Network Working Group   A. Sajassi, Ed.
INTERNET-DRAFT   Cisco
Category: Standards Track

                                                       R. Aggarwal
J. Drake   **Arktan**
Juniper Networks

                                                          N. Bitar
W. Henderickx   **Verizon**
Alcatel-Lucent

                                                      Aldrin Isaac
                                                         Bloomberg

                                                        J. Uttaro
                                                             AT&T

Expires: November 7, 2014   May 7, 2014

                    BGP MPLS Based Ethernet VPN
                      draft-ietf-l2vpn-evpn-07

Status of this Memo

   This Internet-Draft is submitted to IETF 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/1id-abstracts.html

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

Copyright and License Notice

Abstract

This document describes procedures for BGP MPLS based Ethernet VPNs
(EVPN).

Table of Contents

1. Specification of requirements

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

2. Terminology

Broadcast Domain: in a bridged network, it corresponds to a Virtual LAN (VLAN); where a VLAN is typically represented by a single VLAN ID (VID), but can be represented by several VIDs.

Bridge Domain: An instantiation of a broadcast domain on a bridge node

CE: Customer Edge device e.g., host or router or switch

EVI: An EVPN instance spanning across the PEs participating in that EVPN

MAC-VRF: A Virtual Routing and Forwarding table for MAC addresses on a PE for an EVI

Ethernet Segment Identifier (ESI): If a CE is multi-homed to two or more PEs, the set of Ethernet links that attaches the CE to the PEs is an 'Ethernet segment'.   Ethernet segments MUST have a unique non-zero identifier, the '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. Ethernet tag(s) are assigned to the broadcast domains of a given EVPN instance by the provider of that EVPN, and each PE in that EVPN instance performs a mapping between broadcast domain identifier(s) understood by each of its attached CEs and the corresponding Ethernet tag.

LACP: Link Aggregation Control Protocol

MP2MP: Multipoint to Multipoint

P2MP: Point to Multipoint

P2P: Point to Point

Single-Active Redundancy Mode: When only a single PE, among a group of PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet Segment, 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 traffic to/from that Ethernet Segment,
   then the Ethernet segment is defined to be operating in All-Active
   redundancy mode.

3. **Introduction**

   This document describes procedures for BGP MPLS based Ethernet VPNs
   (EVPN). The procedures described here are intended to meet the
   requirements specified in [EVPN-REQ]. Please refer to [EVPN-REQ] for
   the detailed requirements and motivation. EVPN requires extensions to
   existing IP/MPLS protocols as described in this document. In addition
   to these extensions EVPN uses several building blocks from existing
   MPLS technologies.

4. **BGP MPLS Based EVPN Overview**

   This section provides an overview of EVPN. An EVPN instance comprises
   CEs that are connected to PEs that form the edge of the MPLS
   infrastructure. A CE may be a host, a router or a switch. The PEs
   provide virtual Layer 2 bridged connectivity between the CEs. There
   may be multiple EVPN instances in the provider's network.

   The PEs may be connected by an MPLS LSP infrastructure which provides
   the benefits of MPLS technology such as fast-reroute, resiliency,
   etc. The PEs may also be connected by an IP infrastructure in which
   case IP/GRE tunneling or other IP tunneling can be used between the
   PEs. The detailed procedures in this version of this document are
   specified only for MPLS LSPs as the tunneling technology. However
   these procedures are designed to be extensible to IP tunneling as the
   Packet Switched Network (PSN) tunneling technology.

   In an EVPN, MAC learning between PEs occurs not in the data plane (as
   happens with traditional bridging in VPLS [RFC4761] or [RFC4762]) but
   in the control plane. Control plane learning offers greater control
   over the MAC learning process, such as restricting who learns what,
   and the ability to apply policies. Furthermore, the control plane
   chosen for advertising MAC reachability information is multi-protocol
   (MP) BGP (similar to IP VPNs (RFC 4364)). This provides flexibility
   and the ability to preserve the "virtualization" or isolation of
   groups of interacting agents (hosts, servers, virtual machines) from
   each other. In EVPN, PEs advertise the MAC addresses learned from the
   CEs that are connected to them, along with an MPLS label, to other
   PEs in the control plane using MP-BGP. Control plane learning enables

   load balancing of traffic to and from CEs that are multi-homed to
   multiple PEs. This is in addition to load balancing across the MPLS
   core via multiple LSPs between the same pair of PEs. In other words
   it allows CEs to connect to multiple active points of attachment. It
   also improves convergence times in the event of certain network
   failures.

   However, learning between PEs and CEs is done by the method best
   suited to the CE: data plane learning, IEEE 802.1x, LLDP, 802.1aq,
   ARP, management plane or other protocols.

   It is a local decision as to whether the Layer 2 forwarding table on
   a PE is populated with all the MAC destination addresses known to the
   control plane, or whether the PE implements a cache based scheme. For
   instance the MAC forwarding table may be populated only with the MAC
   destinations of the active flows transiting a specific PE.

   The policy attributes of EVPN are very similar to those of IP-VPN. A
   EVPN instance requires a Route-Distinguisher (RD) which is unique per
   PE and one or more globally unique Route-Targets (RTs). A CE attaches
   to a MAC-VRF on a PE, on an Ethernet interface which may be
   configured for one or more Ethernet Tags, e.g., VLAN IDs. Some
   deployment scenarios guarantee uniqueness of VLAN IDs across EVPN
   instances: all points of attachment for a given EVPN instance use the
   same VLAN ID, and no other EVPN instance uses this VLAN ID. This
   document refers to this case as a "Unique VLAN EVPN" and describes
   simplified procedures to optimize for it.

5. Ethernet Segment

   If a CE is multi-homed to two or more PEs, the set of Ethernet links
   constitutes an "Ethernet Segment". An Ethernet segment may appear to
   the CE as a Link Aggregation Group (LAG). Ethernet segments have an
   identifier, called the "Ethernet Segment Identifier" (ESI) which is
   encoded as a ten octets integer. The following two ESI values are
   reserved:

      - ESI 0 denotes a single-homed CE.

      - ESI {0xFF} (repeated 10 times) is known as MAX-ESI and is
   reserved.

   In general, an Ethernet segment MUST have a non-reserved ESI that is
   unique network wide (i.e., across all EVPN instances on all the PEs).
   If the CE(s) constituting an Ethernet Segment is (are) managed by the
   network operator, then ESI uniqueness should be guaranteed; however,
   if the CE(s) is (are) not managed, then the operator MUST configure a

network-wide unique ESI for that Ethernet Segment. This is required
to enable auto-discovery of Ethernet Segments and DF election.

In a network with managed and not-managed CEs, the ESI has the
following format:

```
+---+---+---+---+---+---+---+---+---+---+
| T |           ESI Value               |
+---+---+---+---+---+---+---+---+---+---+
```

Where:

T (ESI Type) is a 1-byte field (most significant octet) that
specifies the format of the remaining nine bytes (ESI Value). The
following 6 ESI types can be used:

- Type 0 (T=0x00) - This type indicates an arbitrary nine-octet ESI
value, which is managed and configured by the operator.

- Type 1 (T=0x01) - When IEEE 802.1AX LACP is used between the PEs
and CEs, this ESI type indicates an auto-generated ESI value
determined from LACP by concatenating the following parameters:

> + CE LACP six octets System MAC address. The CE LACP System MAC
>   address MUST be encoded in the high order six octets of the ESI
>   Value field.

> + CE LACP two octets Port Key. The CE LACP port key MUST be
>   encoded in the two octets next to the System MAC address.

> + The remaining octet will be set to 0x00.

> As far as the CE is concerned, it would treat the multiple PEs
> that it is connected to as the same switch. This allows the CE
> to aggregate links that are attached to different PEs in the
> same bundle.

> This mechanism could be used only if it produces ESIs that satisfy
> the uniqueness requirement specified above.

- Type 2 (T=0x02) - This type is used in the case of indirectly
connected hosts via a bridged LAN between the CEs and the PEs. The
ESI Value is auto-generated and determined based on the Layer 2
bridge protocol as follows: If MST is used in the bridged LAN then
the value of the ESI is derived by listening to BPDUs on the Ethernet
segment. To achieve this the PE is not required to run MST. However
the PE must learn the Root Bridge MAC address and Bridge Priority of
the root of the Internal Spanning Tree (IST) by listening to the

BPDUs. The ESI Value is constructed as follows:

> + Root Bridge six octets MAC address. The Root Bridge MAC
>   address MUST be encoded in the high order six octets of the
>   ESI Value field.
>
> + Root Bridge two octets Priority. The CE Root Bridge Priority
>   MUST be encoded in the two octets next to the Root Bridge
>   MAC address.
>
> + The remaining octet will be set to 0x00.
>
> This mechanism could be used only if it produces ESIs that
> satisfy the uniqueness requirement specified above.

- Type 3 (T=0x03) - This type indicates a MAC-based ESI Value that
can be auto-generated or configured by the operator. The ESI Value is
constructed as follows:

> + System MAC address (six octets). The PE MAC address MUST
>   be encoded in the high order six octets of the ESI Value field.
>
> + Local Discriminator value (three octets). The Local
>   Discriminator MUST be encoded in the low order three octets
>   of the ESI Value.
>
> This mechanism could be used only if it produces ESIs that
> satisfy the uniqueness requirement specified above.

- Type 4 (T=0x04) - This type indicates a router-ID ESI Value that
can be auto-generated or configured by the operator. The ESI Value is
constructed as follows:

> + Router ID (four octets). The system router ID MUST be encoded in
>   the high order four octets of the ESI Value field.
>
> + Local Discriminator value (four octets). The Local
>   Discriminator MUST be encoded in the four octets next to the
>   IP address.
>
> + The low order octet of the ESI Value will be set to 0x00.
>
> This mechanism could be used only if it produces ESIs that
> satisfy the uniqueness requirement specified above.

- Type 5 (T=0x05) - This type indicates an AS-based ESI Value that
can be auto-generated or configured by the operator. The ESI Value is
constructed as follows:

    + AS number (four octets). This is an AS number owned by the
     system and MUST be encoded in the high order four octets of the
     ESI Value field. If a two-octet AS number is used, the high order
     extra two bytes will be 0x0000.

    + Local Discriminator value (four octets). The Local Discriminator
      MUST be encoded in the four octets next to the AS number.

    + The low order octet of the ESI Value will be set to 0x00.

    This mechanism could be used only if it produces ESIs that satisfy
    the uniqueness requirement specified above.

6. Ethernet Tag ID

   An Ethernet Tag ID is a 32-bit field containing either a 12-bit or a
   24-bit identifier that identifies a particular broadcast domain
   (e.g., a VLAN) in an EVPN Instance. The 12-bit identifier is called
   VLAN ID (VID). An EVPN Instance consists of one or more broadcast
   domains (one or more VLANs). VLANs are assigned to a given EVPN
   Instance by the provider of the EVPN service. A given VLAN can itself
   be represented by multiple VLAN IDs (VIDs). In such cases, the PEs
   participating in that VLAN for a given EVPN instance are responsible
   for performing VLAN ID translation to/from locally attached CE
   devices.

   If a VLAN is represented by a single VID across all PE devices
   participating in that VLAN for that EVPN instance, then there is no
   need for VID translation at the PEs. Furthermore, some deployment
   scenarios guarantee uniqueness of VIDs across all EVPN instances;
   all points of attachment for a given EVPN instance use the same VID
   and no other EVPN instances use that VID. This allows the RT(s) for
   each EVPN instance to be derived automatically from the corresponding
   VID, as described in section 7.10.1.

   The following subsections discuss the relationship between broadcast
   domains (e.g., VLANs), Ethernet Tags (e.g., VIDs), and MAC-VRFs as
   well as the setting of the Ethernet Tag Identifier, in the various
   EVPN BGP routes (defined in section 8), for the different types of
   service interfaces described in [EVPN-REQ].

   The following value of Ethernet Tag is reserved:

      - Ethernet Tag {0xFFFFFFFF} is known as MAX-ET

6.1 VLAN Based Service Interface

With this service interface, an EVPN instance consists of only a
single broadcast domain (e.g., a single VLAN). Therefore, there is a
one to one mapping between a VID on this interface and a MAC-VRF.
Since a MAC-VRF corresponds to a single VLAN, it consists of a single
bridge domain corresponding to that VLAN. If the VLAN is represented
by different VIDs on different PEs, then each PE needs to perform VID
translation for frames destined to its attached CEs. In such
scenarios, the Ethernet frames transported over MPLS/IP network
SHOULD remain tagged with the originating VID and a VID translation
MUST be supported in the data path and MUST be performed on the
disposition PE. The Ethernet Tag Identifier in all EVPN routes MUST
be set to 0.

6.2 VLAN Bundle Service Interface

With this service interface, an EVPN instance corresponds to several
broadcast domains (e.g., several VLANs); however, only a single
bridge domain is maintained per MAC-VRF which means multiple VLANs
share the same bridge domain. This implies MAC addresses MUST be
unique across different VLANs for this service to work. In other
words, there is a many-to-one mapping between VLANs and a MAC-VRF,
and the MAC-VRF consists of a single bridge domain. Furthermore, a
single VLAN must be represented by a single VID - e.g., no VID
translation is allowed for this service interface type. The MPLS
encapsulated frames MUST remain tagged with the originating VID. Tag
translation is NOT permitted. The Ethernet Tag Identifier in all EVPN
routes MUST be set to 0.

6.2.1 Port Based Service Interface

This service interface is a special case of the VLAN Bundle service
interface, where all of the VLANs on the port are part of the same
service and map to the same bundle. The procedures are identical to
those described in section 6.2.

6.3 VLAN Aware Bundle Service Interface

With this service interface, an EVPN instance consists of several
broadcast domains (e.g., several VLANs) with each VLAN having its own
bridge domain - i.e., multiple bridge domains (one per VLAN) is
maintained by a single MAC-VRF corresponding to the EVPN instance. In
the case where a single VLAN is represented by different VIDs on
different CEs and thus tag (VID) translation is required, a
normalized Ethernet Tag (VID) MUST be carried in the MPLS
encapsulated frames and a tag translation function MUST be supported
in the data path. This translation MUST be performed in data path on
both the imposition as well as the disposition PEs (translating to
normalized tag on imposition PE and translating to local tag on

disposition PE). The Ethernet Tag Identifier in all EVPN routes MUST
be set to the normalized Ethernet Tag assigned by the EVPN provider.

6.3.1 Port Based VLAN Aware Service Interface

This service interface is a special case of the VLAN Aware Bundle
service interface, where all of the VLANs on the port are part of the
same service and are mapped to a single bundle but without any VID
translation. The procedures are subset of those described in section
6.3.

7. BGP EVPN NLRI

This document defines a new BGP NLRI, called the EVPN NLRI.

Following is the format of the EVPN NLRI:

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

The Route Type field defines encoding of the rest of the EVPN NLRI
(Route Type specific EVPN NLRI).

The Length field indicates the length in octets of the Route Type
specific field of EVPN NLRI.

This document defines the following Route Types:

        + 1 - Ethernet Auto-Discovery (A-D) route
        + 2 - MAC/IP advertisement route
        + 3 - Inclusive Multicast Ethernet Tag Route
        + 4 - Ethernet Segment Route

The detailed encoding and procedures for these route types are
described in subsequent sections.

The EVPN NLRI is carried in BGP [RFC4271] using BGP Multiprotocol
Extensions [RFC4760] with an AFI of 25 (L2VPN) and a SAFI of 70
(EVPN). The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute
contains the EVPN NLRI (encoded as specified above).

In order for two BGP speakers to exchange labeled EVPN NLRI, they
must use BGP Capabilities Advertisement to ensure that they both are

capable of properly processing such NLRI. This is done as specified
in [RFC4760], by using capability code 1 (multiprotocol BGP) with an
AFI of 25 (L2VPN) and a SAFI of 70 (EVPN).

7.1. Ethernet Auto-Discovery Route

A Ethernet A-D route type specific EVPN NLRI consists of the
following:

```
               +--------------------------------------+
               |     RD   (8 octets)                  |
               +--------------------------------------+
               |Ethernet Segment Identifier (10 octets)|
               +--------------------------------------+
               |  Ethernet Tag ID (4 octets)          |
               +--------------------------------------+
               |  MPLS Label (3 octets)               |
               +--------------------------------------+
```

For the purpose of BGP route key processing, only the Ethernet
Segment Identifier and the Ethernet Tag ID are considered to be part
of the prefix in the NLRI.   The MPLS Label field is to be treated as
a route attribute as opposed to being part of the route.

For procedures and usage of this route please see section 8.2 "Fast
Convergence" and section 8.4 "Aliasing".

7.2. MAC/IP Advertisement Route

A MAC/IP advertisement route type specific EVPN NLRI consists of the
following:

```
+---------------------------------------+
|      RD   (8 octets)                  |
+---------------------------------------+
|Ethernet Segment Identifier (10 octets)|
+---------------------------------------+
|  Ethernet Tag ID (4 octets)           |
+---------------------------------------+
|  MAC Address Length (1 octet)         |
+---------------------------------------+
|  MAC Address (6 octets)               |
+---------------------------------------+
|  IP Address Length (1 octet)          |
+---------------------------------------+
|  IP Address (0 or 4 or 16 octets)     |
+---------------------------------------+
|  MPLS Label1 (3 octets)               |
+---------------------------------------+
|  MPLS Label2 (0 or 3 octets)          |
+---------------------------------------+
```

For the purpose of BGP route key processing, only the Ethernet Tag
ID, MAC Address Length, MAC Address, IP Address Length, and IP
Address Address fields are considered to be part of the prefix in the
NLRI. The Ethernet Segment Identifier and MPLS Label fields are to be
treated as route attributes as opposed to being part of the "route".
The IP address length is in bits.

For procedures and usage of this route please see section 9
"Determining Reachability to Unicast MAC Addresses" and section 14
"Load Balancing of Unicast Packets".

7.3. Inclusive Multicast Ethernet Tag Route

An Inclusive Multicast Ethernet Tag route type specific EVPN NLRI
consists of the following:

```
+---------------------------------------+
|      RD   (8 octets)                  |
+---------------------------------------+
|  Ethernet Tag ID (4 octets)           |
+---------------------------------------+
|  IP Address Length (1 octet)          |
+---------------------------------------+
|   Originating Router's IP Addr        |
|          (4 or 16 octets)             |
+---------------------------------------+
```

For procedures and usage of this route please see section 11
"Handling of Multi-Destination Traffic", section 13 "Processing of
Unknown Unicast Traffic" and section 16 "Multicast". The IP address
length is in bits. For the purpose of BGP route key processing, only
the Ethernet Tag ID, IP Address Length, and Originating Router's IP
Address fields are considered to be part of the prefix in the NLRI.

7.4 Ethernet Segment Route

An Ethernet Segment route type specific EVPN NLRI consists of the
following:

```
                +---------------------------------------+
                |      RD   (8 octets)                  |
                +---------------------------------------+
                |Ethernet Segment Identifier (10 octets)|
                +---------------------------------------+
                |  IP Address Length (1 octet)          |
                +---------------------------------------+
                |   Originating Router's IP Addr        |
                |          (4 or 16 octets)             |
                +---------------------------------------+
```

For procedures and usage of this route please see section 8.5
"Designated Forwarder Election". The IP address length is in bits.
For the purpose of BGP route key processing, only the Ethernet
Segment ID, IP Address Length, and Originating Router's IP Address
fields are considered to be part of the prefix in the NLRI.

7.5 ESI Label Extended Community

This extended community is a new transitive extended community with
the Type field is 0x06, and the Sub-Type of 0x01. It may be
advertised along with Ethernet Auto-Discovery routes and it enables
split-horizon procedures for multi-homed sites as described in
section 8.3 "Split Horizon".

Each ESI Label Extended Community is encoded as a 8-octet value as
follows:

```
     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
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | Type=0x06    | Sub-Type=0x01 | Flags(1 Octet)|  Reserved=0   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | Reserved = 0  |         ESI Label                            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The low order bit of the flags octet is defined as the "Single-

   Active" bit. A value of 0 means that the multi-homed site is
   operating in All-Active redundancy mode and a value of 1 means that
   the multi-homed site is operating in Single-Active redundancy mode.

7.6 ES-Import Route Target

   This is a new transitive Route Target extended community carried with
   the Ethernet Segment route. When used, it enables all the PEs
   connected to the same multi-homed site to import the Ethernet Segment
   routes. The value is derived automatically from the ESI by encoding
   the high order 6-byte portion of the 9-byte ESI Value in the ES-
   Import Route Target. The high order 6-byte of the ESI incorporates
   MAC address of ESI (for type 1, 2, and 3) which when encoded in this
   RT and used in the RT constrain feature, it enables proper route-
   target filtering. The format of this extended community is as
   follows:

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Type=0x06     | Sub-Type=0x02 |          ES-Import            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                      ES-Import Cont'd                         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   This document expands the definition of the Route Target extended
   community to allow the value of high order octet (Type field) to be
   0x06 (in addition to the values specified in rfc4360). The value of
   low order octet (Sub-Type field) of 0x02 indicates that this extended
   community is of type "Route Target". The new value for Type field of
   0x06 indicates that the structure of this RT is a six bytes value
   (e.g., a MAC address). A BGP speaker that implements RT-Constrain
   [RFC4684] MUST apply the RT Constraint procedures to the ES-import RT
   as well.

   For procedures and usage of this attribute, please see section 8.1
   "Multi-homed Ethernet Segment Auto-Discovery".

7.7 MAC Mobility Extended Community

   This extended community is a new transitive extended community with
   the Type field of 0x06 and the Sub-Type of 0x00. It may be advertised
   along with MAC Advertisement routes. The procedures for using this
   Extended Community are described in section 15 "MAC Mobility".

   The MAC Mobility Extended Community is encoded as an 8-octet value as
   follows:

```
  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
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Type=0x06    | Sub-Type=0x00 |Flags(1 octet)|  Reserved=0    |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                      Sequence Number                         |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The low order bit of the flags octet is defined as the
"Sticky/static" flag and may be set to 1. A value of 1 means that the
MAC address is static and cannot move. The sequence number is used to
ensure that PEs retain the correct MAC advertisement route when
multiple updates occur for the same MAC address.

7.8 Default Gateway Extended Community

The Default Gateway community is an Extended Community of an Opaque
Type (see 3.3 of rfc4360). It is a transitive community, which means
that the first octet is 0x03. The value of the second octet (Sub-
Type) is 0x0d (Default Gateway) as assigned by IANA. The Value field
of this community is reserved (set to 0 by the senders, ignored by
the receivers).

7.9 Route Distinguisher Assignment per EVI

Route Distinguisher (RD) MUST be set to the RD of the EVI that is
advertising the NLRI. An RD MUST be assigned for a given EVI on a PE.
This RD MUST be unique across all EVIs on a PE. It is RECOMMENDED to
use the Type 1 RD [RFC4364]. The value field comprises an IP address
of the PE (typically, the loopback address) followed by a number
unique to the PE. This number may be generated by the PE. Or in the
Unique VLAN EVPN case, the low order 12 bits may be the 12 bit VLAN
ID, with the remaining high order 4 bits set to 0.

7.10 Route Targets

The EVPN route MAY carry one or more Route Target (RT) attributes.
RTs may be configured (as in IP VPNs), or may be derived
automatically.

If a PE uses RT-Constrain, the PE SHOULD advertise all such RTs using
RT Constraints. The use of RT Constrains allows each Ethernet A-D
route to reach only those PEs that are configured to import at least
one RT from the set of RTs carried in the EVPN route.

7.10.1 Auto-Derivation from the Ethernet Tag ID

For the "Unique VLAN EVPN" scenario, it is highly desirable to auto-

   derive the RT from the Ethernet Tag ID (VLAN ID) for that EVPN
   instance. The following is the procedure for performing such auto-
   derivation.

        +   The Global Administrator field of the RT MUST be set
             to the Autonomous System (AS) number that the PE is
          associated with.

        +   The 12-bit VLAN ID MUST be encoded in the lowest 12 bits of
             the Local Administrator field.

8. Multi-homing Functions

   This section discusses the functions, procedures and associated BGP
   routes used to support multi-homing in EVPN. This covers both multi-
   homed device (MHD) as well as multi-homed network (MHN) scenarios.

8.1 Multi-homed Ethernet Segment Auto-Discovery

   PEs connected to the same Ethernet segment can automatically discover
   each other with minimal to no configuration through the exchange of
   the Ethernet Segment route.

8.1.1 Constructing the Ethernet Segment Route

   The Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value
   field comprises an IP address of the PE (typically, the loopback
   address) followed by 0's.

   The Ethernet Segment Identifier MUST be set to the ten octet ESI
   identifier described in section 5.

   The BGP advertisement that advertises the Ethernet Segment route MUST
   also carry an ES-Import route target, as defined in section 7.6.

   The Ethernet Segment Route filtering MUST be done such that the
   Ethernet Segment Route is imported only by the PEs that are multi-
   homed to the same Ethernet Segment. To that end, each PE that is
   connected to a particular Ethernet segment constructs an import
   filtering rule to import a route that carries the ES-Import extended
   community, constructed from the ESI.

8.2 Fast Convergence

   In EVPN, MAC address reachability is learnt via the BGP control-plane
   over the MPLS network. As such, in the absence of any fast protection
   mechanism, the network convergence time is a function of the number

of MAC Advertisement routes that must be withdrawn by the PE
encountering a failure. For highly scaled environments, this scheme
yields slow convergence.

To alleviate this, EVPN defines a mechanism to efficiently and
quickly signal, to remote PE nodes, the need to update their
forwarding tables upon the occurrence of a failure in connectivity to
an Ethernet segment. This is done by having each PE advertise a set
of Ethernet A-D per Ethernet segment (per ES) routes for each locally
attached Ethernet segment (refer to section 8.2.1 below for details
on how this route is constructed). Upon a failure in connectivity to
the attached segment, the PE withdraws the corresponding Ethernet A-D
route. This triggers all PEs that receive the withdrawal to update
their next-hop adjacencies for all MAC addresses associated with the
Ethernet segment in question. If no other PE had advertised an
Ethernet A-D route for the same segment, then the PE that received
the withdrawal simply invalidates the MAC entries for that segment.
Otherwise, the PE updates the next-hop adjacencies to point to the
backup PE(s).

8.2.1 Constructing the Ethernet A-D per Ethernet Segment (ES) Route

This section describes the procedures used to construct the Ethernet
A-D per ES route, which is used for fast convergence (as discussed
above) and for advertising the ESI label used for split-horizon
filtering (as discussed in section 8.3). Support of this route is
MANDATORY.

The Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value
field comprises an IP address of the PE (typically, the loopback
address) followed by a number unique to the PE.

The Ethernet Segment Identifier MUST be a ten octet entity as
described in section "Ethernet Segment". The Ethernet A-D route is
not needed when the Segment Identifier is set to 0 (e.g., single-
homed scenarios).

The Ethernet Tag ID MUST be set to MAX-ET.

The MPLS label in the NLRI MUST be set to 0.

The "ESI Label Extended Community" MUST be included in the route. If
All-Active redundancy mode is desired, then the "Single-Active" bit
in the flags of the ESI Label Extended Community MUST be set to 0 and
the MPLS label in that extended community MUST be set to a valid MPLS
label value. The MPLS label in this Extended Community is referred to
as the ESI label and MUST have the same value in each Ethernet A-D
per ES route advertised for the ES. This label MUST be a downstream

assigned MPLS label if the advertising PE is using ingress
replication for receiving multicast, broadcast or unknown unicast
traffic from other PEs. If the advertising PE is using P2MP MPLS LSPs
for sending multicast, broadcast or unknown unicast traffic, then
this label MUST be an upstream assigned MPLS label. The usage of this
label is described in section 8.3.

If Single-Active redundancy mode is desired, then the "Single-Active"
bit in the flags of the ESI Label Extended Community MUST be set to 1
and the ESI label MUST be set to zero.

8.2.1.1. Ethernet A-D Route Targets

Each Ethernet A-D per ES route MUST carry one or more Route Target
(RT) attributes. The set of Ethernet A-D routes per ES MUST carry the
entire set of RTs for all the EVPN instances to which the Ethernet
Segment belongs.

8.3 Split Horizon

Consider a CE that is multi-homed to two or more PEs on an Ethernet
segment ES1 operating in All-Active redundancy mode. If the CE sends
a broadcast, unknown unicast, or multicast (BUM) packet to one of the
non-DF (Designated Forwarder) PEs, say PE1, then PE1 will forward
that packet to all or subset of the other PEs in that EVPN instance
including the DF PE for that Ethernet segment. In this case the DF PE
that the CE is multi-homed to MUST drop the packet and not forward
back to the CE. This filtering is referred to as "split horizon"
filtering in this document.

When a set of PEs operating in Single-Active redundancy mode, the use
of this split-horizon filtering mechanism is highly recommended
because it prevents transient loop at the time of failure or recovery
impacting the Ethernet Segment - e.g., when two PEs thinks that both
are DFs for that segment before DF election procedure settles down.

In order to achieve this split horizon function, every BUM packet
originating from a non-DF PE is encapsulated with an MPLS label that
identifies the Ethernet segment of origin (i.e. the segment from
which the frame entered the EVPN network). This label is referred to
as the ESI label, and MUST be distributed by all PEs when operating
in All-Active redundancy mode using a set of Ethernet A-D per ES
routes per section 8.2.1 above. The ESI label SHOULD be distributed
by all PEs when operating in Single-Active redundancy mode using a
set of Ethernet A-D per ES route. This route is imported by the PEs
connected to the Ethernet Segment and also by the PEs that have at
least one EVPN instance in common with the Ethernet Segment in the
route. As described in section 8.1.1, the route MUST carry an ESI

Label Extended Community with a valid ESI label. The disposition PE
rely on the value of the ESI label to determine whether or not a BUM
frame is allowed to egress a specific Ethernet segment.

8.3.1 ESI Label Assignment

The following subsections describe the assignment procedures for the
ESI label, which differ depending on the type of tunnels being used
to deliver multi-destination packets in the EVPN network.

8.3.1.1 Ingress Replication

Each PE attached to a given ES that is operating in All-Active or
Single-Active redundancy mode and that uses ingress replication to
receive BUM traffic advertises a downstream assigned ESI label in the
set of Ethernet A-D per ES routes for that ES. This label MUST be
programmed in the platform label space by the advertising PE and the
forwarding entry for this label must result in NOT forwarding packets
received with this label onto the Ethernet segment for which the
label was distributed.

The rules for the inclusion of the ESI label in a BUM packet by the
ingress PE operating in All-Active redundancy mode are as follows:

A non-DF ingress PE MUST include the ESI label distributed by the DF
egress PE in the copy of a BUM packet sent to it.

An ingress PE (DF or non-DF) SHOULD include the ESI label distributed
by each non-DF egress PE in the copy of a BUM packet sent to it.

The rules for the inclusion of the ESI label in a BUM packet by the
ingress PE operating in Single-Active redundancy mode are as follows:

An ingress DF PE SHOULD include the ESI label distributed by the
egress PE in the copy of a BUM packet sent to it.

In both All-Active and Single-Active redundancy mode, an ingress PE
MUST NOT include an ESI label in the copy of a BUM packet sent to an
egress PE that is not attached to the ES through which the BUM packet
entered the EVI.

As an example, consider PE1 and PE2 that are multi-homed to CE1 on
ES1 and operating in All-Active multi-homing mode. Further consider
that PE1 is using P2P or MP2P LSPs to send packets to PE2. Consider
that PE1 is the non-DF for VLAN1 and PE2 is the DF for VLAN1, and PE1
receives a BUM packet from CE1 on VLAN1 on ES1. In this scenario, PE2
distributes an Inclusive Multicast Ethernet Tag route for VLAN1
corresponding to an EVPN instance. So, when PE1 sends a BUM packet,

that it receives from CE1, it MUST first push onto the MPLS label
stack the ESI label that PE2 has distributed for ES1. It MUST then
push on the MPLS label distributed by PE2 in the Inclusive Multicast
Ethernet Tag route for VLAN1. The resulting packet is further
encapsulated in the P2P or MP2P LSP label stack required to transmit
the packet to PE2. When PE2 receives this packet, it determines the
set of ESIs to replicate the packet to from the top MPLS label, after
any P2P or MP2P LSP labels have been removed. If the next label is
the ESI label assigned by PE2 for ES1, then PE2 MUST NOT forward the
packet onto ES1. If the next label is an ESI label which has not been
assigned by PE2, then PE2 MUST drop the packet. It should be noted
that in this scenario, if PE2 receives a BUM packet for VLAN1 from
CE1, then it SHOULD encapsulate the packet with an ESI label received
from PE1 when sending it to PE1 in order to avoid any transient loop
during a failure scenario impacting ES1 (e.g., port or link failure).

8.3.1.2. P2MP MPLS LSPs

The non-DF PEs attached to a given ES that is operating in All-Active
redundancy mode and that use P2MP LSPs to send BUM traffic advertise
an upstream assigned ESI label in the set of Ethernet A-D per ES
routes for that ES. This label is upstream assigned by the PE that
advertises the route. This label MUST be programmed by the other PEs,
that are connected to the ESI advertised in the route, in the context
label space for the advertising PE. Further the forwarding entry for
this label must result in NOT forwarding packets received with this
label onto the Ethernet segment that the label was distributed for.
This label MUST also be programmed by the other PEs, that import the
route but are not connected to the ESI advertised in the route, in
the context label space for the advertising PE. Further the
forwarding entry for this label must be a POP with no other
associated action.

The DF PE attached to a given ES that is operating in Single-Active
redundancy mode and that use P2MP LSPs to send BUM traffic should
advertise an upstream assigned ESI label in the set of Ethernet A-D
per ES routes for that ES just as above paragraph.

As an example, consider PE1 and PE2 that are multi-homed to CE1 on
ES1 and operating in All-Active multi-homing mode. Also consider PE3
belongs to one of the EVPN instances of ES1. Further, assume that
PE1 which is the non-DF, using P2MP MPLS LSPs to send BUM packets.
When PE1 sends a BUM packet, that it receives from CE1, it MUST first
push onto the MPLS label stack the ESI label that it has assigned for
the ESI that the packet was received on. The resulting packet is
further encapsulated in the P2MP MPLS label stack necessary to
transmit the packet to the other PEs. Penultimate hop popping MUST be

disabled on the P2MP LSPs used in the MPLS transport infrastructure
for EVPN. When PE2 receives this packet, it de-capsulates the top
MPLS label and forwards the packet using the context label space
determined by the top label. If the next label is the ESI label
assigned by PE1 to ES1, then PE2 MUST NOT forward the packet onto
ES1. When PE3 receives this packet, it de-capsulates the top MPLS
label and forwards the packet using the context label space
determined by the top label. If the next label is the ESI label
assigned by PE1 to ES1 and PE3 is not connected to ES1, then PE3 MUST
pop the label and flood the packet over all local ESIs in that EVPN
instance. It should be noted that when PE2 sends a BUM frame over a
P2MP LSP, it should encapsulate the frame with an ESI label even
though it is the DF for that VLAN in order to avoid any transient
loop during a failure scenario impacting ES1 (e.g., port or link
failure).

8.4 Aliasing and Backup-Path

In the case where a CE is multi-homed to multiple PE nodes, using a
LAG with All-Active redundancy, it is possible that only a single PE
learns a set of the MAC addresses associated with traffic transmitted
by the CE. This leads to a situation where remote PE nodes receive
MAC advertisement routes, for these addresses, from a single PE even
though multiple PEs are connected to the multi-homed segment. As a
result, the remote PEs are not able to effectively load-balance
traffic among the PE nodes connected to the multi-homed Ethernet
segment. This could be the case, for e.g. when the PEs perform data-
plane learning on the access, and the load-balancing function on the
CE hashes traffic from a given source MAC address to a single PE.
Another scenario where this occurs is when the PEs 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 address this issue, EVPN introduces the concept of 'Aliasing'
which is the ability of a PE to signal that it has reachability to an
EVPN instance on a given ES even when it has learnt no MAC addresses
from that EVI/ES. The Ethernet A-D per EVI route is used for this
purpose. A remote PE that receives a MAC advertisement route with
non-reserved ESI SHOULD consider the advertised MAC address to be
reachable via all PEs that have advertised reachability to that MAC
address' EVI/ES via the combination of an Ethernet A-D per EVI route
for that EVI/ES (and Ethernet Tag if applicable) AND Ethernet A-D per
ES routes for that ES with the 'Single-Active' bit in the flags of
the ESI Label Extended Community set to 0.

Note that the Ethernet A-D per EVI route may be received by a remote
PE before it receives the set of Ethernet A-D per ES routes.

Therefore, in order to handle corner cases and race conditions, the Ethernet A-D per EVI route MUST NOT be used for traffic forwarding by a remote PE until it also receives the associated set of Ethernet A-D per ES routes.

Backup-path is a closely related function, but it is used in Single-Active redundancy mode. In this case a PE also advertises that it has reachability to a give EVI/ES using same combination of Ethernet A-D per EVI route and Ethernet A-D per ES route as above, but with the 'Single-Active' bit in the flags of the ESI Label Extended Community set to 1.   A remote PE that receives a MAC advertisement route with non-reserved ESI SHOULD consider the advertised MAC address to be reachable via any PE that has advertised this combination of Ethernet A-D routes and it SHOULD install a backup-path for that MAC address.

8.4.1 Constructing the Ethernet A-D per EVPN Instance (EVI) Route

This section describes the procedures used to construct the Ethernet A-D per EVPN Instance (EVI) route, which is used for aliasing (as discussed above). Support of this route is OPTIONAL.

Route-Distinguisher (RD) MUST be set to the RD of the EVI that is advertising the NLRI per section 7.9.

The Ethernet Segment Identifier MUST be a ten octet entity as described in section "Ethernet Segment Identifier". The Ethernet A-D route is not needed when the Segment Identifier is set to 0.

The Ethernet Tag ID is the identifier of an Ethernet Tag on the Ethernet segment. This value may be a 12 bit VLAN ID, in which case the low order 12 bits are set to the VLAN ID and the high order 20 bits are set to 0. Or it may be another Ethernet Tag used by the EVPN. It MAY be set to the default Ethernet Tag on the Ethernet segment or to the value 0.

Note that the above allows the Ethernet A-D route to be advertised with one of the following granularities:

   + One Ethernet A-D route for a given <ESI, Ethernet Tag ID> tuple
     per EVI. This is applicable when the PE uses MPLS-based
     disposition.

   + One Ethernet A-D route per <ESI, EVI> (where the Ethernet
     Tag ID is set to 0). This is applicable when the PE uses
     MAC-based disposition, or when the PE uses MPLS-based
     disposition when no VLAN translation is required.

The usage of the MPLS label is described in the section on "Load
Balancing of Unicast Packets".

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST
be set to the IPv4 or IPv6 address of the advertising PE.

The Ethernet A-D route MUST carry one or more Route Target (RT)
attributes per section 7.10.

8.5 Designated Forwarder Election

Consider a CE that is a host or a router that is multi-homed directly
to more than one PE in an EVPN instance on a given Ethernet segment.
One or more Ethernet Tags may be configured on the Ethernet segment.
In this scenario only one of the PEs, referred to as the Designated
Forwarder (DF), is responsible for certain actions:

- Sending multicast and broadcast traffic, on a given Ethernet
  Tag on a particular Ethernet segment, to the CE.

- Flooding unknown unicast traffic (i.e. traffic for
  which a PE does not know the destination MAC address),
  on a given Ethernet Tag on a particular Ethernet segment
  to the CE, if the environment requires flooding of
  unknown unicast traffic.

Note that this behavior, which allows selecting a DF at the
granularity of <ESI, EVI> for multicast, broadcast and unknown
unicast traffic, is the default behavior in this specification.

Note that a CE always sends packets belonging to a specific flow
using a single link towards a PE. For instance, if the CE is a host
then, as mentioned earlier, the host treats the multiple links that
it uses to reach the PEs as a Link Aggregation Group (LAG). The CE
employs a local hashing function to map traffic flows onto links in
the LAG.

If a bridged network is multi-homed to more than one PE in an EVPN
network via switches, then the support of All-Active redundancy mode
requires the bridged network to be connected to two or more PEs using
a LAG.

If a bridged network does not connect to the PEs using LAG, then only
one of the links between the switched bridged network and the PEs
must be the active link for a given EVPN instance. In this case, the
set of Ethernet A-D per ES routes advertised by each PE MUST have the

'Single-Active' bit in the flags of the ESI Label Extended Community
set to 1.

The default procedure for DF election at the granularity of <ESI,
EVI> is referred to as "service carving". With service carving, it is
possible to elect multiple DFs per Ethernet Segment (one per EVI) in
order to perform load-balancing of multi-destination traffic destined
to a given Segment. The load-balancing procedures carve up the EVI
space among the PE nodes evenly, in such a way that every PE is the
DF for a disjoint set of EVIs. The procedure for service carving is
as follows:

1. When a PE discovers the ESI of the attached Ethernet Segment, it
advertises an Ethernet Segment route with the associated ES-Import
extended community attribute.

2. The PE then starts a timer (default value = 3 seconds) to allow
the reception of Ethernet Segment routes from other PE nodes
connected to the same Ethernet Segment. This timer value MUST be same
across all PEs connected to the same Ethernet Segment.

3. When the timer expires, each PE builds an ordered list of the IP
addresses of all the PE nodes connected to the Ethernet Segment
(including itself), in increasing numeric value. Each IP address in
this list is extracted from the "Originator Router's IP address"
field of the advertised Ethernet Segment route. Every PE is then
given an ordinal indicating its position in the ordered list,
starting with 0 as the ordinal for the PE with the numerically lowest
IP address. The ordinals are used to determine which PE node will be
the DF for a given EVPN instance on the Ethernet Segment using the
following rule: Assuming a redundancy group of N PE nodes, the PE
with ordinal i is the DF for an EVPN instance with an associated
Ethernet Tag value V when (V mod N) = i. In the case where multiple
Ethernet Tags are associated with a single EVPN instance, then the
numerically lowest Ethernet Tag value in that EVPN instance MUST be
used in the modulo function.

It should be noted that using "Originator Router's IP address" field
in the Ethernet Segment route to get the PE IP address needed for the
ordered list, allows for a CE to be multi-homed across different ASes
if such need ever arises.

4. The PE that is elected as a DF for a given EVPN instance will
unblock traffic for the Ethernet Tags associated with that EVPN
instance. Note that the DF PE unblocks multi-destination traffic in
the egress direction towards the Segment. All non-DF PEs continue to
drop multi-destination traffic (for the associated EVPN instances) in
the egress direction towards the Segment.

In the case of link or port failure, the affected PE withdraws its Ethernet Segment route. This will re-trigger the service carving procedures on all the PEs in the RG. For PE node failure, or upon PE commissioning or decommissioning, the PEs re-trigger the service carving. In case of a Single-Active multi-homing, when a service moves from one PE in the RG to another PE as a result of re-carving, the PE, which ends up being the elected DF for the service, SHOULD trigger a MAC address flush notification towards the associated Ethernet Segment. This can be done, for e.g. using IEEE 802.1ak MVRP 'new' declaration.

8.6. Interoperability with Single-homing PEs

Let's refer to PEs that only support single-homed CE devices as single-homing PEs. For single-homing PEs, all the above multi-homing procedures can be omitted; however, to allow for single-homing PEs to fully inter-operate with multi-homing PEs, some of the multi-homing procedures described above SHOULD be supported even by single-homing PEs:

- procedures related to processing Ethernet A-D route for the purpose of Fast Convergence (8.2 Fast Convergence), to let single-homing PEs benefit from fast convergence

- procedures related to processing Ethernet A-D route for the purpose of Aliasing (8.4 Aliasing and Backup-path), to let single-homing PEs benefit from load balancing

- procedures related to processing Ethernet A-D route for the purpose of Backup-path (8.4 Aliasing and Backup-path), to let single-homing PEs to benefit from the corresponding convergence improvement

9. Determining Reachability to Unicast MAC Addresses

PEs forward packets that they receive based on the destination MAC address. This implies that PEs must be able to learn how to reach a given destination unicast MAC address.

There are two components to MAC address learning, "local learning" and "remote learning":

9.1. Local Learning

A particular PE must be able to learn the MAC addresses from the CEs that are connected to it. This is referred to as local learning.

The PEs in a particular EVPN instance MUST support local data plane learning using standard IEEE Ethernet learning procedures. A PE must be capable of learning MAC addresses in the data plane when it receives packets such as the following from the CE network:

- DHCP requests

- ARP request for its own MAC.

- ARP request for a peer.

Alternatively PEs MAY learn the MAC addresses of the CEs in the control plane or via management plane integration between the PEs and the CEs.

There are applications where a MAC address that is reachable via a given PE on a locally attached Segment (e.g. with ESI X) may move such that it becomes reachable via another PE on another Segment (e.g. with ESI Y). This is referred to as a "MAC Mobility". Procedures to support this are described in section "MAC Mobility".

9.2. Remote learning

A particular PE must be able to determine how to send traffic to MAC addresses that belong to or are behind CEs connected to other PEs i.e. to remote CEs or hosts behind remote CEs. We call such MAC addresses "remote" MAC addresses.

This document requires a PE to learn remote MAC addresses in the control plane. In order to achieve this, each PE advertises the MAC addresses it learns from its locally attached CEs in the control plane, to all the other PEs in that EVPN instance, using MP-BGP and specifically the MAC Advertisement route.

9.2.1. Constructing the BGP EVPN MAC/IP Address Advertisement

BGP is extended to advertise these MAC addresses using the MAC/IP Advertisement route type in the EVPN NLRI.

The RD MUST be the RD of the EVI that is advertising the NLRI. The procedures for setting the RD for a given EVI are described in section 7.9.

The Ethernet Segment Identifier is set to the ten octet ESI described in section "Ethernet Segment".

The Ethernet Tag ID may be zero or may represent a valid Ethernet Tag ID. This field may be non-zero when there are multiple bridge

domains in the MAC-VRF (i.e., the PE needs to perform qualified
learning for the VLANs in that MAC-VRF).

When the the Ethernet Tag ID in the NLRI is set to a non-zero value,
for a particular bridge domain, then this Ethernet Tag ID may either
be the CE's Ethernet tag value (e.g., CE VLAN ID) or the EVPN
provider's Ethernet tag value (e.g., provider VLAN ID). The latter
would be the case if the CE Ethernet tags (e.g., CE VLAN ID) for a
particular bridge domain are different on different CEs.

The MAC address length field is in bits and it is set to 48. The MAC
address length values other than 48 bits, are outside the scope of
this document. The encoding of a MAC address MUST be the 6-octet MAC
address specified by [802.1D-ORIG] [802.1D-REV].

The IP Address Field is optional. By default, the IP Address Length
field is set to 0 and the IP address field is omitted from the route.
When a valid IP address needs to be advertised, it is then encoded in
this route. When an IP address is present, the IP Address Length
field is in bits and it is set to 32 or 128 bits. Other IP Address
Length values are outside the scope of this document. The encoding of
an IP address MUST be either 4 octets for IPv4 or 16 octets for IPv6.
The length field of EVPN NLRI (which is in octets and is described in
section 7) is sufficient to determine whether an IP address is
encoded in this route and if so, whether the encoded IP address is
IPV4 or IPv6.

The MPLS label1 field is encoded as 3 octets, where the high-order 20
bits contain the label value. The MPLS label1 MUST be downstream
assigned and it is associated with the MAC address being advertised
by the advertising PE. The advertising PE uses this label when it
receives an MPLS-encapsulated packet to perform forwarding based on
the destination MAC address toward the CE. The forwarding procedures
are specified in sections 13 and 14.

A PE may advertise the same single EVPN label for all MAC addresses
in a given EVI. This label assignment is referred to as a per EVI
label assignment. Alternatively, a PE may advertise a unique EVPN
label per <ESI, Ethernet Tag> combination. This label assignment is
referred to as a per <ESI, Ethernet Tag> label assignment. As a third
option, a PE may advertise a unique EVPN label per MAC address. This
label assignment is referred to as a per MAC label assignment. All of
these label assignment methods have their tradeoffs. The choice of a
particular label assignment methodology is purely local to the PE
that originates the route.

Per EVI label assignment requires the least number of EVPN labels,
but requires a MAC lookup in addition to an MPLS lookup on an egress

PE for forwarding. On the other hand, a unique label per <ESI, Ethernet Tag> or a unique label per MAC allows an egress PE to forward a packet that it receives from another PE, to the connected CE, after looking up only the MPLS labels without having to perform a MAC lookup. This includes the capability to perform appropriate VLAN ID translation on egress to the CE.

The MPLS label2 field is an optional field and if it is present, then it is encoded as 3 octets, where the high-order 20 bits contain the label value.

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

The BGP advertisement for the MAC advertisement route MUST also carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs), or may be derived automatically from the Ethernet Tag ID, in the Unique VLAN case, as described in section 8.4.1.1.1.

It is to be noted that this document does not require PEs to create forwarding state for remote MACs when they are learnt in the control plane. When this forwarding state is actually created is a local implementation matter.

9.2.2 Route Resolution

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to the reserved ESI value of 0 or MAX-ESI, then if the receiving PE decides to install forwarding state for the associated MAC address, it MUST be based on the MAC Advertisement route alone.

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to a non-reserved ESI, and the receiving PE is locally attached to the same ESI, then the PE does not alter its forwarding state based on the received route. This ensures that local routes are preferred to remote routes.

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to a non-reserved ESI, then if the receiving PE decides to install forwarding state for the associated MAC address, it MUST be when both the MAC Advertisement route AND the associated set of Ethernet A-D per ES routes have been received. The dependency of MAC routes installation on Ethernet A-D per ES routes, is to ensure that MAC routes don't get accidentally installed during mass withdraw period.

To illustrate this with an example, consider two PEs (PE1 and PE2)

connected to a multi-homed Ethernet Segment ES1. All-Active
redundancy mode is assumed. A given MAC address M1 is learnt by PE1
but not PE2. On PE3, the following states may arise:

T1- When the MAC Advertisement Route from PE1 and the set of Ethernet
A-D per ES routes and Ethernet A-D per EVI routes from PE1 and PE2
are received, PE3 can forward traffic destined to M1 to both PE1 and
PE2.

T2- If after T1, PE1 withdraws its set of Ethernet A-D per ES routes,
then PE3 forwards traffic destined to M1 to PE2 only.

T2'- If after T1, PE2 withdraws its set of Ethernet A-D per ES
routes, then PE3 forwards traffic destined to M1 to PE1 only.

T2''- If after T1, PE1 withdraws its MAC Advertisement route, then
PE3 treats traffic to M1 as unknown unicast.

T3- PE2 also advertises a MAC route for M1 and then PE1 withdraws its
MAC route for M1. PE3 continues forwarding traffic destined to M1
to both PE1 and PE2. In other words, despite M1 withdrawal by PE1,
PE3 forwards the traffic destined to M1 to both PE1 and PE2. This is
because a flow from the CE, resulting in M1 traffic getting hashed to
PE1, can get terminated resulting in M1 to aged out in PE1; however,
M1 can be reachable by both PE1 and PE2.

[10]. ARP and ND

The IP address field in the MAC advertisement route may optionally
carry one of the IP addresses associated with the MAC address. This
provides an option which can be used to minimize the flooding of ARP
or Neighbor Discovery (ND) messages over the MPLS network and to
remote CEs. This option also minimizes ARP (or ND) message processing
on end-stations/hosts connected to the EVPN network. A PE may learn
the IP address associated with a MAC address in the control or
management plane between the CE and the PE. Or, it may learn this
binding by snooping certain messages to or from a CE. When a PE
learns the IP address associated with a MAC address, of a locally
connected CE, it may advertise this address to other PEs by including
it in the MAC Advertisement route. The IP Address may be an IPv4
address encoded using four octets, or an IPv6 address encoded using
sixteen octets. For ARP and ND purposes, the IP Address length field
MUST be set to 32 for an IPv4 address or to 128 for an IPv6 address.

If there are multiple IP addresses associated with a MAC address,
then multiple MAC advertisement routes MUST be generated, one for
each IP address. For instance, this may be the case when there are
both an IPv4 and an IPv6 address associated with the MAC address.

   When the IP address is dissociated with the MAC address, then the MAC
   advertisement route with that particular IP address MUST be
   withdrawn.

   When a PE receives an ARP request for an IP address from a CE, and if
   the PE has the MAC address binding for that IP address, the PE SHOULD
   perform ARP proxy by responding to the ARP request.

10.1 Default Gateway

   When a PE needs to perform inter-subnet forwarding where each subnet
   is represented by a different broadcast domain (e.g., different VLAN)
   the inter-subnet forwarding is performed at layer 3 and the PE that
   performs such function is called the default gateway for the EVPN
   instance. In this case when the PE receives an ARP Request for the IP
   address configured as the default gateway address, the PE originates
   an ARP Reply.

   Each PE that acts as a default gateway for a given EVPN instance MAY
   advertise in the EVPN control plane its default gateway MAC address
   using the MAC/IP advertisement route, and indicates that such route
   is associated with the default gateway. This is accomplished by
   requiring the route to carry the Default Gateway extended community
   defined in [Section 7.8 Default Gateway Extended Community]. The ESI
   field is set to zero when advertising the MAC route with the Default
   Gateway extended community.

   The IP address field of the MAC/IP advertisement route is set to the
   default GW IP address for that subnet (e.g., EVPN instance). For a
   given subnet (e.g., VLAN or EVPN instance), the default GW IP address
   is the same across all the participant PEs. The inclusion of this IP
   address enables the receiving PE to check its configured default GW
   IP address against the one received in the MAC/IP advertisement route
   for that subnet (or EVPN instance) and if there is a discrepancy,
   then the PE SHOULD notify the operator and log an error message.

   Unless it is known a priori (by means outside of this document) that
   all PEs of a given EVPN instance act as a default gateway for that
   EVPN instance, the MPLS label MUST be set to a valid downstream
   assigned label.

   Furthermore, even if all PEs of a given EVPN instance do act as a
   default gateway for that EVPN instance, but only some, but not all,
   of these PEs have sufficient (routing) information to provide inter-
   subnet routing for all the inter-subnet traffic originated within the
   subnet associated with the EVPN instance, then when such PE
   advertises in the EVPN control plane its default gateway MAC address
   using the MAC advertisement route, and indicates that such route is

associated with the default gateway, the route MUST carry a valid downstream assigned label.

If all PEs of a given EVPN instance act as a default gateway for that EVPN instance, and the same default gateway MAC address is used across all gateway devices, then no such advertisement is needed. However, if each default gateway uses a different MAC address, then each default gateway needs to be aware of other gateways' MAC addresses and thus the need for such advertisement. This is called MAC address aliasing since a single default GW can be represented by multiple MAC addresses.

Each PE that receives this route and imports it as per procedures specified in this document follows the procedures in this section when replying to ARP Requests that it receives.

Each PE that acts as a default gateway for a given EVPN instance that receives this route and imports it as per procedures specified in this document MUST create MAC forwarding state that enables it to apply IP forwarding to the packets destined to the MAC address carried in the route.

11. Handling of Multi-Destination Traffic

Procedures are required for a given PE to send broadcast or multicast traffic, received from a CE encapsulated in a given Ethernet Tag (VLAN) in an EVPN instance, to all the other PEs that span that Ethernet Tag (VLAN) in that EVPN instance. In certain scenarios, described in section "Processing of Unknown Unicast Packets", a given PE may also need to flood unknown unicast traffic to other PEs.

The PEs in a particular EVPN instance may use ingress replication, P2MP LSPs or MP2MP LSPs to send unknown unicast, broadcast or multicast traffic to other PEs.

Each PE MUST advertise an "Inclusive Multicast Ethernet Tag Route" to enable the above. The following subsection provides the procedures to construct the Inclusive Multicast Ethernet Tag route. Subsequent subsections describe in further detail its usage.

11.1. Construction of the Inclusive Multicast Ethernet Tag Route

The RD MUST be the RD of the EVI that is advertising the NLRI. The procedures for setting the RD for a given EVPN instance on a PE are described in section 7.9.

The Ethernet Tag ID is the identifier of the Ethernet Tag. It may be

     set to 0 or to a valid Ethernet Tag value.

     The Originating Router's IP address MUST be set to an IP address of
     the PE. This address SHOULD be common for all the EVIs on the PE
     (e.,g., this address may be PE's loopback address). The IP Address
     Length field is in bits.

     The Next Hop field of the MP_REACH_NLRI attribute of the route MUST
     be set to the same IP address as the one carried in the Originating
     Router's IP Address field.

     The BGP advertisement for the Inclusive Multicast Ethernet Tag route
     MUST also carry one or more Route Target (RT) attributes. The
     assignment of RTs described in the [section 7.10](#) MUST be followed.

11.2. P-Tunnel Identification

     In order to identify the P-Tunnel used for sending broadcast, unknown
     unicast or multicast traffic, the Inclusive Multicast Ethernet Tag
     route MUST carry a "PMSI Tunnel Attribute" as specified in [BGP
     MVPN].

     Depending on the technology used for the P-tunnel for the EVPN
     instance on the PE, the PMSI Tunnel attribute of the Inclusive
     Multicast Ethernet Tag route is constructed as follows.

          + If the PE that originates the advertisement uses a
            P-Multicast tree for the P-tunnel for EVPN, the PMSI
            Tunnel attribute MUST contain the identity of the tree
            (note that the PE could create the identity of the
            tree prior to the actual instantiation of the tree).

          + A PE that uses a P-Multicast tree for the P-tunnel MAY
            aggregate two or more EVPN instances (EVIs) present
            on the PE onto the same tree. In this case, in addition
            to carrying the identity of the tree, the PMSI Tunnel
            attribute MUST carry an MPLS upstream assigned label which
            the PE has bound uniquely to the EVI associated with this
            update (as determined by its RTs).

            If the PE has already advertised Inclusive Multicast
            Ethernet Tag routes for two or more EVIs that it now
            desires to aggregate, then the PE MUST re-advertise
            those routes. The re-advertised routes MUST be the same
            as the original ones, except for the PMSI Tunnel attribute
            and the label carried in that attribute.

          + If the PE that originates the advertisement uses ingress

replication for the P-tunnel for EVPN, the route MUST
include the PMSI Tunnel attribute with the Tunnel Type set to
Ingress Replication and Tunnel Identifier set to a routable
address of the PE. The PMSI Tunnel attribute MUST carry a
downstream assigned MPLS label. This label is used to
demultiplex the broadcast, multicast or unknown unicast EVPN
traffic received over a MP2P tunnel by the PE.

+ The Leaf Information Required flag of the PMSI Tunnel
  attribute MUST be set to zero, and MUST be ignored on receipt.

12. Processing of Unknown Unicast Packets

The procedures in this document do not require the PEs to flood
unknown unicast traffic to other PEs. If PEs learn CE MAC addresses
via a control plane protocol, the PEs can then distribute MAC
addresses via BGP, and all unicast MAC addresses will be learnt prior
to traffic to those destinations.

However, if a destination MAC address of a received packet is not
known by the PE, the PE may have to flood the packet. When flooding,
one must take into account "split horizon forwarding" as follows: The
principles behind the following procedures are borrowed from the
split horizon forwarding rules in VPLS solutions [RFC4761] and
[RFC4762]. When a PE capable of flooding (say PEx) receives an
unknown destination MAC address, it floods the frame. If the frame
arrived from an attached CE, PEx must send a copy of that frame on
every Ethernet Segment (belonging to that EVI) for which it is the
DF, other than the Ethernet Segment on which it received the frame.
In addition, the PE must flood the frame to all other PEs
participating in that EVPN instance. If, on the other hand, the frame
arrived from another PE (say PEy), PEx must send a copy of the packet
on each Ethernet Segment (belonging to that EVI) for which it is the
DF. PEx MUST NOT send the frame to other PEs, since PEy would have
already done so. Split horizon forwarding rules apply to unknown MAC
addresses.

Whether or not to flood packets to unknown destination MAC addresses
should be an administrative choice, depending on how learning happens
between CEs and PEs.

The PEs in a particular EVPN instance may use ingress replication
using RSVP-TE P2P LSPs or LDP MP2P LSPs for sending unknown unicast
traffic to other PEs. Or they may use RSVP-TE P2MP or LDP P2MP for
sending such traffic to other PEs.

12.1. Ingress Replication

If ingress replication is in use, the P-Tunnel attribute, carried in the Inclusive Multicast Ethernet Tag routes for the EVPN instance, specifies the downstream label that the other PEs can use to send unknown unicast, multicast or broadcast traffic for that EVPN instance to this particular PE.

The PE that receives a packet with this particular MPLS label MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the PE MUST treat the packet as an unknown unicast packet.

12.2. P2MP MPLS LSPs

The procedures for using P2MP LSPs are very similar to VPLS procedures [VPLS-MCAST]. The P-Tunnel attribute used by a PE for sending unknown unicast, broadcast or multicast traffic for a particular EVPN instance is advertised in the Inclusive Ethernet Tag Multicast route as described in section "Handling of Multi-Destination Traffic".

The P-Tunnel attribute specifies the P2MP LSP identifier. This is the equivalent of an Inclusive tree in [VPLS-MCAST]. Note that multiple Ethernet Tags, which may be in different EVPN instances, may use the same P2MP LSP, using upstream labels [VPLS-MCAST]. This is the equivalent of an Aggregate Inclusive tree in [VPLS-MCAST]. When P2MP LSPs are used for flooding unknown unicast traffic, packet re-ordering is possible.

The PE that receives a packet on the P2MP LSP specified in the PMSI Tunnel Attribute MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the PE MUST treat the packet as an unknown unicast packet.

13. Forwarding Unicast Packets

This section describes procedures for forwarding unicast packets by PEs, where such packets are received from either directly connected CEs, or from some other PEs.

13.1. Forwarding packets received from a CE

When a PE receives a packet from a CE, on a given Ethernet Tag, it must first look up the source MAC address of the packet. In certain environments the source MAC address MAY be used to authenticate the CE and determine that traffic from the host can be allowed into the network. Source MAC lookup MAY also be used for local MAC address learning.

If the PE decides to forward the packet, the destination MAC address
of the packet must be looked up. If the PE has received MAC address
advertisements for this destination MAC address from one or more
other PEs or learned it from locally connected CEs, it is considered
as a known MAC address. Otherwise, the MAC address is considered as
an unknown MAC address.

For known MAC addresses the PE forwards this packet to one of the
remote PEs or to a locally attached CE. When forwarding to a remote
PE, the packet is encapsulated in the EVPN MPLS label advertised by
the remote PE, for that MAC address, and in the MPLS LSP label stack
to reach the remote PE.

If the MAC address is unknown and if the administrative policy on the
PE requires flooding of unknown unicast traffic then:

- The PE MUST flood the packet to other PEs. The PE MUST first
encapsulate the packet in the ESI MPLS label as described in section
8.3. If ingress replication is used, the packet MUST be replicated to
each remote PE with the VPN label being an MPLS label determined as
follows: This is the MPLS label advertised by the remote PE in a PMSI
Tunnel Attribute in the Inclusive Multicast Ethernet Tag route for an
<EVPN instance, Ethernet Tag> combination. The Ethernet Tag in the
route may be the same as the Ethernet Tag associated with the
interface on which the ingress PE receives the packet. If P2MP LSPs
are being used the packet MUST be sent on the P2MP LSP that the PE is
the root of for the Ethernet Tag in the EVPN instance. If the same
P2MP LSP is used for all Ethernet Tags, then all the PEs in the EVPN
instance MUST be the leaves of the P2MP LSP. If a distinct P2MP LSP
is used for a given Ethernet Tag in the EVPN instance, then only the
PEs in the Ethernet Tag MUST be the leaves of the P2MP LSP. The
packet MUST be encapsulated in the P2MP LSP label stack.

If the MAC address is unknown then, if the administrative policy on
the PE does not allow flooding of unknown unicast traffic:

- The PE MUST drop the packet.

13.2. Forwarding packets received from a remote PE

This section described the procedures for forwarding known and
unknown unicast packets received from a remote PE.

13.2.1. Unknown Unicast Forwarding

When a PE receives an MPLS packet from a remote PE then, after
processing the MPLS label stack, if the top MPLS label ends up being
a P2MP LSP label associated with an EVPN instance or in case of

ingress replication the downstream label advertised in the P-Tunnel
attribute, and after performing the split horizon procedures
described in [section 8.3](#):

- If the PE is the designated forwarder of BUM traffic on a
particular set of ESIs for the Ethernet Tag, the default behavior is
for the PE to flood the packet on these ESIs. In other words, the
default behavior is for the PE to assume that for BUM traffic, it is
not required to perform a destination MAC address lookup. As an
option, the PE may perform a destination MAC lookup to flood the
packet to only a subset of the CE interfaces in the Ethernet Tag. For
instance the PE may decide to not flood an BUM packet on certain
Ethernet segments even if it is the DF on the Ethernet segment, based
on administrative policy.

- If the PE is not the designated forwarder on any of the ESIs for
the Ethernet Tag, the default behavior is for it to drop the packet.

13.2.2. Known Unicast Forwarding

If the top MPLS label ends up being an EVPN label that was advertised
in the unicast MAC advertisements, then the PE either forwards the
packet based on CE next-hop forwarding information associated with
the label or does a destination MAC address lookup to forward the
packet to a CE.

14. Load Balancing of Unicast Frames

This section specifies the load balancing procedures for sending
known unicast frames to a multi-homed CE.

14.1. Load balancing of traffic from a PE to remote CEs

Whenever a remote PE imports a MAC advertisement for a given <ESI,
Ethernet Tag> in an EVI, it MUST examine all imported Ethernet A-D
routes for that ESI in order to determine the load-balancing
characteristics of the Ethernet segment.

14.1.1 Single-Active Redundancy Mode

For a given ES, if the remote PE has imported the set of Ethernet A-D
per ES routes from at least one PE, where the "Single-Active" flag in
the ESI Label Extended Community is set, then the remote PE MUST
deduce that the ES is operating in Single-Active redundancy mode. As
such, the MAC address will be reachable only via the PE announcing
the associated MAC Advertisement route - this is referred to as the
primary PE. The other PEs advertising the set of Ethernet A-D per ES
routes for the same ES provide backup paths for that ES, in case the

primary PE encounters a failure, and are referred to as backup PEs.
It should be noted that the primary PE for a given <ES, EVI> is the
DF for that <ES, EVI>.

If the primary PE encounters a failure, it MAY withdraw its set of
Ethernet A-D per ES routes for the affected ES prior to withdrawing
it set of MAC Advertisement routes.

If there is only one backup PE for a given ES, the remote PE MAY use
the primary PE's withdrawal of its set of Ethernet A-D per ES routes
as a trigger to update its forwarding entries, for the associated MAC
addresses, to point towards the backup PE. As the backup PE starts
learning the MAC addresses over its attached ES, it will start
sending MAC Advertisement routes while the failed PE withdraws its
routes. This mechanism minimizes the flooding of traffic during fail-
over events.

If there is more than one backup PE for a given ES, the remote PE
MUST use the primary PE's withdrawal of its set of Ethernet A-D per
ES routes as a trigger to start flooding traffic for the associated
MAC addresses (as long as flooding of unknown unicast is
administratively allowed), as it is not possible to select a single
backup PE.

14.1.2 All-Active Redundancy Mode

For a given ES, if the remote PE has imported the set of Ethernet A-D
per ES routes from one or more PEs and none of them have the "Single-
Active" flag in the ESI Label Extended Community set, then the remote
PE MUST deduce that the ES is operating in All-Active redundancy
mode. A remote PE that receives a MAC advertisement route with non-
reserved ESI SHOULD consider the advertised MAC address to be
reachable via all PEs that have advertised reachability to that MAC
address' EVI/ES via the combination of an Ethernet A-D per EVI route
for that EVI/ES (and Ethernet Tag if applicable) AND an Ethernet A-D
per ES route for that ES. The remote PE MUST use received MAC
Advertisement routes and Ethernet A-D per EVI/per ES routes to
construct the set of next-hops for the advertised MAC address.

Each next-hop comprises an MPLS label stack that is to be used by the
egress PE to forward the packet. This label stack is determined as
follows:

-If the next-hop is constructed as a result of a MAC route then this
label stack MUST be used. However, if the MAC route doesn't exist for
that PE, then the next-hop and MPLS label stack is constructed as a
result of the Ethernet A-D routes. Note that the following
description applies to determining the label stack for a particular

next-hop to reach a given PE, from which the remote PE has received
and imported Ethernet A-D routes that have the matching ESI and
Ethernet Tag as the one present in the MAC advertisement. The
Ethernet A-D routes mentioned in the following description refer to
the ones imported from this given PE.

-If a set of Ethernet A-D per ES routes for that ES AND an Ethernet
A-D route per EVI exist, only then the label from that latter route
must be used.

The following example explains the above.

Consider a CE (CE1) that is dual-homed to two PEs (PE1 and PE2) on a
LAG interface (ES1), and is sending packets with source MAC address
MAC1 on VLAN1 (mapped to EVI1). A remote PE, say PE3, is able to
learn that MAC1 is reachable via PE1 and PE2. Both PE1 and PE2 may
advertise MAC1 in BGP if they receive packets with MAC1 from CE1. If
this is not the case, and if MAC1 is advertised only by PE1, PE3
still considers MAC1 as reachable via both PE1 and PE2 as both PE1
and PE2 advertise a set of Ethernet A-D per ES routes for ES1 as well
as an Ethernet A-D per EVI route for <EVI1, ES1>.

The MPLS label stack to send the packets to PE1 is the MPLS LSP stack
to get to PE1 and the EVPN label advertised by PE1 for CE1's MAC.

The MPLS label stack to send packets to PE2 is the MPLS LSP stack to
get to PE2 and the MPLS label in the Ethernet A-D route advertised by
PE2 for <ES1, VLAN1>, if PE2 has not advertised MAC1 in BGP.

We will refer to these label stacks as MPLS next-hops.

The remote PE (PE3) can now load balance the traffic it receives from
its CEs, destined for CE1, between PE1 and PE2. PE3 may use N-Tuple
flow information to hash traffic into one of the MPLS next-hops for
load balancing of IP traffic. Alternatively PE3 may rely on the
source MAC addresses for load balancing.

Note that once PE3 decides to send a particular packet to PE1 or PE2
it can pick one out of multiple possible paths to reach the
particular remote PE using regular MPLS procedures. For instance, if
the tunneling technology is based on RSVP-TE LSPs, and PE3 decides to
send a particular packet to PE1, then PE3 can choose from multiple
RSVP-TE LSPs that have PE1 as their destination.

When PE1 or PE2 receive the packet destined for CE1 from PE3, if the
packet is a known unicast, it is forwarded to CE1. If it is a BUM
packet then only one of PE1 or PE2 must forward the packet to the CE.
Which of PE1 or PE2 forward this packet to the CE is determined based

on which of the two is the DF.

[14.2](#). Load balancing of traffic between a PE and a local CE

   A CE may be configured with more than one interface connected to
   different PEs or the same PE for load balancing, using a technology
   such as LAG. The PE(s) and the CE can load balance traffic onto these
   interfaces using one of the following mechanisms.

[14.2.1](#). Data plane learning

   Consider that the PEs perform data plane learning for local MAC
   addresses learned from local CEs. This enables the PE(s) to learn a
   particular MAC address and associate it with one or more interfaces,
   if the technology between the PE and the CE supports multi-pathing.
   The PEs can now load balance traffic destined to that MAC address on
   the multiple interfaces.

   Whether the CE can load balance traffic that it generates on the
   multiple interfaces is dependent on the CE implementation.

[14.2.2](#). Control plane learning

   The CE can be a host that advertises the same MAC address using a
   control protocol on all interfaces. This enables the PE(s) to learn
   the host's MAC address and associate it with all interfaces. The PEs
   can now load balance traffic destined to the host on all these
   interfaces. The host can also load balance the traffic it generates
   onto these interfaces and the PE that receives the traffic employs
   EVPN forwarding procedures to forward the traffic.

[15](#). MAC Mobility

   It is possible for a given host or end-station (as defined by its MAC
   address) to move from one Ethernet segment to another; this is
   referred to as 'MAC Mobility' or 'MAC move' and it is different from
   the multi-homing situation in which a given MAC address is reachable
   via multiple PEs for the same Ethernet segment. In a MAC move, there
   would be two sets of MAC Advertisement routes, one set with the new
   Ethernet segment and one set with the previous Ethernet segment, and
   the MAC address would appear to be reachable via each of these
   segments.

   In order to allow all of the PEs in the EVPN instance to correctly
   determine the current location of the MAC address, all advertisements
   of it being reachable via the previous Ethernet segment MUST be
   withdrawn by the PEs, for the previous Ethernet segment, that had

advertised it.

If local learning is performed using the data plane, these PEs will
not be able to detect that the MAC address has moved to another
Ethernet segment and the receipt of MAC Advertisement routes, with
the MAC Mobility extended community attribute, from other PEs serves
as the trigger for these PEs to withdraw their advertisements. If
local learning is performed using the control or management planes,
these interactions serve as the trigger for these PEs to withdraw
their advertisements.

In a situation where there are multiple moves of a given MAC,
possibly between the same two Ethernet segments, there may be
multiple withdrawals and re-advertisements. In order to ensure that
all PEs in the EVPN instance receive all of these correctly through
the intervening BGP infrastructure, it is necessary to introduce a
sequence number into the MAC Mobility extended community attribute.

An implementation MUST handle the scenarios where the sequence number
wraps around to process mobility event correctly.

Every MAC mobility event for a given MAC address will contain a
sequence number that is set using the following rules:

- A PE advertising a MAC address for the first time advertises it
with no MAC Mobility extended community attribute.

- A PE detecting a locally attached MAC address for which it had
previously received a MAC Advertisement route with a different
Ethernet segment identifier advertises the MAC address in a MAC
Advertisement route tagged with a MAC Mobility extended community
attribute with a sequence number one greater than the sequence number
in the MAC mobility attribute of the received MAC Advertisement
route. In the case of the first mobility event for a given MAC
address, where the received MAC Advertisement route does not carry a
MAC Mobility attribute, the value of the sequence number in the
received route is assumed to be 0 for purpose of this processing.

- A PE detecting a locally attached MAC address for which it had
previously received a MAC Advertisement route with the same non-zero
Ethernet segment identifier advertises it with:
    i. no MAC Mobility extended community attribute, if the received
    route did not carry said attribute.

    ii. a MAC Mobility extended community attribute with the sequence
    number equal to the highest of the sequence number(s) in the
    received MAC Advertisement route(s), if the received route(s) is
    (are) tagged with a MAC Mobility extended community attribute.

- A PE detecting a locally attached MAC address for which it had
previously received a MAC Advertisement route with the same zero
Ethernet segment identifier (single-homed scenarios) advertises it
with MAC mobility extended community attribute with the sequence
number set properly. In case of single-homed scenarios, there is no
need for ESI comparison. The reason ESI comparison is done for multi-
homing, is to prevent false detection of MAC move among the PEs
attached to the same multi-homed site.

A PE receiving a MAC Advertisement route for a MAC address with a
different Ethernet segment identifier and a higher sequence number
than that which it had previously advertised, withdraws its MAC
Advertisement route. If two (or more) PEs advertise the same MAC
address with same sequence number but different Ethernet segment
identifiers, a PE that receives these routes selects the route
advertised by the PE with lowest IP address as the best route. If the
PE is the originator of the MAC route and it receives the same MAC
address with the same sequence number that it generated, it will
compare its own IP address with the IP address of the remote PE and
will select the lowest IP. If its own route is not the best one, it
will withdraw the route.

15.1. MAC Duplication Issue

A situation may arise where the same MAC address is learned by
different PEs in the same VLAN because of two (or more hosts) being
mis-configured with the same (duplicate) MAC address. In such
situation, the traffic originating from these hosts would trigger
continuous MAC moves among the PEs attached to these hosts. It is
important to recognize such situation and avoid incrementing the
sequence number (in the MAC Mobility attribute) to infinity. In order
to remedy such situation, a PE that detects a MAC mobility event by
way of local learning starts an M-second timer (default value of M =
180) and if it detects N MAC moves before the timer expires (default
value for N = 5), it concludes that a duplicate MAC situation has
occurred. The PE MUST alert the operator and stop sending and
processing any BGP MAC Advertisement routes for that MAC address till
a corrective action is taken by the operator. The values of M and N
MUST be configurable to allow for flexibility in operator control.
Note that the other PEs in the E-VPN instance will forward the
traffic for the duplicate MAC address to one of the PEs advertising
the duplicate MAC address.

15.2. Sticky MAC addresses

There are scenarios in which it is desired to configure some MAC

addresses as static so that they are not subjected to MAC move. In
such scenarios, these MAC addresses are advertised with MAC Mobility
Extended Community where static flag is set to 1 and sequence number
is set to zero. If a PE receives such advertisements and later learns
the same MAC address(es) via local learning, then the PE MUST alert
the operator.

16. Multicast & Broadcast

   The PEs in a particular EVPN instance may use ingress replication or
   P2MP LSPs to send multicast traffic to other PEs.

16.1. Ingress Replication

   The PEs may use ingress replication for flooding BUM traffic as
   described in section "Handling of Multi-Destination Traffic". A given
   broadcast packet must be sent to all the remote PEs. However a given
   multicast packet for a multicast flow may be sent to only a subset of
   the PEs. Specifically a given multicast flow may be sent to only
   those PEs that have receivers that are interested in the multicast
   flow. Determining which of the PEs have receivers for a given
   multicast flow is done using explicit tracking described below.

16.2. P2MP LSPs

   A PE may use an "Inclusive" tree for sending an BUM packet. This
   terminology is borrowed from [VPLS-MCAST].

   A variety of transport technologies may be used in the SP network.
   For inclusive P-Multicast trees, these transport technologies include
   point-to-multipoint LSPs created by RSVP-TE or mLDP.

16.2.1. Inclusive Trees

   An Inclusive Tree allows the use of a single multicast distribution
   tree, referred to as an Inclusive P-Multicast tree, in the SP network
   to carry all the multicast traffic from a specified set of EVPN
   instances on a given PE. A particular P-Multicast tree can be set up
   to carry the traffic originated by sites belonging to a single EVPN
   instance, or to carry the traffic originated by sites belonging to
   several EVPN instances. The ability to carry the traffic of more than
   one EVPN instance on the same tree is termed 'Aggregation' and the
   tree is called an Aggregate Inclusive P-Multicast tree or Aggregate
   Inclusive tree for short. The Aggregate Inclusive tree needs to
   include every PE that is a member of any of the EVPN instances that
   are using the tree. This implies that a PE may receive BUM traffic

even if it doesn't have any receivers that are interested in
receiving that traffic.

An Inclusive or Aggregate Inclusive tree as defined in this document
is a P2MP tree. A P2MP tree is used to carry traffic only for EVPN
CEs that are connected to the PE that is the root of the tree.

The procedures for signaling an Inclusive tree are the same as those
in [VPLS-MCAST] with the VPLS-AD route replaced with the Inclusive
Multicast Ethernet Tag route. The P-Tunnel attribute [VPLS-MCAST] for
an Inclusive tree is advertised with the Inclusive Multicast Ethernet
Tag route as described in section "Handling of Multi-Destination
Traffic". Note that for an Aggregate Inclusive tree, a PE can
"aggregate" multiple EVPN instances on the same P2MP LSP using
upstream labels. The procedures for aggregation are the same as those
described in [VPLS-MCAST], with VPLS A-D routes replaced by EVPN
Inclusive Multicast Ethernet Tag routes.

17. Convergence

This section describes failure recovery from different types of
network failures.

17.1. Transit Link and Node Failures between PEs

The use of existing MPLS Fast-Reroute mechanisms can provide failure
recovery in the order of 50ms, in the event of transit link and node
failures in the infrastructure that connects the PEs.

17.2. PE Failures

Consider a host CE1 that is dual homed to PE1 and PE2. If PE1 fails,
a remote PE, PE3, can discover this based on the failure of the BGP
session. This failure detection can be in the sub-second range if
BFD is used to detect BGP session failure. PE3 can update its
forwarding state to start sending all traffic for CE1 to only PE2.

17.3. PE to CE Network Failures

If the connectivity between the multi-homed CE and one of the PEs
that it is attached to, fails, the PE MUST withdraw the set of
Ethernet A-D per ES routes that had been previously advertised for
that ES. When the MAC entry on the PE ages out, the PE MUST withdraw
the MAC address from BGP. Note that to aid convergence, the Ethernet
A-D per EVI routes MAY be withdrawn before the MAC routes. This
enables the remote PEs to remove the MPLS next-hop to this particular
PE from the set of MPLS next-hops that can be used to forward traffic

to the CE.

When a Ethernet Tag is decommissioned on an Ethernet segment, then the PE MUST withdraw the Ethernet A-D per EVI route(s) announced for the <ESI, Ethernet Tags> that are impacted by the decommissioning. In addition, the PE MUST also withdraw the MAC advertisement routes that are impacted by the decommissioning.

The Ethernet A-D per ES routes should be used by an implementation to optimize the withdrawal of MAC advertisement routes. When a PE receives a withdrawal of a particular Ethernet A-D route from a PE it SHOULD consider all the MAC advertisement routes, that are learned from the same ESI as in the Ethernet A-D route, from the advertising PE, as having been withdrawn. This optimizes the network convergence times in the event of PE to CE failures.

[18]. Frame Ordering

In a MAC address, if the value of the 1st nibble (bits 8 thorough 5) of the most significant byte of the destination MAC address (which follows the last MPLS label) happens to be 0x4 or 0x6, then the Ethernet frame can be misinterpreted as an IPv4 or IPv6 packet by intermediate P nodes performing ECMP based on deep packet inspection, thus resulting in load balancing packets belonging to the same flow on different ECMP paths and subjecting them to different delays. Therefore, packets belonging to the same flow can arrive at the destination out of order. This out of order delivery can happen during steady state in absence of any failures resulting in significant impact to the network operation.

In order to avoid any such mis-ordering, the following rules are applied:

- If a network uses deep packet inspection for its ECMP, then the control word SHOULD be used when sending EVPN encapsulated packets over a MP2P LSP.

- If a network uses Entropy label [RFC6790], then the control word SHOULD NOT be used when sending EVPN encapsulated packet over a MP2P LSP.

- When sending EVPN encapsulated packets over a P2MP LSP or P2P LSP, then the control world SHOULD NOT be used.

The control word is defined as follows:

```
   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
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |0 0 0 0|   Reserved            |       Sequence Number         |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   In the above diagram the first 4 bits MUST be set to 0. The rest of
   the first 16 bits are reserved for future use. They MUST be set to 0
   when transmitting, and MUST be ignored upon receipt. The next 16 bits
   provide a sequence number that MUST also be set to zero by default.

19. Acknowledgements

   Special thanks to Yakov Rekhter for reviewing this draft several
   times and providing valuable comments and for his very engaging
   discussions on several topics of this draft that helped shape this
   document. We would also like to thank Pedro Marques, Kaushik Ghosh,
   Nischal Sheth, Robert Raszuk, Amit Shukla, and Nadeem Mohammed for
   discussions that helped shape this document. We would also like to
   thank Han Nguyen for his comments and support of this work. We would
   also like to thank Steve Kensil and Reshad Rahman for their reviews.
   We would like to thank Jorge Rabadan for his contribution to section
   5 of this draft. We like to thank Thomas Morin for his review of this
   draft and his contribution of section 8.6. Many thanks to Jakob Heitz
   for his help to improve several sections of this draft.

   We would also like to thank Clarence Filsfils, Dennis Cai, Quaizar
   Vohra, Kireeti Kompella, Apurva Mehta for their contributions to this
   document.

   Last but not least, special thanks to Giles Heron (our WG chair) for
   his detailed review of this document in preparation for WG LC and
   making many valuable suggestions.

20. Security Considerations

   Security considerations discussed in [RFC4761] and [RFC4762] apply to
   this document for MAC learning in data-plane over an Attachment
   Circuit (AC) and for flooding of unknown unicast and ARP messages
   over the MPLS/IP core. Security considerations discussed in [RFC4364]
   apply to this document for MAC learning in control-plane over the
   MPLS/IP core. This section describes additional considerations.

   As mentioned in [RFC4761], there are two aspects to achieving data
   privacy and protecting against denial-of-service attacks in a VPN:

securing the control plane and protecting the forwarding path.
Compromise of the control plane could result in a PE sending customer
data belonging to some EVPN to another EVPN, or black-holing EVPN
customer data, or even sending it to an eavesdropper; none of which
are acceptable from a data privacy point of view. In addition,
compromise of the control plane could result in black-holing EVPN
customer data and could provide opportunities for unauthorized EVPN
data usage (e.g., exploiting traffic replication within a multicast
tree to amplify a denial-of-service attack based on sending large
amounts of traffic).

The mechanisms in this document use BGP for the control plane. Hence,
techniques such as in [RFC5925] help authenticate BGP messages,
making it harder to spoof updates (which can be used to divert EVPN
traffic to the wrong EVPN instance) or withdrawals (denial-of-service
attacks). In the multi-AS methods (b) and (c), this also means
protecting the inter-AS BGP sessions, between the ASBRs, the PEs, or
the Route Reflectors.

Note that [RFC5925] will not help in keeping MPLS labels private --
knowing the labels, one can eavesdrop on EVPN traffic. However, this
requires access to the data path within an SP network, which is
assumed to be composed of trusted nodes/links.

One of the requirements for protecting the data plane is that the
MPLS labels be accepted only from valid interfaces. For a PE, valid
interfaces comprise links from other routers in the PE's own AS. For
an ASBR, valid interfaces comprise links from other routers in the
ASBR's own AS, and links from other ASBRs in ASes that have instances
of a given EVPN. It is especially important in the case of multi-AS
EVPN instances that one accept EVPN packets only from valid
interfaces.

It is also important to help limit malicious traffic into a network
for an imposter MAC address. The mechanism described in section 15.1,
shows how duplicate MAC addresses can be detected and continous false
MAC mobility can be prevented. The mechanism described in section
15.2, shows how MAC addresses can be pinned to a given Ethernet
Segment, such that if they appear behind any other Ethernet Segments,
the traffic for those MAC addresses be prevented from entering the
EVPN network from the other Ethernet Segments.

21. Co-authors

In addition to the authors listed on the front page, the following
individuals have also helped to shape this document:

   Keyur Patel

Samer Salam
Sami Boutros
Cisco

Yakov Rekhter
Ravi Shekhar
Juniper Networks

Florin Balus
Nuage Networks

22. IANA Considerations

This document defines a new NLRI, called "EVPN", to be carried in BGP
using multiprotocol extensions. This NLRI uses the existing AFI of
25 (L2VPN). IANA has assigned it a SAFI value of 70.

23. References

23.1 Normative References

[RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, et. al., February 2006

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service
          (VPLS) Using BGP for Auto-Discovery and Signaling", RFC
          4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service
          (VPLS) Using Label Distribution Protocol (LDP) Signaling",
          RFC 4762, January 2007.

[RFC4271] Y. Rekhter et. al., "A Border Gateway Protocol 4 (BGP-4)",
          RFC 4271, January 2006

[RFC4760] T. Bates et. al., "Multiprotocol Extensions for BGP-4", RFC
          4760, January 2007

23.2 Informative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
          Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC7209] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet
          VPN", draft-ietf-l2vpn-evpn-req-04.txt, July 2013.

[RFC7117] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-
          vpls-mcast-14.txt, July 2013.

   [RFC4684] P. Marques et. al., "Constrained Route Distribution for
            Border Gateway Protocol/MultiProtocol Label Switching
            (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks
            (VPNs)", RFC 4684, November 2006.

   [RFC6790] K. Kompella et. al, "The Use of Entropy Labels in MPLS
            Forwarding", RFC 6790, November 2012.

24. Author's Address

     Ali Sajassi
     Cisco
     Email: sajassi@cisco.com

     Rahul Aggarwal
     Email: raggarwa_1@yahoo.com

     Nabil Bitar
     Verizon Communications
     Email : nabil.n.bitar@verizon.com

     Aldrin Isaac
     Bloomberg
     Email: aisaac71@bloomberg.net

     James Uttaro
     AT&T
     Email: uttaro@att.com

     John Drake
     Juniper Networks
     Email: jdrake@juniper.net

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