Authors: A. Sajassi   LA. Burdet, Ed.   J. Drake   J. Rabadan
         Cisco   Cisco   Juniper   Nokia

BGP MPLS-Based Ethernet VPN

Abstract

This document describes procedures for BGP MPLS-based Ethernet VPNs
(EVPN). The procedures described here meet the requirements
specified in RFC 7209 -- "Requirements for Ethernet VPN (EVPN)".

Note to Readers

RFC EDITOR: please remove this section before publication

The complete and detailed set of all changes between this version
and RFC7432 may be found as an Annotated Diff (rfcdiff) here.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

   Virtual Private LAN Service (VPLS), as defined in [RFC4664],
   [RFC4761], and [RFC4762], is a proven and widely deployed
   technology. However, the existing solution has a number of
   limitations when it comes to multihoming and redundancy, multicast
   optimization, provisioning simplicity, flow-based load balancing,
   and multipathing; these limitations are important considerations for
   Data Center (DC) deployments. [RFC7209] describes the motivation for
   a new solution to address these limitations. It also outlines a set
   of requirements that the new solution must address.

   This document describes procedures for a BGP MPLS-based solution
   called Ethernet VPN (EVPN) to address the requirements specified in
   [RFC7209]. Please refer to [RFC7209] 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.

1.1. Summary of changes from RFC 7432

   This section describes the significant changes between [RFC4762] and
   this document.

     *Updates to Terminology i.a. BD, EVI, Ethernet Tag ID, P-tunnel,
      DF/BDF/NDF, DCB;

     *Added Section 6.4 for description and disambiguation of EVPN
      bridging terminology;

     *Added ES-Import route target auto-derivation for ESI types 0,4,5;

     *Precision of 'encoding' language for all references to 'Label'
      fields;

     *Added Section 7.11 for usage of EVPN Layer 2 Attributes Extended
      Community in EVPN Bridging;

     *Added Section 7.12 proposes relative order-of-magnitude route
      priority and processing to help achieve fast convergence;

     *Corrected Section 8.2.1 to include reference to E-TREE exception;

*Updated [Section 8.5](#) to include Backup- and Non-Designated
  Forwarder roles to DF-Election algorithm, description of those
  roles and signaling updates;

 *Updated [Section 8.5](#) to specify DF Election behaviour for
  Originating IP in different family

 *Added [Section 8.3.1.3](#) for MP2MP MPLS LSPs and updated [Section
  12.2](#);

 *Address conflicts in Best Path algorithm for Default Gateway in
  [Section 10.1.1](#);

 *Update to [Section 14.1.1](#) redundancy mode description;

 *Added [Section 15.3](#) describing a loop detection and protection
  mechanism;

 *Added [Section 18.1](#) describing Flow-label usage and signaling (see
  also new [Section 7.11](#));

 *[Section 19](#) specifies use of Domain-wide Common Block (DCB) for
  several cases;

 *Restructuring, namely [Section 8.5](#) to [Section 5](#), simplify all
  Ethernet Tag ID references to [Section 6](#) ; and

 *Corrected Route Target and other extcomm 'attributes' references
  to 'extended communities';

 *Cross-references and editorial changes; [[RFC7991](#)] and xml2rfc-v3
  update (source).

2. 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](#)].

3. Terminology

BD: Broadcast Domain. In a bridged network, the broadcast domain
  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 where Shared VLAN Learning (SVL) is used per [[IEEE.
  802.1Q_2014](#)].

Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.

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

EVI:
An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN. An EVI may be comprised of one BD (VLAN-based, VLAN Bundle, or Port-based services) or multiple BDs (VLAN-aware Bundle or Port-based VLAN-Aware services).

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

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

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

VID: VLAN Identifier.

Ethernet Tag: Used to represent a BD that is configured on a given ES for the purposes of DF election and <EVI, BD> identification for frames received from the CE. Note that any of the following may be used to represent a BD: VIDs (including Q-in-Q tags), configured IDs, VNIs (Virtual Extensible Local Area Network (VXLAN) Network Identifiers), normalized VIDs, I-SIDs (Service Instance Identifiers), etc., as long as the representation of the BDs is configured consistently across the multihomed PEs attached to that ES.

Ethernet Tag ID: Normalized network wide ID that is used to identify a BD within an EVI and carried in EVPN routes.

LACP: Link Aggregation Control Protocol.

MP2MP: Multipoint to Multipoint.

MP2P: Multipoint to Point.

P2MP: Point to Multipoint.

P2P: Point to Point.

P-tunnel: A tunnel through the network of one or more SPs. In this document, P-tunnels are instantiated as bidirectional multicast distribution trees.

PE: Provider Edge device.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward

traffic to/from that Ethernet segment for a given VLAN, then the
Ethernet segment is defined to be operating in Single-Active
redundancy mode.

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

BUM: Broadcast, unknown unicast, and multicast.

DF: Designated Forwarder.

Backup-DF (BDF): Backup-Designated Forwarder.

Non-DF (NDF): Non-Designated Forwarder.

DCB: Domain-wide Common Block (of labels), as in [I-D.ietf-bess-
mvpn-evpn-aggregation-label].

AC: Attachment Circuit.

4. BGP MPLS-Based EVPN Overview

This section provides an overview of EVPN. An EVPN instance
comprises Customer Edge devices (CEs) that are connected to Provider
Edge devices (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 Label Switched Path (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 (Generic Routing
Encapsulation) tunneling or other IP tunneling can be used between
the PEs. The detailed procedures in 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] [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 [RFC4364]). 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 Multiprotocol BGP (MP-BGP). Control-plane learning enables
load balancing of traffic to and from CEs that are multihomed 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, the Link Layer
Discovery Protocol (LLDP), IEEE 802.1aq, Address Resolution Protocol
(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.
An EVPN instance requires a Route Distinguisher (RD) that is unique
per MAC-VRF and one or more globally unique Route Targets (RTs). A
CE attaches to a BD on a PE, on an Ethernet interface that may be
configured for one or more Ethernet tags. If the Ethernet tags are
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. See for
example Section 7.10.1 which describes deriving automatically the
RT(s) for each EVPN instance from the corresponding VID.

5. **Ethernet Segment**

As indicated in [RFC7209], each Ethernet segment needs a unique
identifier in an EVPN. This section defines how such identifiers are
assigned and how they are encoded for use in EVPN signaling. Later
sections of this document describe the protocol mechanisms that
utilize the identifiers.

When a customer site is connected to one or more PEs via a set of
Ethernet links, then this set of Ethernet links constitutes an
"Ethernet segment". For a multihomed site, each Ethernet segment
(ES) is identified by a unique non-zero identifier called an
Ethernet Segment Identifier (ESI). An ESI is encoded as a 10-octet

integer in line format with the most significant octet sent first. The following two ESI values are reserved:

- ESI 0 denotes a single-homed site.

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

In general, an Ethernet segment SHOULD 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 Designated Forwarder (DF) election.

In a network with managed and non-managed CEs, the ESI has the following format:

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

Where:

T (ESI Type) is a 1-octet field (most significant octet) that specifies the format of the remaining 9 octets (ESI Value). The following six ESI types can be used:

  *Type 0 (T=0x00) - This type indicates an arbitrary 9-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 System MAC address (6 octets). The CE LACP System MAC
      address MUST be encoded in the high-order 6 octets of the ESI
      Value field.

     -CE LACP Port Key (2 octets). The CE LACP port key MUST be
      encoded in the 2 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 the Multiple Spanning Tree
Protocol (MSTP) is used in the bridged LAN, then the value of the
ESI is derived by listening to Bridge PDUs (BPDUs) on the
Ethernet segment. To achieve this, the PE is not required to run
MSTP. 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 MAC address (6 octets). The Root Bridge MAC
   address MUST be encoded in the high-order 6 octets of the ESI
   Value field.

  -Root Bridge Priority (2 octets). The CE Root Bridge Priority
   MUST be encoded in the 2 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 (6 octets). The PE MAC address MUST be
   encoded in the high-order 6 octets of the ESI Value field.

  -Local Discriminator value (3 octets). The Local Discriminator
   value MUST be encoded in the low-order 3 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 (4 octets). The system router ID MUST be encoded in
   the high-order 4 octets of the ESI Value field.

-Local Discriminator value (4 octets). The Local Discriminator
 value MUST be encoded in the 4 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 Autonomous System
(AS)-based ESI Value that can be auto-generated or configured by
the operator. The ESI Value is constructed as follows:

-AS number (4 octets). This is an AS number owned by the system
 and MUST be encoded in the high-order 4 octets of the ESI
 Value field. If a 2-octet AS number is used, the high-order
 extra octets will be 0x0000.

-Local Discriminator value (4 octets). The Local Discriminator
 value MUST be encoded in the 4 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.

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 multihomed 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 a LAG, then
only one of the links between the bridged network and the PEs must
be the active link for a given <ES, EVI>. 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.

6. Ethernet Tag ID

An Ethernet Tag ID is a 32-bit field containing either a 12-bit or
24-bit identifier that identifies a particular broadcast domain
(e.g., a VLAN) in an EVPN instance. The 12-bit identifier is called
the 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 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.

The following subsections discuss the relationship between broadcast
domains (e.g., VLANs), Ethernet Tag IDs (e.g., VIDs), and MAC-VRFs
as well as the setting of the Ethernet Tag ID, in the various EVPN
BGP routes (defined in Section 8), for the different types of
service interfaces described in [RFC7209].

The following Ethernet Tag ID value is reserved:

   *Ethernet Tag ID {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 table corresponding to that VLAN. If the VLAN is
represented by multiple VIDs (e.g., a different VID per Ethernet
segment per PE), then each PE needs to perform VID translation for
frames destined to its Ethernet segment(s). In such scenarios, the
Ethernet frames transported over an 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 ID in all EVPN routes MUST be set to 0.

6.2. VLAN Bundle Service Interface

With this service interface, an EVPN instance corresponds to
multiple broadcast domains (e.g., multiple VLANs); however, only a
single bridge table is maintained per MAC-VRF, which means multiple
VLANs share the same bridge table. This implies that MAC addresses
MUST be unique across all VLANs for that EVI in order 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 table. 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 ID 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 multiple
broadcast domains (e.g., multiple VLANs) with each VLAN having its
own bridge table -- i.e., multiple bridge tables (one per VLAN) are
maintained by a single MAC-VRF corresponding to the EVPN instance.

Broadcast, unknown unicast, or multicast (BUM) traffic is sent only
to the CEs in a given broadcast domain; however, the broadcast
domains within an EVI either MAY each have their own P-Tunnel or MAY
share P-Tunnels -- e.g., all of the broadcast domains in an EVI MAY
share a single P-Tunnel.

In the case where a single VLAN is represented by a single VID and
thus no VID translation is required, an MPLS-encapsulated packet
MUST carry that VID. The Ethernet Tag ID in all EVPN routes MUST be
set to that VID. The advertising PE MAY advertise the MPLS Label1 in
the MAC/IP Advertisement route representing ONLY the EVI or
representing both the Ethernet Tag ID and the EVI. This decision is
only a local matter by the advertising PE (which is also the
disposition PE) and doesn't affect any other PEs.

In the case where a single VLAN is represented by different VIDs on
different CEs and thus VID translation is required, a normalized
Ethernet Tag ID (VID) MUST be carried in the EVPN BGP routes.
Furthermore, the advertising PE advertises the MPLS Label1 in the
MAC/IP Advertisement route representing both the Ethernet Tag ID and
the EVI, so that upon receiving an MPLS-encapsulated packet, it can
identify the corresponding bridge table from the MPLS EVPN label and
perform Ethernet Tag ID translation ONLY at the disposition PE --
i.e., the Ethernet frames transported over the MPLS/IP network MUST
remain tagged with the originating VID, and VID translation is
performed on the disposition PE. The Ethernet Tag ID in all EVPN
routes MUST be set to the normalized Ethernet Tag ID 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 a subset of those described in
[Section 6.3](#).

6.4. EVPN PE Model

Since this document discusses EVPN operation in relationship to MAC-
VRF, EVI, Broadcast Domain (BD), and Bridge Table (BT), it is
important to understand the relationship between these terms.
Therefore, the following PE model is depicted below to illustrate
the relationship among them.

```
    +----------------------------------------------------+
    |                                                    |
    |                 +-----------------+      EVPN PE |
    | Attachment    | +-----------------+             |
    | Circuit(AC1) | |  +----------+     |      MPLS/NVO tnl
  ---------------------*Bridge      |     |         +-----
    |               | |  |Table(BT1)|     |          / \     \
    |               | |  |          |      |<------------------> |Eth|
    |               | |  |  VLAN x  |     |          \ /    /
    |               | |  +----------+     |         +-----
    |               | |    ...          |            |
    |               | |  +----------+    |      MPLS/NVO tnl
    |               | |  |Bridge    |    |         +-----
    |               | |  |Table(BT2)|    |          / \     \
    |               | |  |          |      |<------------------> |Eth|
  ---------------------*  VLAN y  |    |          \ /    /
    |   AC2         | |  +----------+    |         +-----
    |               | |    MAC-VRF1      |            |
    |               +-+    RD1/RT1       |            |
    |                 +-----------------+             |
    |                                                 |
    |                                                 |
    +----------------------------------------------------+
```

                    Figure 1: EVPN PE Model

A tenant configured for an EVPN service instance (i.e, EVI) on a PE,
is instantiated by a single MAC Virtual Routing and Forwarding table
(MAC-VRF) on that PE. A MAC-VRF consists of one or more Bridge
Tables (BTs) where each BT corresponds to a VLAN (broadcast domain -
BD). If a service interface for an EVPN PE is configured in VLAN-
Based mode (i.e., section 6.1), then there is only a single BT per
MAC-VRF (per EVI) - i.e., there is only one tenant VLAN per EVI.
However, if a service interface for an EVPN PE is configured in
VLAN-Aware Bundle mode (i.e., section 6.3), then there are several

BTs per MAC-VRF (per EVI) - i.e., there are several tenant VLANs per
EVI. The relationship among these terms can be summarized as follow:

  *An EVI consists of one or more BDs and a MAC-VRF consists of one
   or more BTs, one for each BD. A BD is identified by an Ethernet
   Tag ID which is typically represented by a single VLAN ID (VID);
   however, it can be represented by multiple VIDs (i.e., Shared
   VLAN Learning (SVL) mode in 802.1Q).

  *In VLAN-based mode, there is one EVI per VLAN and thus one BD/BT
   per VLAN. Furthermore, there is one BT per MAC-VRF.

  *In VLAN-bundle service, it can be considered as analogous to SVL
   mode in 802.1Q i.e., one BD per EVI and one BT per MAC-VRF with
   multiple VIDs representing that BD.

  *In VLAN-aware bundle service, there is one EVI with multiple BDs
   where each BD is represented by a VLAN. Furthermore, there are
   multiple BTs in a single MAC-VRF.

Since a single tenant subnet is typically (and in this document)
represented by a VLAN (and thus supported by a single BT), for a
given tenant there are as many BTs as there are subnets as shown in
the PE model above.

MAC-VRF is identified by its corresponding route target and route
distinguisher. If operating in EVPN VLAN-Based mode, then a
receiving PE that receives an EVPN route with MAC-VRF route target
can identify the corresponding BT; however, if operating in EVPN
VLAN-Aware Bundle mode, then the receiving PE needs both the MAC-VRF
route target and Ethernet Tag ID in order to identify the
corresponding BT.

7. BGP EVPN Routes

This document defines a new BGP Network Layer Reachability
Information (NLRI) called the EVPN NLRI.

The format of the EVPN NLRI is as follows:

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

The Route Type field defines the 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 the 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 Address Family Identifier (AFI) of 25 (L2VPN) and a Subsequent Address Family Identifier (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 Advertisements 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

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

```
            +---------------------------------------+
            |  Route Distinguisher (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.

The MPLS Label field is encoded as 3 octets, where the high-order
20 bits contain the label value.

For procedures and usage of this route, please see Sections 8.2
("Fast Convergence") and 8.4 ("Aliasing and Backup Path").

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, 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 fields are considered to be part of the prefix in the NLRI.
The Ethernet Segment Identifier, MPLS Label1, and MPLS Label2 fields
are to be treated as route attributes as opposed to being part of
the "route". Both the IP and MAC address lengths are in bits.

The MPLS Label1 and MPLS Label2 fields are encoded as 3 octets,
where the high-order 20 bits contain the label value.

For procedures and usage of this route, please see Sections 9
("Determining Reachability to Unicast MAC Addresses") and 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 Address      |
|          (4 or 16 octets)             |
+---------------------------------------+
```

For procedures and usage of this route, please see Sections 11
("Handling of Multi-destination Traffic"), 12 ("Processing of
Unknown Unicast Packets"), and 16 ("Multicast and Broadcast"). 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 Address      |
|          (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
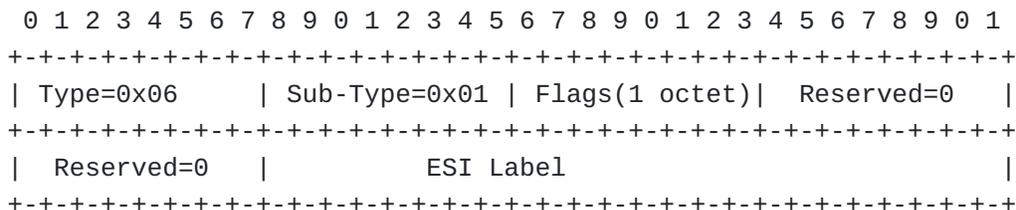fields are considered to be part of the prefix in the NLRI.

7.5. ESI Label Extended Community

This Extended Community is a new transitive Extended Community having a Type field value of 0x06 and the Sub-Type 0x01. It may be advertised along with Ethernet Auto-discovery routes, and it enables split-horizon procedures for multihomed sites as described in Section 8.3 ("Split Horizon"). The ESI Label field represents an ES by the advertising PE, and it is used in split-horizon filtering by other PEs that are connected to the same multihomed Ethernet segment.
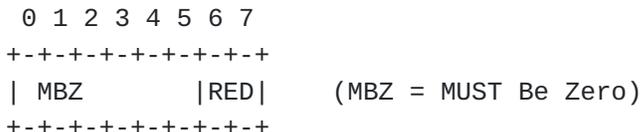
The ESI Label field is encoded as 3 octets, where the high-order 20 bits contain the label value.

The ESI label value MAY be zero if no split-horizon filtering procedures are required in any of the VLANs of the Ethernet Segment. This is the case in [RFC8214] or Ethernet Segments using Local Bias procedures in [I-D.ietf-bess-evpn-mh-split-horizon].

Each ESI Label 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=0x01 | Flags(1 octet)|  Reserved=0   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Reserved=0   |             ESI Label                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

This document creates an IANA registry called "EVPN ESI Multihoming Attributes" (Section 21, where the following bits in Flags octet are initially defined:

```
  0 1 2 3 4 5 6 7
 +-+-+-+-+-+-+-+-+
 | MBZ       |RED|    (MBZ = MUST Be Zero)
 +-+-+-+-+-+-+-+-+

 Name     Meaning
 ----------------------------------------------------------------
 RED      Multihomed site redundancy mode
```

Multihomed site redundancy mode:

RED = 00:
            A value of 00 means that the multihomed site is operating
       in All-Active redundancy mode.

   RED = 01: A value of 01 means that the multihomed site is operating
       in Single-Active redundancy mode.
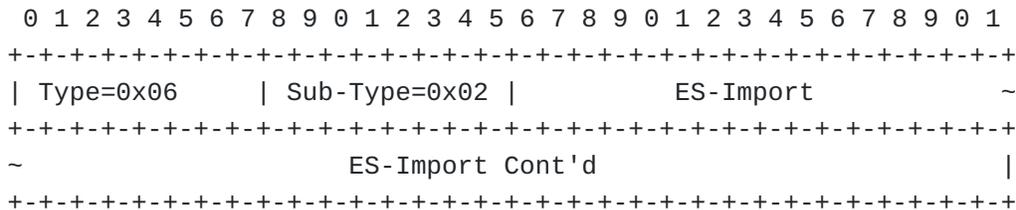
7.6. ES-Import Route Target

   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 multihomed site to import the Ethernet Segment
   routes.

     *The value MAY be derived automatically for ESI Type 0 by encoding
      the high-order 6-octet portion of the 9-octet ESI Value, which
      corresponds to part of the arbitrary value configured, in the ES-
      Import Route Target.

     *The value is derived automatically for ESI Types 1, 2, and 3, by
      encoding the high-order 6-octet portion of the 9-octet ESI Value,
      which corresponds to a MAC address, in the ES-Import Route
      Target.

     *The value MAY be derived automatically for ESI Types 4 and 5, by
      encoding the high-order 6-octet portion of the 9-octet ESI Value,
      which corresponds to a Router ID or AS number (4-octets)
      respectively, and 2-octets of Local Discriminator, in the
      ES-Import Route Target.

   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                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
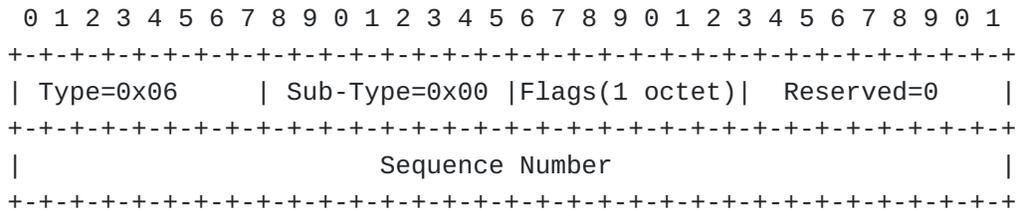
   This document expands the definition of the Route Target extended
   community to allow the value of the high-order octet (Type field) to
   be 0x06 (in addition to the values specified in [RFC4360]). The
   low-order octet (Sub-Type field) value 0x02 indicates that this
   Extended Community is of type "Route Target". The new Type field
   value 0x06 indicates that the structure of this RT is a 6-octet
   value (e.g., a MAC address). A BGP speaker that implements RT
   Constraint [RFC4684] MUST apply the RT Constraint procedures to the
   ES-Import RT as well.

For procedures and usage of this extended community, please see
Section 8.1 ("Multihomed Ethernet Segment Auto-discovery").

7.7. MAC Mobility Extended Community

This Extended Community is a new transitive Extended Community
having a Type field value of 0x06 and the Sub-Type 0x00. It may be
advertised along with MAC/IP 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/IP 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 Section 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). For procedures and usage of this extended
community, please see Section 10.1 ("Default Gateway").

7.9. Route Distinguisher Assignment per MAC-VRF

The Route Distinguisher (RD) MUST be set to the RD of the MAC-VRF
that is advertising the NLRI. An RD MUST be assigned for a given
MAC-VRF on a PE. This RD MUST be unique across all MAC-VRFs 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. In case of VLAN-based or VLAN Bundle services, this
number may also be generated out of the Ethernet Tag ID for the BD
as long as the value does not exceed a length of 16 bits. 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) extended
communities. RTs may be configured (as in IP VPNs) or may be derived
automatically.

If a PE uses RT Constraint, the PE advertises all such RTs using RT
Constraints per [RFC4684]. The use of RT Constraints allows each
EVPN 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 (VLAN ID)

For the "Unique VLAN EVPN" scenario (Section 4), it is highly
desirable to auto-derive the RT from the Ethernet Tag (VLAN ID). The
procedure for performing such auto-derivation is as follows:

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

  *The 12-bit VLAN ID MUST be encoded in the lowest 12 bits of the
   Local Administrator field, with the remaining bits set to zero.

For VLAN-based and VLAN Bundle services, the RT may also be auto-
derived as per the above rules but replacing the 12-bit VLAN ID with
a 16-bit Ethernet Tag ID configured for the BD. If the Ethernet Tag
ID length is 24 bits, the RT for the MAC-VRF can be auto-derived as
per [RFC8365] section 5.1.2.1.

7.11. EVPN Layer 2 Attributes Extended Community

[RFC8214] defines this extended community ("L2-Attr"), to be
included with per-EVI Ethernet A-D routes and mandatory if
multihoming is enabled.

Usage and applicability of this Extended community to Bridging is
clarified here.

```
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |    MBZ        |RSV|RSV|F|C|P|B|  (MBZ = MUST Be Zero)
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following bits in Control Flags from [RFC8214] are listed here
for completeness only:

```
Name      Meaning
    ------------------------------------------------------------------
    P         If set to 1 in multihoming Single-Active scenarios,
              this flag indicates that the advertising PE is the
              primary PE.  MUST be set to 1 for multihoming
              All-Active scenarios by all active PE(s).

    B         If set to 1 in multihoming Single-Active scenarios,
              this flag indicates that the advertising PE is the
              backup PE.

    C         If set to 1, a control word [RFC4448] MUST be present
              when sending EVPN packets to this PE.  It is
              recommended that the control word be included in the
              absence of an entropy label [RFC6790].</t>
```

The bits in Control Flags are extended by the following defined
bits:

```
Name      Meaning
    ------------------------------------------------------------------
    F         If set to 1, a Flow Label MUST be present
              when sending EVPN packets to this PE.
              If set to 0, a Flow Label MUST NOT be present
              when sending EVPN packets to this PE.
```

For procedures and usage of this extended community, with respect to
Control Word and Flow Label, please see Section 18. ("Frame
Ordering").

For procedures and usage of this extended community, with respect to
Primary-Backup bits, please see Section 8.5. ("Designated Forwarder
Election").

7.11.1. EVPN Layer 2 Attributes Partitioning

The information carried in the L2-Attr Extended Community may be
ESI-specific or BD/MAC-VRF-specific. In order to minimize the
processing overhead of configuration-time items such as MTU not
expected to change at runtime based on failures, the Extended
Community from [RFC8214] is partitioned, with a subset of
information carried over each Ethernet A-D per EVI and Inclusive
Multicast routes.

The EVPN Layer 2 Attributes Extended Community, when added to
Inclusive Multicast route:

  *BD/MAC-VRF attributes MTU, Control Word and Flow Label are
   conveyed, and;

  *per-ESI attributes P, B MUST be zero.

```
 +-------------------------------------------+
 | Type (0x06) / Sub-type (0x04) (2 octets)  |
 +-------------------------------------------+
 | Control Flags (2 octets)                  |
 +-------------------------------------------+
 | L2 MTU (2 octets)                         |
 +-------------------------------------------+
 | Reserved (2 octets)                       |
 +-------------------------------------------+


                     1 1 1 1 1
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | MBZ           |  MBZ  |F|C|MBZ|    (MBZ = MUST Be Zero)
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The EVPN Layer 2 Attributes Extended Community is included on
Ethernet A-D per EVI route and:

  *per-ESI attributes P, B are conveyed, and;

  *BD/MAC-VRF attributes MTU, Control Word and Flow Label MUST be
   zero.

```
 +-------------------------------------------+
 | Type (0x06) / Sub-type (0x04) (2 octets)  |
 +-------------------------------------------+
 | Control Flags (2 octets)                  |
 +-------------------------------------------+
 | MBZ (2 octets)                            |
 +-------------------------------------------+
 | Reserved (2 octets)                       |
 +-------------------------------------------+


                     1 1 1 1 1
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | MBZ           |   MBZ   |P|B|    (MBZ = MUST Be Zero)
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Note that in both of the above cases, the values conveyed in this extended community are at the granularity of an individual EVI (or [EVI, BD] for VLAN-aware bundle) and hence may vary for different EVIs.

7.12. Route Prioritization

In order to achieve the Fast Convergence referred to in (Section 8.2 ("Fast Convergence")), BGP speakers SHOULD prioritise advertisement, processing and redistribution of routes based on relative scale of priority vs. expected or average scale.

1. Ethernet A-D per ES (Mass-Withdraw Route Type 1) and Ethernet Segment (Route Type 4) are lower scale and highly convergence affecting, and SHOULD be handled in first order of priority

2. Ethernet A-D per EVI, Inclusive Multicast Ethernet Tag route, and IP Prefix route defined in [RFC9136] are sent for each Bridge or AC at medium scale and may be convergence affecting, and SHOULD be handled in second order of priority

3. MAC advertisement route (zero and nonzero IP portion), Multicast Join Sync and Multicast Leave Sync routes defined in [I-D.ietf-bess-evpn-igmp-mld-proxy] are considered 'individual routes' and very-high scale or of relatively low importance for fast convergence and SHOULD be handled in last order of priority.

8. Multihoming Functions

This section discusses the functions, procedures, and associated BGP routes used to support multihoming in EVPN. This covers both multihomed device (MHD) and multihomed network (MHN) scenarios.

8.1. Multihomed 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 a number unique to the PE.

The Ethernet Segment Identifier (ESI) MUST be set to the 10-octet value 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
multihomed 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 Route
Target, constructed from the ESI.

8.2. Fast Convergence

In EVPN, MAC address reachability is learned 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/IP 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 one or more Ethernet A-D per ES routes for each locally
attached Ethernet segment (refer to Section 8.2.1 below for details
on how these routes are constructed). A PE may need to advertise
more than one Ethernet A-D per ES route for a given ES because the
ES may be in a multiplicity of EVIs and the RTs for all of these
EVIs may not fit into a single route. Advertising a set of Ethernet
A-D per ES routes for the ES allows each route to contain a subset
of the complete set of RTs. Each Ethernet A-D per ES route is
differentiated from the other routes in the set by a different Route
Distinguisher (RD).

Upon a failure in connectivity to the attached segment, the PE
withdraws the corresponding set of Ethernet A-D per ES routes. 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
per ES route for the same segment, then the PE that received the
withdrawal simply invalidates the MAC entries for that segment.
Otherwise, the PE updates its next-hop adjacencies accordingly.

8.2.1. Constructing Ethernet A-D per Ethernet Segment 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 REQUIRED.

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 10-octet entity as described in Section 5 ("Ethernet Segment"). The Ethernet A-D route is not needed when the Segment Identifier is set to 0 (e.g., single-homed scenarios). An exception to this rule is described in [RFC8317].

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, unless DCB allocated labels are used. 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 SHOULD be set to a valid MPLS label value.

8.2.1.1. Ethernet A-D Route Targets

Each Ethernet A-D per ES route MUST carry one or more Route Target (RT) extended communities. 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 multihomed 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-Designated Forwarder (Non-DF) PEs, say PE1, then PE1 will

forward that packet to all or a subset of the other PEs in that EVPN
instance, including the DF PE for that Ethernet segment. In this
case, the DF PE to which the CE is multihomed 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 are operating in Single-Active redundancy mode,
the use of this split-horizon filtering mechanism is highly
recommended because it prevents transient loops at the time of
failure or recovery that would impact the Ethernet segment -- e.g.,
when two PEs think that both are DFs for that segment before the 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 routes. These routes are 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 relies 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 that operates 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 its attached 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 rule for the inclusion of the ESI label in a BUM packet by the
ingress PE operating in Single-Active redundancy mode is 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, which are multihomed to CE1 on
ES1 and operating in All-Active multihoming 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 onto the MPLS label stack 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, from the top MPLS label, the
set of ESIs to which it will replicate the packet 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 that 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 loops during a failure scenario that would impact ES1
(e.g., port or link failure).

8.3.1.2. P2MP MPLS LSPs

The Non-DF PEs that operate 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 their common attached 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 for which the label was distributed. 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 label pop with no other associated action.

The DF PE that operates in Single-Active redundancy mode and that uses P2MP LSPs to send BUM traffic should advertise an upstream assigned ESI label in the set of Ethernet A-D per ES routes for its attached ES, just as described in the previous paragraph.

As an example, consider PE1 and PE2, which are multihomed to CE1 on ES1 and operating in All-Active multihoming mode. Also, consider that PE3 belongs to one of the EVPN instances of ES1. Further, assume that PE1, which is the Non-DF, is 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 on which the packet was received. 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 decapsulates 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 decapsulates 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 loops during a failure scenario that would impact ES1 (e.g., port or link failure).

8.3.1.3. MP2MP MPLS LSPs

The procedures for MP2MP tunnels follow Section 8.3.1.2, with the exceptions described in this section.

When MP2MP tunnels are used, ESI-labels MUST be allocated from a DCB
and the same label must be used by all the PEs attached to the same
Ethernet Segment.

In that way, any egress PE with local Ethernet Segments can identify
the source ES of the received BUM packets.

8.4. Aliasing and Backup Path

In the case where a CE is multihomed to multiple PE nodes, using a
Link Aggregation Group (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/IP Advertisement routes
for these addresses from a single PE, even though multiple PEs are
connected to the multihomed segment. As a result, the remote PEs are
not able to effectively load balance traffic among the PE nodes
connected to the multihomed Ethernet segment. This could be the
case, for example, 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 learned 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/IP Advertisement
route with a non-reserved ESI SHOULD consider the advertised MAC
address to be reachable via all PEs that have advertised
reachability to that MAC address's EVI/ES/Ethernet Tag ID via the
combination of an Ethernet A-D per EVI route for that EVI/ES/
Ethernet Tag ID 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.

The 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 given EVI/ES using the same

combination of Ethernet A-D per EVI route and Ethernet A-D per ES route as discussed 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/IP Advertisement route with a 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.

Please see Section 14.1.1 for a description of the backup paths operation.

8.4.1. Constructing Ethernet A-D per EVPN Instance 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.

The Route Distinguisher (RD) MUST be set per Section 7.9.

The Ethernet Segment Identifier MUST be a 10-octet entity as described in Section 5 ("Ethernet Segment"). The Ethernet A-D route is not needed when the Segment Identifier is set to 0.

The Ethernet Tag ID is set as defined in Section 6.

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

  *One Ethernet A-D route per <ESI, Ethernet Tag ID> tuple per MAC-VRF. This is applicable when the PE uses MPLS-based disposition with VID translation or may be applicable when the PE uses MAC-based disposition with VID translation.

  *One Ethernet A-D route for each <ESI> per MAC-VRF (where the Ethernet Tag ID is set to 0). This is applicable when the PE uses MAC-based disposition or MPLS-based disposition without VID translation.

The usage of the MPLS label is described in Section 14 ("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 per EVI route MUST carry one or more Route Target (RT) extended communities, per Section 7.10.

8.5. Designated Forwarder Election

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

  *Sending broadcast and multicast traffic for a given EVI to that CE.

  *If the flooding of unknown unicast traffic (i.e., traffic for which a PE does not know the destination MAC address, see Section 12) is allowed, sending unknown unicast traffic for a given EVI to that CE.

  *If the multihoming mode is Single-Active, sending (known) unicast traffic for a given EVI to that CE.

Note that this behavior, which allows selecting a DF at the granularity of <ES, EVI> for is the default behavior in this specification.

In this same scenario, a second PE referred to as the Backup-Designated Forwarder (Backup-DF or BDF), is responsible for assuming the role of DF in the event of DF's failure. Until this occurs, the Backup-DF PE is a subset of, and behaves like, a Non-DF PE for all forwarding considerations.

All other PEs, referred to as Non-Designated Forwarder (Non-DF or NDF) are not responsible for any forwarding nor of assuming any functionality from the DF in the event of its failure.

The default procedure for DF election at the granularity of <ES, EVI> is referred to as "service carving". With service carving, it is possible to perform load-balancing of traffic destined to a given segment. The load-balancing procedure carves the set of EVIs on that ES among the PEs nodes evenly such that every PE is the DF for a disjoint and distinct set of EVIs for that ES. The procedure for service carving is as follows according to the DF Election Finite State Machine as defined in Section 2.1 of [RFC8584]:

  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.

  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 should be the 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 "IP Address
   length" and "Originating Router's IP address" fields 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 lowest IP address
   length and numeric value tuple. 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 <ES, EVI> when (V mod N) = i, where V is the
   Ethernet tag for that EVI. For VLAN-Aware Bundle service, then
   the numerically lowest Ethernet tag in that EVI MUST be used in
   the modulo function.
   It should be noted that using the "Originating 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
   multihomed across different ASes if such a need ever arises.

4. For each EVPN instance, a second list of the IP addresses of
   all the PE nodes connected to the Ethernet segment is built.
   The PE which was determined as DF above is removed from that
   ordered candidate list, forming a backup redundancy group of M
   PE nodes. Every remaining PE is then given a second ordinal
   indicating its position in the secondary ordered list according
   to the same criteria as in step 3 above.
   The second ordinals are used to determine which PE nodes will
   be the BDF for a given EVPN instance on the Ethernet segment,
   using the same modulo rule as above, (V mod M) = i.

5. The PE that is elected as a DF for a given <ES, EVI> will
   unblock BUM traffic, or all traffic if in Single-Active mode,
   for that EVI on the corresponding ES. Note that the DF PE
   unblocks BUM traffic in the egress direction towards the
   segment. All Non-DF PEs, including the Backup-DF PE, continue
   to drop multi-destination traffic in the egress direction
   towards that <ES, EVI>.
   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 redundancy group: the
   expected new-DF will be BDF previously calculated in step 5.
   For PE node failure, or upon PE commissioning or
   decommissioning, the PEs re-trigger the service carving. In the
   case of Single-Active multihoming, when a service moves from
   one PE in the redundancy group 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
example, using the IEEE 802.1ak Multiple VLAN Registration
Protocol (MVRP) 'new' declaration.

It is RECOMMENDED that all future DF Election algorithms specify an
algorithm to select one Designated Forwarder (DF) PE, one Backup-DF
PE and a residual number of Non-DF PE(s).

8.6. Signaling Primary and Backup DF Elected PEs

Once the Primary and Backup DF Elected PEs for a given <ES, EVI> are
determined, the multi-homed PEs for that ES will each advertise an
Ethernet A-D per EVI route for that EVI and each will include an L2-
Attr extended community with the P and B bits set to reflect the
advertising PE's role for that EVI.

It should be noted if L2-Attr extended community is included for
All-Active mode, then the P bit must be set for all PEs in the
redundancy group.

8.7. 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 multihoming
procedures can be omitted; however, to allow for single-homing PEs
to fully interoperate with multihoming PEs, some of the multihoming
procedures described above SHOULD be supported even by single-
homing PEs:

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

  *procedures related to processing Ethernet A-D routes for the
   purpose of aliasing (Section 8.4 ("Aliasing and Backup Path")),
   to let single-homing PEs benefit from load balancing

  *procedures related to processing Ethernet A-D routes for the
   purpose of a backup path (Section 8.4 ("Aliasing and Backup
   Path")), to let single-homing PEs 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
  *An ARP Request for its own MAC
  *An 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 "MAC Mobility".
Procedures to support this are described in Section 15 ("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/IP Advertisement route.

9.2.1. Constructing 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 set per Section 7.9.

The Ethernet Segment Identifier is set to the 10-octet ESI described
in Section 5 ("Ethernet Segment").

The Ethernet Tag ID is set as defined in Section 6.

The MAC Address Length field is in bits, and it is set to 48. 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 [IEEE.802.1Q_2014] and [IEEE.802.1D_2004].

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 the 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 MAC-VRF. This label assignment is referred to as a per MAC-VRF label assignment. Alternatively, a PE may advertise a unique EVPN label per <MAC-VRF, Ethernet tag> combination. This label assignment is referred to as a per <MAC-VRF, Ethernet tag> label assignment. As a third option, 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 fourth 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 trade-offs. The choice of a particular label assignment methodology is purely local to the PE that originates the route.

An assignment per MAC-VRF label 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. 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/IP Advertisement route MUST also carry one or more Route Target (RT) extended communities. RTs may be configured (as in IP VPNs) or may be derived automatically in the "Unique VLAN EVPN" case from the Ethernet Tag (VLAN ID), as described in [Section 7.10.1](#).

It is to be noted that this document does not require PEs to create forwarding state for remote MACs when they are learned 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/IP 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/IP Advertisement route alone.

If the Ethernet Segment Identifier field in a received MAC/IP 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/IP 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/IP Advertisement route AND the associated set of Ethernet A-D per ES routes have been received. The dependency of MAC route installation on Ethernet A-D per ES routes is to ensure that MAC routes don't get accidentally installed during a mass withdraw period.

To illustrate this with an example, consider two PEs (PE1 and PE2) connected to a multihomed Ethernet segment ES1. All-Active redundancy mode is assumed. A given MAC address M1 is learned by PE1 but not PE2. On PE3, the following states may arise:

T1 When the MAC/IP 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/IP 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 being 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/IP Advertisement route may optionally carry one of the IP addresses associated with the MAC address. This provides an option that 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/IP Advertisement route. The IP address may be an IPv4 address encoded using 4 octets or an IPv6 address encoded using 16 octets. For ARP and ND purposes, the IP Address Length field MUST be set to 32 for an IPv4 address or 128 for an IPv6 address.

If there are multiple IP addresses associated with a MAC address, then multiple MAC/IP 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 same MAC address for dual-IP-stack scenarios. When the IP address is dissociated with the MAC address, then the MAC/IP Advertisement route with that particular IP address MUST be withdrawn.

Note that a MAC-only route can be advertised along with, but independent from, a MAC/IP route for scenarios where the MAC learning over an access network/node is done in the data plane and independent from ARP snooping that generates a MAC/IP route. In such

scenarios, when the ARP entry times out and causes the MAC/IP to be
withdrawn, then the MAC information will not be lost. In scenarios
where the host MAC/IP is learned via the management or control
plane, then the sender PE may only generate and advertise the MAC/IP
route. If the receiving PE receives both the MAC-only route and the
MAC/IP route, then when it receives a withdraw message for the MAC/
IP route, it MUST delete the corresponding entry from the ARP table
but not the MAC entry from the MAC-VRF table, unless it receives a
withdraw message for the MAC-only route.

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.

In the same way, when a PE receives a Neighbor Solicitation for an
IP address from a CE, the PE SHOULD perform ND proxy and respond if
the PE has the binding information for the IP.

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., a different
VLAN), the inter-subnet forwarding is performed at Layer 3, and the
PE that performs such a 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 each such PE indicates
that such a 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 gateway IP address for that subnet (e.g., an EVPN instance).
For a given subnet (e.g., a VLAN or EVPN instance), the default
gateway 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 gateway 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 a PE advertises in the EVPN control plane its default gateway MAC address using the MAC/IP Advertisement route and indicates that such a 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 an advertisement. This is called MAC address aliasing, since a single default gateway 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.

10.1.1. Best Path selection for Default Gateway

Default gateway MAC address that is assigned to an IRB interface (for a subnet) in a PE MUST be unique in context of that subnet. In other words, the same MAC address cannot be used by a host either intentionally or accidently. Therefore, in case such conflicts arises, there needs to be scheme to detect it and resolve it. In order to properly detect such conflicts, the following BGP best path selection MUST be applied.

  *When comparing two routes, the route which has Default Gateway
   extended community is preferred over a route which does not have
   the extended comunity. The PE that has advertised the MAC route
   withnout Default Gateway extended community, upon receiving the
   route with Default Gateway extended community, SHALL withdraw its
   route and raise an alarm.

*When comparing two routes where both routes have the Default
    Gateway extended community, normal BGP best path processing is be
    applied.

   *When comparing local and remote routes with Default Gateway
    extended community, the local route is always preferred.

   *MAC Mobility extended community SHALL NOT be attached to routes
    which also have Default Gateway extended community on the sending
    side and SHALL be ignored on the receiving side.

11. Handling of Multi-destination Traffic

   Procedures are required for a given PE to flood broadcast or
   multicast traffic received from a CE and with a given Ethernet tag
   to the other PEs in the associated [EVI, BD] (EVPN instance). In
   certain scenarios, as described in Section 12 ("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 its usage in further detail.

11.1. Constructing Inclusive Multicast Ethernet Tag Route

   The RD MUST be set per Section 7.9.

   The Ethernet Tag ID is set as defined in Section 6.

   The Originating Router's IP Address field value MUST be set to an IP
   address of the PE that should be common for all the EVIs on the PE
   (e.g., this address may be the 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 IPv4 or IPv6 address of the advertising PE.

   The BGP advertisement for the Inclusive Multicast Ethernet Tag route
   MUST also carry one or more Route Target (RT) extended communities.
   The assignment of RTs as described in 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 Provider Multicast Service Interface
(PMSI) Tunnel attribute as specified in [RFC6514].

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 Broadcast Domains (BDs) 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 label,
   which the PE has bound uniquely to the BD associated with this
   update (as determined by its RTs and Ethernet Tag ID). The
   assigned MPLS label is upstream allocated unless the procedures
   in section 19 (Use of Domain-wide Common Block (DCB) Labels) are
   followed. If the PE has already advertised Inclusive Multicast
   Ethernet Tag routes for two or more BDs 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 the 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 an MP2P tunnel by the PE.

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 learned
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] [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 or MP2MP LSPs are very similar to the VPLS procedures described in [RFC7117]. 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 Multicast Ethernet Tag route as described in Section 11 ("Handling of Multi-destination Traffic").

The P-tunnel attribute specifies the P2MP or MP2MP LSP identifier. This is the equivalent of an Inclusive tree as described in [RFC7117]. Note that multiple BDs in the same or different EVIs may use the same P2MP or MP2MP LSP, using upstream labels [RFC7117] or DCB labels [I-D.ietf-bess-mvpn-evpn-aggregation-label]. This is the equivalent of an Aggregate Inclusive tree [RFC7117]. When P2MP or MP2MP LSPs are used for flooding unknown unicast traffic, packet reordering is possible.

The PE that receives a packet on the P2MP or MP2MP 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 some other PEs.

13.1. Forwarding Packets Received from a CE

When a PE receives a packet from a CE with a given Ethernet Tag, it must first look up the packet's source MAC address. In certain environments that enable MAC security, the source MAC address MAY be used to validate the host identity 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 has learned it from locally connected CEs, the MAC address is considered a known MAC address. Otherwise, it is considered 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 the MPLS label
advertised by the remote PE in a PMSI Tunnel attribute in the
Inclusive Multicast Ethernet Tag route for the [EVI, BD]
associated with the received packet's Ethernet tag.
If P2MP LSPs are being used, the packet MUST be sent on the P2MP
LSP of which the PE is the root, for the [EVI, BD] associated
with the received packet's Ethernet tag. If the same P2MP LSP is
used for all the BD's in the EVI, then all the PEs in the EVI
MUST be the leaves of the P2MP LSP. If a different P2MP LSP is
used for a given BD in the EVI, then only the PEs in that BD 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 describes 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 the 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 ESes for the [EVI, BD], the default behavior is
    for the PE to flood that traffic to these ESes. 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 these ESes. For instance,
    the PE may decide to not flood a 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 for any ES associated
    with the [EVI, BD], the default behavior is for it to drop the
    BUM traffic.

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 Packets

This section specifies the load-balancing procedures for sending
known unicast packets to a multihomed CE.

14.1. Load Balancing of Traffic from a PE to Remote CEs

When a remote PE imports a MAC/IP Advertisement route for a given ES
in a MAC-VRF, 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 a 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 that remote PE MUST
deduce that the ES is operating in Single-Active redundancy mode.

This means that for a given [EVI, BD], a given MAC address is only
reachable only via the PE announcing the associated MAC/IP
Advertisement route - this PE will also have advertised an Ethernet
A-D per EVI route for that [EVI, BD] with an L2-Attr extended
community in which the P bit is set. I.e., the Primary DF Elected PE
is also responsible for sending known unicast frames to the CE and
receiving unicast and BUM frames from it. Similarly, the Backup DF
Elected PE will have advertised an Ethernet AD per EVI route for
[EVI, BD] with an L2-Attr extended community in which the B bit is
set.

If the Primary DF Elected PE loses connectivity to the CE it SHOULD
withdraw its set of Ethernet A-D per ES routes for the affected ES
prior to withdrawing the affected MAC/IP Advertisement routes. The
Backup DF Elected PE (which is now the Primary DF Elected PE) needs
to advertise an Ethernet A-D per EVI route for [EVI, BD] with an L2-
Attr extended community in which the P bit is set. Furthermore, the
new Backup DF Elected PE needs to advertise an Ethernet A-D per EVI
route for [EVI, BD] with an L2-Attr extended community in which the
B bit is set.

A remote PE SHOULD use the Primary DF Elected 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 at the
Backup DF Elected PE. As the Backup DF Elected PE starts learning
the MAC addresses over its attached ES, it will start sending MAC/IP
Advertisement routes while the failed PE withdraws its routes. This
mechanism minimizes the flooding of traffic during fail-over events.

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/IP Advertisement
route with a non-reserved ESI SHOULD consider the advertised MAC
address to be reachable via all PEs that have advertised
reachability to that MAC address's EVI/ES/Ethernet Tag ID via the
combination of an Ethernet A-D per EVI route for that EVI/ES/
Ethernet Tag ID AND an Ethernet A-D per ES route for that ES. The
remote PE MUST use received MAC/IP 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 to
reach a given egress PE and allow it to forward a packet. The
portion of the MPLS label stack that is to be used by that egress PE
to forward a packet is constructed by the remote PE as follows:

  *If a MAC/IP Advertisement route was received from that PE, then
   its label stack MUST be used in the next hop.

  *Otherwise, the label stack from the Ethernet A-D per EVI route
   that matches the MAC address' EVI/ES/Ethernet Tag ID MUST be used
   in the next hop.

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 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 (at the top of the stack) followed by 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 (at the top of the stack) followed by the MPLS label in the Ethernet A-D route advertised by PE2 for <EVI1, ES1>, 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 receives the packet destined for CE1 from PE3, if the packet is a known unicast, it is forwarded to CE1.

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 a 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 multipathing. 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 multihoming 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/IP 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/IP Advertisement routes,
with the MAC Mobility extended community, 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, introducing a sequence number
into the MAC Mobility extended community is necessary.

In order to process mobility events correctly, an implementation
MUST handle scenarios in which sequence number wraparound occurs.

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.

  *A PE detecting a locally attached MAC address for which it had
   previously received a MAC/IP Advertisement route with a different
   Ethernet segment identifier advertises the MAC address in a MAC/
   IP Advertisement route tagged with a MAC Mobility extended

community with a sequence number one greater than the sequence
number in the MAC Mobility extended community of the received
MAC/IP Advertisement route. In the case of the first mobility
event for a given MAC address, where the received MAC/IP
Advertisement route does not carry a MAC Mobility extended
community, the value of the sequence number in the received route
is assumed to be 0 for the purpose of this processing.

*A PE detecting a locally attached MAC address for which it had
 previously received a MAC/IP Advertisement route with the same
 non-zero Ethernet segment identifier advertises it with:

   1. no MAC Mobility extended community, if the received route
      did not carry said extended community.

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

*A PE detecting a locally attached MAC address for which it had
 previously received a MAC/IP Advertisement route with the same
 zero Ethernet segment identifier (single-homed scenarios)
 advertises it with a MAC Mobility extended community with the
 sequence number set properly. In the case of single-homed
 scenarios, there is no need for ESI comparison. ESI comparison is
 done for multihoming in order to prevent false detection of MAC
 moves among the PEs attached to the same multihomed site.

A PE receiving a MAC/IP 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/IP
Advertisement route. If two (or more) PEs advertise the same MAC
address with the same sequence number but different Ethernet segment
identifiers, a PE that receives these routes selects the route
advertised by the PE with the 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
misconfigured with the same (duplicate) MAC address. In such a
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 a situation and avoid incrementing the
sequence number (in the MAC Mobility extended community) to
infinity. In order to remedy such a situation, a PE that detects a
MAC mobility event via local learning starts an M-second timer (with
a default value of M = 180), and if it detects N MAC moves before
the timer expires (with a default value of 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/IP
Advertisement routes for that MAC address until 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 EVPN 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 moves. In
such scenarios, these MAC addresses are advertised with a MAC
Mobility extended community where the static flag is set to 1 and
the 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.

15.3. Loop Protection

The EVPN MAC Duplication procedure in Section 15.1 prevents an
endless EVPN MAC/IP route advertisement exchange for a duplicate MAC
between two (or more) PEs. While this helps the control plane
settle, in case there is backdoor link (loop) between two or more
PEs attached to the same BD, BUM frames being sent by a CE are still
endlessly looped within the BD through the backdoor link and among
the PEs. This may cause unpredictable issues in the CEs connected to
the affected BD.

The EVPN MAC Duplication Mechanism in Section 15.1 MAY be extended
with a Loop-protection action that is applied on the duplicate-MAC
addresses. This additional mechanism resolves loops created by
accidental or intentional backdoor links and SHOULD be enabled in
all the PEs attached to the BD.

After following the procedure in Section 15.1, when a PE detects a
MAC M as duplicate, the PE behaves as follows:

a)
Stops advertising M and logs a duplicate event.

b)   Initializes a retry-timer, R seconds.

c)   Since Loop Protection is enabled, the PE executes a Loop
Protection action, which we refer to as "Black-Holing" M.

When the PE programs M as a Black-Hole MAC in the Bridge Table, M is
no longer associated to the backdoor Attachment Circuit (AC), but to
a Black-Hole destination.

At this point and while M is in Black-Hole state:

a)   If a new frame is received (from the EVPN network or the
backdoor AC) with MAC SA = M, the PE identifies M to be Black-
Holed and discards the frame, ending the loop.

b)   Optionally, instead of simply discarding the frame with MAC SA
= M, the PE MAY bring down the AC on which the offending frame
is seen last.

c)   Optionally, any frame that arrives at the PE with MAC DA = M
SHOULD be discarded too.

When the retry-timer R for M expires, the PE flushes M from the
Bridge Table and the process is restarted. In general, a Black-Hole
MAC M can be flushed from the Bridge Table if any of the following
events occur:

  *Retry-timer R for duplicate-MAC M expires (as discussed). R is
   initialized when M is detected as duplicate-MAC. Its value is
   configurable and SHOULD be at least three times the EVPN MAC
   Duplication M-timer window.

  *The operator manually flushes a Black-Hole MAC M. This should be
   done only if the conditions under which M was identified as
   duplicate have been cleared.

  *The remote PE withdraws the MAC/IP route for M and there are no
   other remote MAC/IP routes for M.

  *The remote PE sends a MAC/IP route update for M with the
   sticky-bit set (in the MAC Mobility extended community).

16. Multicast and Broadcast

The PEs in a particular EVPN instance may use ingress replication or
P2MP or MP2MP 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 11 ("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 the procedures of [I-D.ietf-bess-evpn-igmp-mld-proxy].

16.2. P2MP or MP2MP LSPs

A PE may use an "Inclusive" tree for sending a BUM packet. This terminology is borrowed from [RFC7117].

A variety of transport technologies may be used in the service provider (SP) network. For Inclusive P-multicast trees, these transport technologies include point-to-multipoint LSPs created by RSVP-TE or Multipoint LDP (mLDP) or BIER.

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 or MP2MP 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 [RFC7117], with the VPLS A-D route replaced with the Inclusive Multicast Ethernet Tag route. The P-tunnel attribute [RFC7117] for an Inclusive tree is advertised with the Inclusive Multicast Ethernet Tag route as described in Section 11 ("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 or DCB allocated labels [I-D.ietf-bess-mvpn-
evpn-aggregation-label]. The procedures for aggregation are the same
as those described in [RFC7117], 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 on the order of 50 ms, 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
Bidirectional Forwarding Detection (BFD) is used to detect BGP
session failures. 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 multihomed CE and one of the PEs to
which it is attached fails, the PE MUST withdraw the set of Ethernet
A-D per ES routes that had been previously advertised for that ES.
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 the MAC entry on the PE ages out,
the PE MUST withdraw the MAC address from BGP.

When an EVI is decommissioned on an Ethernet segment the PE MUST
withdraw the Ethernet A-D per EVI route(s) announced for that <EVI,
ES>. In addition, the PE MUST also withdraw the MAC/IP 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/IP Advertisement routes. When a PE
receives a withdrawal of a particular Ethernet A-D route from an
advertising PE, it SHOULD consider all the MAC/IP 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 first nibble (bits 8 through
5) of the most significant octet 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 those packets
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 the absence of any
failures, resulting in significant impact on network operations.

In order to avoid frame misordering described in Section 18, the
following network-wide rules are applied:

  *If a network uses deep packet inspection for its ECMP, then the
   "Preferred PW MPLS Control Word" [RFC4385] MUST be used with the
   value 0 (e.g., a 4-octet field with a value of zero) when sending
   unicast EVPN-encapsulated packets over an MP2P LSP.

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

  *When sending EVPN-encapsulated packets over a P2MP LSP (e.g.,
   using mLDP signaling), then the control word SHOULD be used.

  *If a network uses entropy labels per [RFC6790], then the control
   word SHOULD NOT be used when sending EVPN-encapsulated packets
   over an MP2P LSP.

18.1. Flow Label

Flow label is used to add entropy to divisible flows, and creates
ECMP load-balancing in the network. The Flow Label MAY be used in
EVPN networks to achieve better load-balancing in the network, when
transit nodes perform deep packet inspection for ECMP hashing. The
following rules apply:

  *When F-bit is set to 1, the PE announces the capability of both
   sending and receiving flow label for known unicast. If the PE is
   capable of supporting Flow Label, then :

     -upon receiving the F-bit set (F=1) from a remote PE, it MUST
      send known unicast packets to that PE with Flow labels;

     -alternately, upon receiving the F-bit unset (F=0) from a
      remote PE, it MUST NOT send known unicast packets to that PE
      with Flow labels.

*The Flow Label MUST NOT be used for EVPN-encapsulated BUM
   packets.

  *An ingress PE will push the Flow Label at the bottom of the stack
   of the EVPN-encapsulated known unicast packets sent to an egress
   PE that previously signaled F-bit set to 1.

  *If a PE receives a unicast packet with two labels, then it can
   differentiate between [VPN label + ESI label] and [VPN label +
   Flow label] and there should be no ambiguity between ESI and Flow
   labels even if they overlap. The reason for this is that the
   downstream assigned VPN label for known unicast is different than
   for BUM traffic and ESI label (if present) comes after BUM VPN
   label. Therefore, from the VPN label, the receiving PE knows
   whether the next label is a ESI label or a Flow label - i.e., if
   the VPN label is for known unicast, then the next label MUST be a
   flow label and if the VPN label is for BUM traffic, then the next
   label MUST be an ESI label because BUM packets are not sent with
   Flow labels.

  *When sending EVPN-encapsulated packets over a P2MP LSP (either
   RSVP-TE or mLDP), flow label SHOULD NOT be used. This is
   independant of any F-bit signalling in the L2-Attr Extended
   Community which would still apply to unicast.

19. Use of Domain-wide Common Block (DCB) Labels

   The use of DCB labels as in [I-D.ietf-bess-mvpn-evpn-aggregation-
   label] is RECOMMENDED in the following cases:

  *Aggregate P-multicast trees: A P-multicast tree MAY aggregate the
   traffic of two or more BDs on a given ingress PE. When
   aggregation is needed, DCB Labels [I-D.ietf-bess-mvpn-evpn-
   aggregation-label] MAY be used in the MPLS label field of the
   Inclusive Multicast Ethernet Tag routes PMSI Tunnel Attribute.
   The use of DCB Labels, instead of upstream allocated labels, can
   greatly reduce the number of labels that the egress PEs need to
   process when P-multicast tunnel aggregation is used in a network
   with a large number of BDs.

  *BIER tunnels: As described in [I-D.ietf-bier-evpn], the use of
   labels with BIER tunnels in EVPN networks is similar to aggregate
   tunnels, since the ingress PE uses upstream allocated labels to
   identify the BD. As described in [I-D.ietf-bier-evpn], DCB labels
   can be allocated instead of upstream labels in the PMSI Tunnel
   Attribute so that the number of labels required on the egress PEs
   can be reduced.

  *ESI labels: The ESI labels advertised with EVPN A-D per ES routes
   MAY be allocated as DCB labels in general, and are RECOMMENDED to

be allocated as DCB labels when used in combination with P2MP/
BIER tunnels.

When MP2MP tunnels are used, ESI-labels MUST be allocated from a DCB
and the same label must be used by all the PEs attached to the same
Ethernet Segment. In that way, any egress PE with local Ethernet
Segments can identify the source ES of the received BUM packets.

20. Security Considerations

Security considerations discussed in [RFC4761] and [RFC4762] apply
to this document for MAC learning in the 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 the 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 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 those discussed 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 backbone
options (b) and (c) [RFC4364], this also means protecting the
inter-AS BGP sessions between the Autonomous System Border Routers
(ASBRs), the PEs, or the Route Reflectors.

Further discussion of security considerations for BGP may be found
in the BGP specification itself [RFC4271] and in the security
analysis for BGP [RFC4272]. The original discussion of the use of
the TCP MD5 signature option to protect BGP sessions is found in
[RFC5925], while [RFC6952] includes an analysis of BGP keying and
authentication issues.

Note that [RFC5925] will not help in keeping MPLS labels private --
knowing the labels, one can eavesdrop on EVPN traffic. Such
eavesdropping additionally requires access to the data path within

an SP network. Users of VPN services are expected to take
appropriate precautions (such as encryption) to protect the data
exchanged over a VPN.

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 impostor MAC address. The mechanism described in Section 15.1
shows how duplicate MAC addresses can be detected and continuous
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 can be prevented from
entering the EVPN network from the other Ethernet segments.

21. 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 BGP EVPNs a SAFI value of 70.

IANA has allocated the following EVPN Extended Community sub-types
in [RFC7153], and this document is the only reference for them, in
addition to [RFC7432].

0x00   MAC Mobility   [RFC7432]
0x01   ESI Label   [RFC7432]
0x02   ES-Import Route Target   [RFC7432]

This document creates a registry called "EVPN Route Types". New
registrations will be made through the "RFC Required" procedure
defined in [RFC8126]. The registry has a maximum value of 255.
Registrations carried forward from [RFC7432] are as follows:

0   Reserved   [RFC7432]
1   Ethernet Auto-discovery   [RFC7432]
2   MAC/IP Advertisement   [RFC7432]
3   Inclusive Multicast Ethernet Tag [RFC7432]
4   Ethernet Segment   [RFC7432]

This document creates a registry called "EVPN ESI Multihoming
Attributes" for the 1-octet Flags field in the ESI Label Extended

Community. New registrations will be made through the "RFC Required" procedure defined in [RFC8126].

Initial registrations are as follows:

RED   Multihomed site redundancy mode
         00 = All-Active
         01 = Single-Active

This document requests allocation of bit 3 in the "EVPN Layer 2 Attributes Control Flags" registry with name F:

F   Flow Label MUST be present

22. References

22.1. Normative References

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

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

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

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

[RFC4760]   Bates, T., Chandra, R., Katz, D., and Y. Rekhter,
            "Multiprotocol Extensions for BGP-4", RFC 4760, DOI
            10.17487/RFC4760, January 2007, <https://www.rfc-
            editor.org/info/rfc4760>.

[RFC4761]   Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private
            LAN Service (VPLS) Using BGP for Auto-Discovery and
            Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007,
            <https://www.rfc-editor.org/info/rfc4761>.

[RFC4762]   Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private
            LAN Service (VPLS) Using Label Distribution Protocol

                  (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January
                  2007, <https://www.rfc-editor.org/info/rfc4762>.

   [RFC7153]   Rosen, E. and Y. Rekhter, "IANA Registries for BGP
                  Extended Communities", RFC 7153, DOI 10.17487/RFC7153,
                  March 2014, <https://www.rfc-editor.org/info/rfc7153>.

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

   [RFC8214]   Boutros, S., Sajassi, A., Salam, S., Drake, J., and J.
                  Rabadan, "Virtual Private Wire Service Support in
                  Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August
                  2017, <https://www.rfc-editor.org/info/rfc8214>.

   [RFC8584]   Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake,
                  J., Nagaraj, K., and S. Sathappan, "Framework for
                  Ethernet VPN Designated Forwarder Election
                  Extensibility", RFC 8584, DOI 10.17487/RFC8584, April
                  2019, <https://www.rfc-editor.org/info/rfc8584>.

22.2. Informative References

   [I-D.ietf-bess-evpn-igmp-mld-proxy] Sajassi, A., Thoria, S., Mishra,
                  M. P., Drake, J., and W. Lin, "IGMP and MLD Proxy for
                  EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-
                  evpn-igmp-mld-proxy-18, 16 February 2022, <https://
                  www.ietf.org/archive/id/draft-ietf-bess-evpn-igmp-mld-
                  proxy-18.txt>.

   [I-D.ietf-bess-evpn-mh-split-horizon] Rabadan, J., Nagaraj, K., Lin,
                  W., and A. Sajassi, "EVPN Multi-Homing Extensions for
                  Split Horizon Filtering", Work in Progress, Internet-
                  Draft, draft-ietf-bess-evpn-mh-split-horizon-02, 15
                  October 2021, <https://www.ietf.org/archive/id/draft-
                  ietf-bess-evpn-mh-split-horizon-02.txt>.

   [I-D.ietf-bess-mvpn-evpn-aggregation-label] Zhang, Z. J., Rosen, E.
                  C., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel
                  Aggregation with Common Labels", Work in Progress,
                  Internet-Draft, draft-ietf-bess-mvpn-evpn-aggregation-
                  label-06, 19 April 2021, <https://www.ietf.org/archive/
                  id/draft-ietf-bess-mvpn-evpn-aggregation-label-06.txt>.

   [I-D.ietf-bier-evpn] Zhang, Z., Przygienda, T., Sajassi, A., and J.
                  Rabadan, "EVPN BUM Using BIER", Work in Progress,
                  Internet-Draft, draft-ietf-bier-evpn-04, 2 December 2020,

<https://www.ietf.org/internet-drafts/draft-ietf-bier-evpn-04.txt>.

[IEEE.802.1D_2004] IEEE, "IEEE Standard for Local and metropolitan
                    area networks: Media Access Control (MAC) Bridges", IEEE
                    802.1D-2004, DOI 10.1109/ieeestd.2004.94569, 6 July 2004,
                    <http://ieeexplore.ieee.org/servlet/opac?punumber=9155>.

[IEEE.802.1Q_2014] IEEE, "IEEE Standard for Local and metropolitan
                    area networks--Bridges and Bridged Networks", IEEE
                    802.1Q-2014, DOI 10.1109/ieeestd.2014.6991462, 18
                    December 2014, <http://ieeexplore.ieee.org/servlet/opac?
                    punumber=6991460>.

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

[RFC4385]   Bryant, S., Swallow, G., Martini, L., and D. McPherson,
            "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word
            for Use over an MPLS PSN", RFC 4385, DOI 10.17487/
            RFC4385, February 2006, <https://www.rfc-editor.org/info/
            rfc4385>.

[RFC4664]   Andersson, L., Ed. and E. Rosen, Ed., "Framework for
            Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, DOI
            10.17487/RFC4664, September 2006, <https://www.rfc-
            editor.org/info/rfc4664>.

[RFC4684]   Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk,
            R., Patel, K., and J. Guichard, "Constrained Route
            Distribution for Border Gateway Protocol/MultiProtocol
            Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual
            Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684,
            November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC5925]   Touch, J., Mankin, A., and R. Bonica, "The TCP
            Authentication Option", RFC 5925, DOI 10.17487/RFC5925,
            June 2010, <https://www.rfc-editor.org/info/rfc5925>.

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

[RFC6790]   Kompella, K., Drake, J., Amante, S., Henderickx, W., and
            L. Yong, "The Use of Entropy Labels in MPLS Forwarding",
            RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://
            www.rfc-editor.org/info/rfc6790>.

[RFC6952]   Jethanandani, M., Patel, K., and L. Zheng, "Analysis of
            BGP, LDP, PCEP, and MSDP Issues According to the Keying
            and Authentication for Routing Protocols (KARP) Design
            Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013,
            <https://www.rfc-editor.org/info/rfc6952>.

[RFC7117]   Aggarwal, R., Ed., Kamite, Y., Fang, L., Rekhter, Y.,
            and C. Kodeboniya, "Multicast in Virtual Private LAN
            Service (VPLS)", RFC 7117, DOI 10.17487/RFC7117, February
            2014, <https://www.rfc-editor.org/info/rfc7117>.

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

[RFC7991]   Hoffman, P., "The "xml2rfc" Version 3 Vocabulary", RFC
            7991, DOI 10.17487/RFC7991, December 2016, <https://
            www.rfc-editor.org/info/rfc7991>.

[RFC8126]   Cotton, M., Leiba, B., and T. Narten, "Guidelines for
            Writing an IANA Considerations Section in RFCs", BCP 26,
            RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://
            www.rfc-editor.org/info/rfc8126>.

[RFC8317]   Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J.,
            Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree)
            Support in Ethernet VPN (EVPN) and Provider Backbone
            Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/
            RFC8317, January 2018, <https://www.rfc-editor.org/info/
            rfc8317>.

[RFC8365]   Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R.,
            Uttaro, J., and W. Henderickx, "A Network Virtualization
            Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365,
            DOI 10.17487/RFC8365, March 2018, <https://www.rfc-
            editor.org/info/rfc8365>.

[RFC9136]   Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W.,
            and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN
            (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021,
            <https://www.rfc-editor.org/info/rfc9136>.

Appendix A. Acknowledgments for This Document (2021)

Appendix B. Contributors for This Document (2021)

   In addition to the authors listed on the front page, the following
   co-authors have also contributed to this document:

Appendix C. Acknowledgments from the First Edition (2015)

Special thanks to Yakov Rekhter for reviewing this document several
times and providing valuable comments, and for his very engaging
discussions on several topics of this document 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 document. We would like to thank
Thomas Morin for his review of this document and his contribution of
Section 8.7. Many thanks to Jakob Heitz for his help to improve
several sections of this document.

We would also like to thank Clarence Filsfils, Dennis Cai, Quaizar
Vohra, Kireeti Kompella, and 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 Last Call
and for making many valuable suggestions.

C.1. Contributors from the First Edition (2015)

In addition to the authors listed on the front page, the following
co-authors have also contributed to this document:

Keyur Patel
Samer Salam
Sami Boutros
Cisco

Yakov Rekhter
Ravi Shekhar
Juniper Networks

Florin Balus
Nuage Networks

C.2. Authors from the First Edition (2015)

Original Authors:

Ali Sajassi
Cisco

EMail: sajassi@cisco.com

         Rahul Aggarwal
         Arktan

         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

         EMail: wim.henderickx@alcatel-lucent.com

Authors' Addresses

         Ali Sajassi
         Cisco

         Email: sajassi@cisco.com

         Luc Andre Burdet (editor)
         Cisco

         Email: lburdet@cisco.com

         John Drake
         Juniper

         Email: jdrake@juniper.net

         Jorge Rabadan
         Nokia

         Email: jorge.rabadan@nokia.com