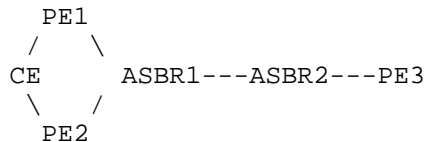
                    Figure 3: Inter-AS Option B

Now, when the AC between the PE2 and the CE fails and PE2 sends
Network Layer Reachability Information (NLRI) withdrawal for Ethernet
A-D per ES route, and this withdrawal gets propagated and received by
the PE3, the BGP process in PE3 removes the corresponding BGP route;
however, it doesn't remove the associated information (namely ESI and

BGP next hop) from the L2 routing table (L2 RIB) because it still has
the other Ethernet A-D per ES route (originated from PE1) with the
same information. That is why the mass withdrawal mechanism does not
work when doing DCI with inter-AS Option B. However, as described
previously, the Aliasing function works and so does "mass withdrawal
per EVI" (which is associated with withdrawing the EVPN route
associated with Aliasing, i.e., Ethernet A-D per EVI route).

In the above example, the PE3 receives two Aliasing routes with the
same BGP next hop (ASBR2) but different RDs. One of the Aliasing
route has the same RD as the advertised MAC route (M1). PE3 follows
the route resolution procedure specified in [RFC7432] upon receiving
the two Aliasing routes; that is, it resolves M1 to <ES, EVI1>, and,
subsequently, it resolves <ES, EVI1> to a BGP path list with two
paths along with the corresponding VNIs/MPLS labels (one associated
with PE1 and the other associated with PE2). It should be noted that
even though both paths are advertised by the same BGP next hop
(ASRB2), the receiving PE3 can handle them properly. Therefore, M1
is reachable via two paths. This creates two end-to-end LSPs, from
PE3 to PE1 and from PE3 to PE2, for M1 such that when PE3 wants to
forward traffic destined to M1, it can load-balance between the two
LSPs. Although route resolution for Aliasing routes with the same
BGP next hop is not explicitly mentioned in [RFC7432], this is the
expected operation; thus, it is elaborated here.

When the AC between the PE2 and the CE fails and PE2 sends NLRI
withdrawal for Ethernet A-D per EVI routes, and these withdrawals get
propagated and received by the PE3, the PE3 removes the Aliasing
route and updates the path list; that is, it removes the path
corresponding to the PE2. Therefore, all the corresponding MAC
routes for that <ES, EVI> that point to that path list will now have
the updated path list with a single path associated with PE1. This
action can be considered to be the mass withdrawal at the per-EVI
level. The mass withdrawal at the per-EVI level has a longer
convergence time than the mass withdrawal at the per-ES level;
however, it is much faster than the convergence time when the
withdrawal is done on a per-MAC basis.

If a PE becomes detached from a given ES, then, in addition to
withdrawing its previously advertised Ethernet A-D per ES routes, it
MUST also withdraw its previously advertised Ethernet A-D per EVI
routes for that ES. For a remote PE that is separated from the
withdrawing PE by one or more EVPN inter-AS Option B ASBRs, the
withdrawal of the Ethernet A-D per ES routes is not actionable.
However, a remote PE is able to correlate a previously advertised
Ethernet A-D per EVI route with any MAC/IP Advertisement routes also
advertised by the withdrawing PE for that <ES, EVI, BD>. Hence, when

it receives the withdrawal of an Ethernet A-D per EVI route, it
SHOULD remove the withdrawing PE as a next hop for all MAC addresses
associated with that <ES, EVI, BD>.

In the previous example, when the AC between PE2 and the CE fails,
PE2 will withdraw its Ethernet A-D per ES and per EVI routes. When
PE3 receives the withdrawal of an Ethernet A-D per EVI route, it
removes PE2 as a valid next hop for all MAC addresses associated with
the corresponding <ES, EVI, BD>. Therefore, all the MAC next hops
for that <ES, EVI, BD> will now have a single next hop, viz. the LSP
to PE1.

In summary, it can be seen that Aliasing (and Backup Path)
functionality should work as is for inter-AS Option B without
requiring any additional functionality in ASBRs or PEs. However, the
mass withdrawal functionality falls back from per-ES mode to per-EVI
mode for inter-AS Option B. That is, PEs receiving a mass withdrawal
route from the same AS take action on Ethernet A-D per ES route;
whereas, PEs receiving mass withdrawal routes from different ASes
take action on the Ethernet A-D per EVI route.

11. Security Considerations

   This document uses IP-based tunnel technologies to support data-plane
   transport. Consequently, the security considerations of those tunnel
   technologies apply. This document defines support for VXLAN
   [RFC7348] and NVGRE encapsulations [RFC7637]. The security
   considerations from those RFCs apply to the data-plane aspects of
   this document.

   As with [RFC5512], any modification of the information that is used
   to form encapsulation headers, to choose a tunnel type, or to choose
   a particular tunnel for a particular payload type may lead to user
   data packets getting misrouted, misdelivered, and/or dropped.

   More broadly, the security considerations for the transport of IP
   reachability information using BGP are discussed in [RFC4271] and
   [RFC4272] and are equally applicable for the extensions described in
   this document.

12. IANA Considerations

   This document registers the following in the "BGP Tunnel
   Encapsulation Attribute Tunnel Types" registry.

   Value   Name
   -----   -----------------------
   8   VXLAN Encapsulation
   9   NVGRE Encapsulation
   10   MPLS Encapsulation
   11   MPLS in GRE Encapsulation
   12   VXLAN GPE Encapsulation

13. References

13.1. Normative References

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

   [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
              2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
              May 2017, <https://www.rfc-editor.org/info/rfc8174>.

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

   [RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger,
              L., Sridhar, T., Bursell, M., and C. Wright, "Virtual
              eXtensible Local Area Network (VXLAN): A Framework for
              Overlaying Virtualized Layer 2 Networks over Layer 3
              Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014,
              <https://www.rfc-editor.org/info/rfc7348>.

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

   [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed.,
              "Encapsulating MPLS in IP or Generic Routing Encapsulation
              (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005,
              <https://www.rfc-editor.org/info/rfc4023>.

   [RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network
              Virtualization Using Generic Routing Encapsulation",
              RFC 7637, DOI 10.17487/RFC7637, September 2015,
              <https://www.rfc-editor.org/info/rfc7637>.

13.2. Informative References

   [RFC7209] Sajassi, A., Aggarwal, R., Uttaro, J., Bitar, N.,
              Henderickx, W., and A. Isaac, "Requirements for Ethernet
              VPN (EVPN)", RFC 7209, DOI 10.17487/RFC7209, May 2014,
              <https://www.rfc-editor.org/info/rfc7209>.

   [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis",
              RFC 4272, DOI 10.17487/RFC4272, January 2006,
              <https://www.rfc-editor.org/info/rfc4272>.

   [RFC7364] Narten, T., Ed., Gray, E., Ed., Black, D., Fang, L.,
              Kreeger, L., and M. Napierala, "Problem Statement:
              Overlays for Network Virtualization", RFC 7364,
              DOI 10.17487/RFC7364, October 2014,
              <https://www.rfc-editor.org/info/rfc7364>.

   [RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y.
              Rekhter, "Framework for Data Center (DC) Network
              Virtualization", RFC 7365, DOI 10.17487/RFC7365, October
              2014, <https://www.rfc-editor.org/info/rfc7365>.

   [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP
              Encodings and Procedures for Multicast in MPLS/BGP IP
              VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012,
              <https://www.rfc-editor.org/info/rfc6514>.

   [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,
              <https://www.rfc-editor.org/info/rfc4271>.

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

   [TUNNEL-ENCAP]
              Rosen, E., Ed., Patel, K., and G. Velde, "The BGP Tunnel
              Encapsulation Attribute", Work in Progress draft-ietf-idr-
              tunnel-encaps-09, February 2018.

   [DCI-EVPN-OVERLAY]
             Rabadan, J., Ed., Sathappan, S., Henderickx, W., Sajassi,
             A., and J. Drake, "Interconnect Solution for EVPN Overlay
             networks", Work in Progress, draft-ietf-bess-dci-evpn-
             overlay-10, March 2018.

   [EVPN-GENEVE]
             Boutros, S., Sajassi, A., Drake, J., and J. Rabadan, "EVPN
             control plane for Geneve", Work in Progress,
             draft-boutros-bess-evpn-geneve-02, March 2018.

   [VXLAN-GPE]
             Maino, F., Kreeger, L., Ed., and U. Elzur, Ed., "Generic
             Protocol Extension for VXLAN", Work in Progress,
             draft-ietf-nvo3-vxlan-gpe-05, October 2017.

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

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

Contributors

   S. Salam
   K. Patel
   D. Rao
   S. Thoria
   D. Cai
   Cisco

   Y. Rekhter
   A. Issac
   W. Lin
   N. Sheth
   Juniper

   L. Yong
   Huawei

Authors' Addresses

   Ali Sajassi (editor)
   Cisco
   United States of America

   Email: sajassi@cisco.com

   John Drake (editor)
   Juniper Networks
   United States of America

   Email: jdrake@juniper.net

   Nabil Bitar
   Nokia
   United States of America

   Email: nabil.bitar@nokia.com

   R. Shekhar
   Juniper
   United States of America

   Email: rshekhar@juniper.net

   James Uttaro
   AT&T
   United States of America

   Email: uttaro@att.com

   Wim Henderickx
   Nokia
   Copernicuslaan 50
   2018 Antwerp
   Belgium

   Email: wim.henderickx@nokia.com