E-TREE Support in EVPN & PBB-EVPN
draft-ietf-bess-evpn-etree-04

Abstract

   The Metro Ethernet Forum (MEF) has defined a rooted-multipoint
   Ethernet service known as Ethernet Tree (E-Tree). [ETREE-FMWK]
   proposes a solution framework for supporting this service in MPLS
   networks. This document discusses how those functional requirements
   can be easily met with (PBB-)EVPN and how (PBB-)EVPN offers a more
   efficient implementation of these functions.

http://www.ietf.org/shadow.html

Table of Contents

1 Introduction

The Metro Ethernet Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet Tree (E-Tree). In an E-Tree service, endpoints are labeled as either Root or Leaf sites. Root sites can communicate with all other sites. Leaf sites can communicate with Root sites but not with other Leaf sites.

[ETREE-FMWK] proposes the solution framework for supporting E-Tree service in MPLS networks. The document identifies the functional components of the overall solution to emulate E-Tree services in addition to Ethernet LAN (E-LAN) services on an existing MPLS network.

[EVPN] is a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reach-ability information over the MPLS/IP network. [PBB-EVPN] combines the functionality of EVPN with [802.1ah] Provider Backbone Bridging for MAC address scalability.

This document discusses how the functional requirements for E-Tree service can be easily met with (PBB-)EVPN and how (PBB-)EVPN offers a more efficient implementation of these functions.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [KEYWORDS].

2 E-Tree Scenarios and EVPN / PBB-EVPN Support

In this section, we will categorize support for E-Tree into three different scenarios, depending on the nature of the site association (Root/Leaf) per PE or per Ethernet Segment:

- Leaf OR Root site(s) per PE

- Leaf OR Root site(s) per AC

- Leaf OR Root site(s) per MAC

2.1 Scenario 1: Leaf OR Root site(s) per PE

In this scenario, a PE may receive traffic from either Root sites OR Leaf sites for a given MAC-VRF/bridge table, but not both

concurrently. In other words, a given EVI on a PE is either
associated with a root or leaf. The PE may have both Root and Leaf
sites albeit for different EVIs.

```
                    +---------+          +---------+
                    |   PE1   |          |   PE2   |
   +---+            |  +---+  |  +------+ |  +---+  |              +---+
   |CE1+---ES1----+--+   |  |  | MPLS |  |  |   +--+----ES2-----+CE2|
   +---+  (Root)  |  |MAC|  |  | /IP  |  |  |MAC|  |   (Leaf)    +---+
                  |  |VRF|  |  |      |  |  |VRF|  |
                  |  |   |  |  |      |  |  |   |  |              +---+
                  |  |   |  |  |      |  |  |   |  +--+----ES3-----+CE3|
                  |  +---+  |  +------+ |  +---+  |   (Leaf)    +---+
                    +---------+          +---------+
```
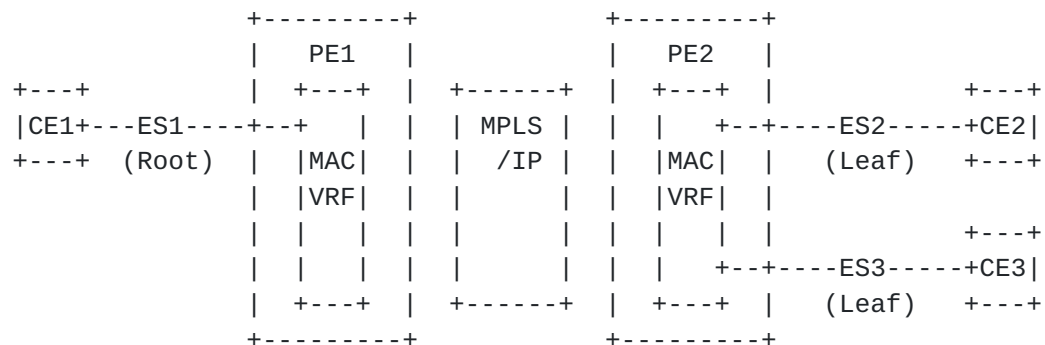
Figure 1: Scenario 1

In such scenario, an EVPN PE implementation MAY provide E-TREE
service using topology constraint among the PEs belonging to the same
EVI. The purpose of this topology constraint is to avoid having PEs
with only Leaf sites importing and processing BGP MAC routes from
each other. To support such topology constrain in EVPN, two BGP
Route-Targets (RTs) are used for every EVPN Instance (EVI): one RT is
associated with the Root sites and the other is associated with the
Leaf sites. On a per EVI basis, every PE exports the single RT
associated with its type of site(s). Furthermore, a PE with Root
site(s) imports both Root and Leaf RTs, whereas a PE with Leaf
site(s) only imports the Root RT. If the number of EVIs is very large
(e.g., more than 64K), then RT type 0 as defined in [RFC4360] SHOULD
be used; otherwise, RT type 2 is sufficient [RFC7153].

2.2 Scenario 2: Leaf OR Root site(s) per AC

In this scenario, a PE receives traffic from either Root OR Leaf
sites (but not both) on a given Attachment Circuit (AC) of an EVI. In
other words, an AC (ES or ES/VLAN) is either associated with a Root
or Leaf (but not both).

```
                    +---------+              +---------+
                    |   PE1   |              |   PE2   |
      +---+         |  +---+  |  +------+  |  +---+  |            +---+
      |CE1+-----ES1----+--+   |  |      |  |  |   +--+---ES2/AC1--+CE2|
      +---+    (Leaf)  | |MAC| |  | MPLS |  |  |MAC| |   (Leaf)   +---+
                    |  |VRF| |  | /IP  |  |  |VRF| |
                    |  |   | |  |      |  |  |   | |            +---+
                    |  |   | |  |      |  |  |   +--+---ES2/AC2--+CE3|
                    |  +---+  |  +------+  |  +---+  |   (Root)   +---+
                    +---------+              +---------+
```
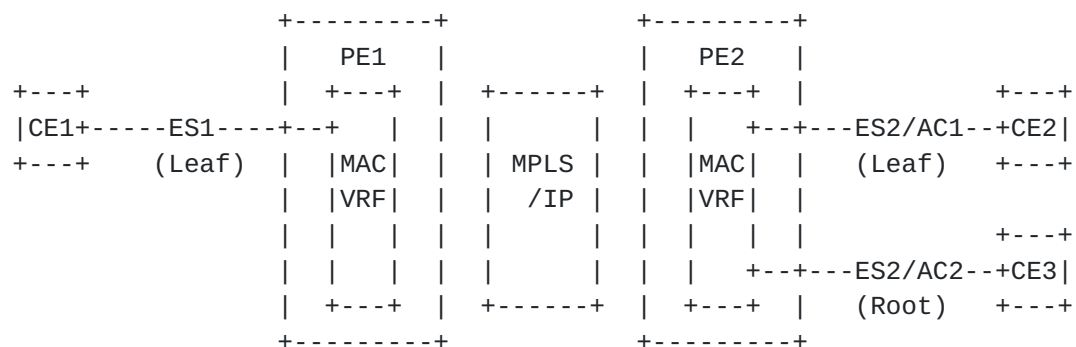
Figure 2: Scenario 2

In this scenario, if there are PEs with only root (or leaf) sites per
EVI, then the RT constrain procedures described in section 2.1 can
also be used here. However, when a Root site is added to a Leaf PE,
then that PE needs to process MAC routes from all other Leaf PEs and
add them to its forwarding table. For this scenario, if for a given
EVI, the majority of PEs will eventually have both Leaf and Root
sites attached, even though they may start as Root-only or Leaf-only
PEs, then it is recommended to use a single RT per EVI and avoid
additional configuration and operational overhead.

2.3 **Scenario 3: Leaf OR Root site(s) per MAC**

In this scenario, a PE may receive traffic from both Root AND Leaf
sites on a given Attachment Circuit (AC) of an EVI. Since an
Attachment Circuit (ES or ES/VLAN) carries traffic from both Root and
Leaf sites, the granularity at which Root or Leaf sites are
identifies is on a per MAC address. This scenario is considered in
this draft for EVPN service with only known unicast traffic - i.e.,
there is no BUM traffic.

```
                    +---------+              +---------+
                    |   PE1   |              |   PE2   |
      +---+         |  +---+  |  +------+  |  +---+  |            +---+
      |CE1+-----ES1----+--+   |  |      |  |  |   +--+---ES2/AC1--+CE2|
      +---+    (Root)  | | E | |  | MPLS |  |  | E | | (Leaf/Root)+---+
                    |  | V | |  | /IP  |  |  | V | |
                    |  | I | |  |      |  |  | I | |            +---+
                    |  |   | |  |      |  |  |   +--+---ES2/AC2--+CE3|
                    |  +---+  |  +------+  |  +---+  |   (Leaf)   +---+
                    +---------+              +---------+
```

Figure 3: Scenario 3

3 Operation for EVPN

[EVPN] defines the notion of ESI MPLS label used for split-horizon filtering of BUM traffic at the egress PE. Such egress filtering capabilities can be leveraged in provision of E-TREE 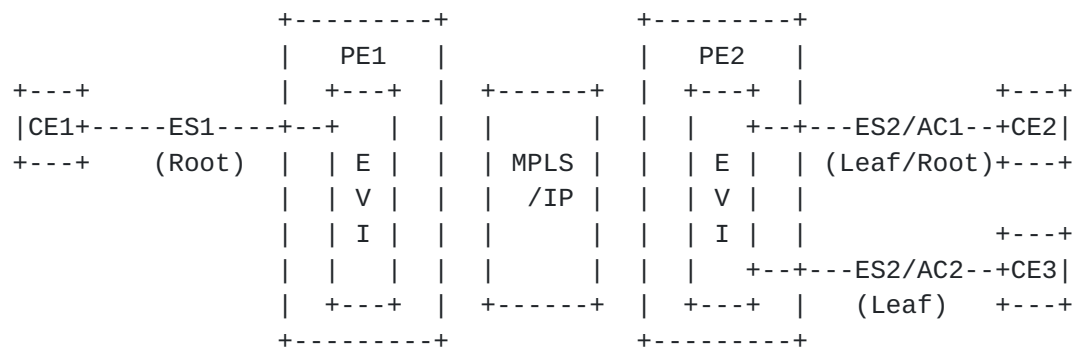services as seen shortly. In other words, [EVPN] has inherent capability to support E-TREE services without defining any new BGP routes but by just defining a new BGP Extended Community for leaf indication as shown later in this document.

3.1 Known Unicast Traffic

Since in EVPN, MAC learning is performed in control plane via advertisement of BGP routes, the filtering needed by E-TREE service for known unicast traffic can be performed at the ingress PE, thus providing very efficient filtering and avoiding sending known unicast traffic over MPLS/IP core to be filtered at the egress PE as done in traditional E-TREE solutions (e.g., E-TREE for VPLS).

To provide such ingress filtering for known unicast traffic, a PE MUST indicate to other PEs what kind of sites (root or leaf) its MAC addresses are associated with by advertising a leaf indication flag (via an Extended Community) along with each of its MAC/IP Advertisement route. The lack of such flag indicates that the MAC address is associated with a root site. This scheme applies to all scenarios described in section 2.

Furthermore, for multi-homing scenario of section 2.2, where an AC is either root or leaf (but not both), the PE MAY advertise leaf indication along with the Ethernet A-D per EVI route. This advertisement is used for sanity checking in control-plane to ensure that there is no discrepancy in configuration among different PEs of the same redundancy group. For example, if a leaf site is multi-homed to PE1 an PE2, and PE1 advertises the Ethernet A-D per EVI corresponding to this leaf site with the leaf-indication flag but PE2 does not, then the receiving PE notifies the operator of such discrepancy and ignore the leaf-indication flag on PE1. In other words, in case of discrepancy, the multi-homing for that pair of PEs is assumed to be in default "root" mode for that <ESI, EVI> or <ESI, EVI/VLAN>. The leaf indication flag on Ethernet A-D per EVI route tells the receiving PEs that all MAC addresses associated with this <ESI, EVI> or <ESI, EVI/VLAN> are from a leaf site. Therefore, if a PE receives a leaf indication for an AC via the Ethernet A-D per EVI route but doesn't receive a leaf indication in the corresponding MAC route, then it notify the operator and ignore the leaf indication on the Ethernet A-D per EVI route.

Tagging MAC addresses with a leaf indication enables remote PEs to perform ingress filtering for known unicast traffic - i.e., on the ingress PE, the MAC destination address lookup yields, in addition to the forwarding adjacency, a flag which indicates whether the target MAC is associated with a Leaf site or not. The ingress PE cross-checks this flag with the status of the originating AC, and if both are Leafs, then the packet is not forwarded.

To support the above ingress filtering functionality, a new E-TREE Extended Community with a Leaf indication flag is introduced [section 5.2]. This new Extended Community MUST be advertised with MAC/IP Advertisement route and MAY be advertised with an Ethernet A-D per EVI route as described above.

3.2 BUM Traffic

For BUM traffic, it is not possible to perform filtering on the ingress PE, as is the case with known unicast, because of the multi-destination nature of the traffic. As such, the solution relies on egress filtering. In order to apply the proper egress filtering, which varies based on whether a packet is sent from a Leaf AC or a root AC, the MPLS-encapsulated frames MUST be tagged with an indication when they originated from a Leaf AC. In other words, leaf indication for BUM traffic is done at the granularity of AC. This can be achieved in EVPN through the use of a MPLS label where it can be used to either identify the Ethernet segment of origin per [RFC 7432] (i.e., ESI label) or it can be used to indicate that the packet is originated from a leaf site (Leaf label).

BUM traffic sent over a P2MP LSP or ingress replication, may need to carry an upstream assigned or downstream assigned MPLS label (respectively) for the purpose of egress filtering to indicate to the egress PEs whether this packet is originated from a leaf AC.

The main difference between downstream and upstream assigned MPLS label is that in case of downstream assigned not all egress PE devices need to receive the label just like ingress replication procedures defined in [RFC7432].

There are four scenarios to consider as follow. In all these scenarios, the imposition PE imposes the right MPLS label associated with the originated Ethernet Segment (ES) depending on whether the Ethernet frame originated from a Root or a Leaf site on that Ethernet Segment (ESI or Leaf label). The mechanism by which the PE identifies whether a given frame originated from a Root or a Leaf site on the segment is based on the Ethernet Tag associated with the frame (e.g., whether the frame received on a leaf or a root AC). Other mechanisms for identifying whether an ingress AC is a root or leaf is beyond the

scope of this document.

3.2.1 BUM traffic originated from a single-homed site on a leaf AC

In this scenario, the ingress PE adds a special MPLS label indicating
a Leaf site. This special Leaf MPLS label, used for single-homing
scenarios, is not on a per ES basis but rather on a per PE basis -
i.e., a single Leaf MPLS label is used for all single-homed ES's on
that PE. This Leaf label is advertised to other PE devices, using a
new EVPN Extended Community called E-TREE Extended Community (section
5.1) along with an Ethernet A-D per ES route with ESI of zero and a
set of Route Targets (RTs) corresponding to all EVIs on the PE with
at least one leaf site per EVI. The set of Ethernet A-D per ES routes
may be needed if the number of Route Targets (RTs) that need to be
sent exceed the limit on a single route per [RFC 7432]. The ESI for
the Ethernet A-D per ES route is set to zero to indicate single-homed
sites.

When a PE receives this special Leaf label in the data path, it
blocks the packet if the destination AC is of type Leaf; otherwise,
it forwards the packet.

3.2.2 BUM traffic originated from a single-homed site on a root AC

In this scenario, the ingress PE does not add any ESI or Leaf label
and it operates per [RFC7432] procedures.

3.2.3 BUM traffic originated from a multi-homed site on a leaf AC

In this scenario, it is assumed that While different ACs (VLANs) on
the same ES could have different root/leaf designation (some being
roots and some being leaves), the same VLAN does have the same
root/leaf designation on all PEs on the same ES. Furthermore, it is
assumed that there is no forwarding among subnets - ie, the service
is EVPN L2 and not EVPN IRB. IRB use case is outside the scope of
this document.

In such scenarios, If a multicast packet is originated from a leaf
AC, then it only needs to carry Leaf label described in section
3.2.1. This label is sufficient in providing the necessary egress
filtering of BUM traffic from getting sent to leaf ACs including the
leaf AC on the same Ethernet Segment.

3.2.4 BUM traffic originated from a multi-homed site on a root AC

In this scenario, both the ingress and egress PE devices follows the
procedure defined in [RFC 7432] for adding and/or processing an ESI
MPLS label.

3.3 E-TREE Traffic Flows for EVPN

Per [ETREE-FMWK], a generic E-Tree service supports all of the
following traffic flows:

- Ethernet Unicast from Root to Roots & Leaf
- Ethernet Unicast from Leaf to Root
- Ethernet Broadcast/Multicast from Root to Roots & Leafs
- Ethernet Broadcast/Multicast from Leaf to Roots

A particular E-Tree service may need to support all of the above
types of flows or only a select subset, depending on the target
application. In the case where unicast flows need not be supported,
the L2VPN PEs can avoid performing any MAC learning function.

In the subsections that follow, we will describe the operation of
EVPN to support E-Tree service with and without MAC learning.

3.3.1 E-Tree with MAC Learning

The PEs implementing an E-Tree service must perform MAC learning when
unicast traffic flows must be supported among Root and Leaf sites. In
this case, the PE with Root sites performs MAC learning in the data-
path over the Ethernet Segments, and advertises reachability in EVPN
MAC Advertisement routes. These routes will be imported by all PEs
for that EVI (i.e., PEs that have Leaf sites as well as PEs that have
Root sites). Similarly, the PEs with Leaf sites perform MAC learning
in the data-path over their Ethernet Segments, and advertise
reachability in EVPN MAC Advertisement routes. For the scenario
described in section 2.1 (or possibly section 2.2), these routes are
imported only by PEs with at least one Root site in the EVI - i.e., a
PE with only Leaf sites will not import these routes. PEs with Root
and/or Leaf sites may use the Ethernet A-D routes for aliasing (in
the case of multi-homed segments) and for mass MAC withdrawal per
[RFC 7432].

To support multicast/broadcast from Root to Leaf sites, either a P2MP
tree rooted at the PE(s) with the Root site(s) or ingress replication
can be used. The multicast tunnels are set up through the exchange of
the EVPN Inclusive Multicast route, as defined in [RFC7432].

To support multicast/broadcast from Leaf to Root sites, ingress
replication should be sufficient for most scenarios where there are
only a few Roots (typically two). Therefore, in a typical scenario, a
root PE needs to support both a P2MP tunnel in transmit direction
from itself to leaf PEs and at the same time it needs to support
ingress-replication tunnels in receive direction from leaf PEs to

itself. In order to signal this efficiently from the root PE, a new
composite tunnel type is defined per section 5.3. This new composite
tunnel type is advertised by the root PE to simultaneously indicate a
P2MP tunnel in transmit direction and an ingress-replication tunnel
in the receive direction for the BUM traffic.

If the number of Roots is large, P2MP tunnels originated at the PEs
with Leaf sites may be used and thus there will be no need to use the
modified PMSI tunnel attribute in section 5.2 for composite tunnel
type.

3.3.2 E-Tree without MAC Learning

The PEs implementing an E-Tree service need not perform MAC learning
when the traffic flows between Root and Leaf sites are only multicast
or broadcast. In this case, the PEs do not exchange EVPN MAC
Advertisement routes. Instead, the Inclusive Multicast Ethernet Tag
(IMET) routes are used to support BUM traffic.

The fields of the IMET route are populated per the procedures defined
in [RFC7432], and the multicast tunnel setup criteria are as
described in the previous section.

Just as in the previous section, if the number of PEs with root sites
are only a few and thus ingress replication is desired from leaf PEs
to these root PEs, then the modified PMSI attribute as defined in
section 5.3 should be used.

4 Operation for PBB-EVPN

In PBB-EVPN, the PE must advertise a Root/Leaf indication along with
each B-MAC Advertisement route, to indicate whether the associated B-
MAC address corresponds to a Root or a Leaf site. Similar to the EVPN
case, this flag will be added to the new E-TREE Extended Community
defined in section [5.2], and advertised with each MAC Advertisement
route.

In the case where a multi-homed Ethernet Segment has both Root and
Leaf sites attached, two B-MAC addresses are allocated and
advertised: one B-MAC address implicitly denoting Root and the other
explicitly denoting Leaf. The former B-MAC address is not advertised
with the E-TREE extended community but the latter B-MAC denoting Leaf
is advertised with the new E-TREE extended community.

The ingress PE uses the right B-MAC source address depending on
whether the Ethernet frame originated from the Root or Leaf site on
that Ethernet Segment. The mechanism by which the PE identifies
whether a given frame originated from a Root or Leaf site on the

segment is based on the Ethernet Tag associated with the frame. Other
mechanisms of identification, beyond the Ethernet Tag, are outside
the scope of this document. It should be noted that support for both
Root and Leaf sites on a single Ethernet Segment requires that the PE
performs the Ethernet Segment split-horizon check on a per Ethernet
Tag basis.

In the case where a multi-homed Ethernet Segment has only Root OR
Leaf sites attached, then a single B-MAC address is allocated and
advertised per segment.

Furthermore, a PE advertises two special global B-MAC addresses: one
for Root and another for Leaf, and tags the Leaf one as such in the
MAC Advertisement route. These B-MAC addresses are used as source
addresses for traffic originating from single-homed segments.

4.1 Known Unicast Traffic

For known unicast traffic, the PEs perform ingress filtering: On the
ingress PE, the C-MAC destination address lookup yields, in addition
to the target B-MAC address and forwarding adjacency, a flag which
indicates whether the target B-MAC is associated with a Root or a
Leaf site. The ingress PE cross-checks this flag with the status of
the originating site, and if both are a Leaf, then the packet is not
forwarded.

4.2 BUM Traffic

For BUM traffic, the PEs must perform egress filtering. When a PE
receives a MAC advertisement route, it updates its Ethernet Segment
egress filtering function (based on the B-MAC source address), as
follows:

- If the MAC Advertisement route indicates that the advertised B-MAC
is a Leaf, and the local Ethernet Segment is a Leaf as well, then the
source B-MAC address is added to the B-MAC filtering list.

- Otherwise, the B-MAC filtering list is not updated.

When the egress PE receives the packet, it examines the B-MAC source
address to check whether it should filter or forward the frame. Note
that this uses the same filtering logic as baseline [PBB-EVPN] and
does not require any additional flags in the data-plane.

The PE places all Leaf Ethernet Segments of a given bridge domain in
a single split-horizon group in order to prevent intra-PE forwarding
among Leaf segments. This split-horizon function applies to BUM

traffic.

5 BGP Encoding

This document defines two new BGP Extended Community for EVPN.

5.1 E-TREE Extended Community

This Extended Community is a new transitive Extended Community having
a Type field value of 0x06 (EVPN) and the Sub-Type 0x04. It is used
for leaf indication of known unicast and BUM traffic. For BUM
traffic, the Leaf Label field is set to a valid MPLS label and this
EC is advertised along with Ethernet A-D per ES route with an ESI of
zero to enable egress filtering on disposition PEs per section 3.2.1
and 3.2.3. For known unicast traffic, the Leaf flag bit is set to one
and this EC is advertised along with MAC/IP Advertisement route per
section 3.1.

The E-TREE Extended Community is encoded as an 8-octet value as
follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type=0x06     | Sub-Type=0x04 | Flags(1 Octet)|  Reserved=0   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Reserved=0   |            Leaf Label                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The low-order bit of the Flags octet is defined as the "Leaf-
Indication" bit. A value of one indicates a Leaf AC/Site.

When this EC is advertised along with MAC/IP Advertisement route, the
Leaf-Indication flag MUST be set to one and Leaf Label is set to
zero. The received PE should ignore Leaf Label and only processes
Leaf-Indication flag. A value of zero for Leaf-Indication flag is
invalid when sent along with MAC/IP advertisement route and an error
should be logged.

When this EC is advertised along with Ethernet A-D per ES route (with
ESI of zero), the Leaf Label MUST be set to a valid MPLS label and
the Leaf-Indication flag should be set to zero. The received PE
should ignore the Leaf-Indication flag. A non-valid MPLS label when
sent along with the Ethernet A-D per ES route, should be logged as an
error.

5.2 **PMSI Tunnel Attribute**

[RFC 6514] defines PMSI Tunnel attribute which is an optional
transitive attribute with the following format:

```
     +---------------------------------+
     |  Flags (1 octet)                |
     +---------------------------------+
     |  Tunnel Type (1 octets)         |
     +---------------------------------+
     |  MPLS Label (3 octets)          |
     +---------------------------------+
     |  Tunnel Identifier (variable)   |
     +---------------------------------+
```

This draft uses all the fields per existing definition except for the
following modifications to the Tunnel Type and Tunnel Identifier:

When receiver ingress-replication label is needed, the high-order bit
of the tunnel type field (C bit - Composite tunnel bit) is set while
the remaining low-order seven bits indicate the tunnel type as
before. When this C bit is set, the "tunnel identifier" field would
begin with a three-octet label, followed by the actual tunnel
identifier for the transmit tunnel. PEs that don't understand the
new meaning of the high-order bit would treat the tunnel type as an
invalid tunnel type. For the PEs that do understand the new meaning
of the high-order, if ingress replication is desired when sending BUM
traffic, the PE will use the the label in the Tunnel Identifier field
when sending its BUM traffic.

6 **Acknowledgement**

We would like to thank Dennis Cai, Antoni Przygienda, and Jeffrey
Zhang for their valueable comments.

7 **Security Considerations**

Same security considerations as [RFC7432].

8 **IANA Considerations**

This document requests the allocation of value 4 in the "EVPN
Extended Community Sub-Types" registry defined in [RFC7153] and
modification of the registry as follow:

```
        SUB-TYPE VALUE      NAME                      Reference

        0x04                E-TREE Extended Community   This document
        6-255               Unassigned
```

9 References

9.1 Normative References

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

   [RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", February,
             2015.

9.2 Informative References

   [ETREE-FMWK] Key et al., "A Framework for E-Tree Service over MPLS
   Network", draft-ietf-l2vpn-etree-frwk-03, work in progress, September
   2013.

   [PBB-EVPN] Sajassi et al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-
   05.txt, work in progress, October, 2013.

   [RFC4360] S. Sangli et al, "BGP Extended Communities Attribute",
   February, 2006.

   [RFC7153] Rosen et al., "IANA Registries for BGP Extended
   Communities", March, 2014.

Authors' Addresses

   Ali Sajassi
   Cisco
   Email: sajassi@cisco.com

   Samer Salam
   Cisco
   Email: ssalam@cisco.com

   Wim Henderickx

Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

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

Aldrin
Bloomberg Issac
Email: aisaac71@bloomberg.net

Sami Boutros
Cisco
Email: sboutros@cisco.com