BESS Workgroup
INTERNET-DRAFT
Intended Status: Standards Track
Updates: RFC7385

A. Sajassi, Ed.
S. Salam
Cisco
J. Drake
Juniper
J. Uttaro
ATT
S. Boutros
VMware
J. Rabadan
Nokia

Expires: June 12, 2017 January 12, 2017

E-TREE Support in EVPN & PBB-EVPN
draft-ietf-bess-evpn-etree-09

Abstract

The Metro Ethernet Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet Tree (E-Tree). A solution framework for supporting this service in MPLS networks is proposed in and RFC called "A Framework for E-Tree Service over MPLS Network". 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. This document makes use of the most significant bit of the scope governed by the IANA registry created by RFC7385, and hence updates that RFC accordingly.

Copyright and License Notice

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.

[RFC7387] 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.

[RFC7432] 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. [RFC7623] 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 root(s) or leaf(s). 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, using tailored BGP Route Target (RT) import/export policies among the PEs belonging to the same EVI, can be used to restrict the communications among Leaf PEs. To restrict the communications among Leaf sites connected to the same PE and belonging to the same EVI, split-horizon filtering is used to block traffic from one Leaf interface to another Leaf interface of a given E-TREE 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.

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 a Root AC or a Leaf AC (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, just like the previous scenario (in section 2.1), two Route Targets (one for Root and another for Leaf) can be used. However, the difference is that on a PE with both Root and Leaf ACs, all remote MAC routes are imported and thus there needs to be a way to differentiate remote MAC routes associated with Leaf ACs versus the ones associated with Root ACs in order to apply the proper ingress filtering.

In order to support such ingress filtering on the ingress PE with both Leaf and Root ACs, one the following two approaches can be used:

A) To use two MAC-VRFs (two bridge tables per VLANs if a given VLAN exists on the PE for both Leaf and Root ACs of an EVI) - one for Root ACs and another for Leaf ACs.

B) To color MAC addresses with Leaf or Root color before distributing them in BGP to other PEs depending on whether they are learned on a Leaf AC or a Root AC.

Maintaining two MAC-VRFs (two bridge tables) per VLAN (when both Leaf and Root ACs exists for that VLAN) would either require two lookups be performed per MAC address in each direction in case of a miss, or duplicating many MAC addresses between the two bridge tables belonging to the same VLAN (same E-TREE instance). Unless two lookups are made, duplication of MAC addresses would be needed for both locally learned and remotely learned MAC addresses. Locally learned MAC addresses from Leaf ACs need to be duplicated onto Root bridge table and locally learned MAC addresses from Root ACs need to be duplicated onto Leaf bridge table. Remotely learned MAC addresses from Root ACs need to be copied onto both Root and Leaf bridge tables. Because of potential inefficiencies associated with data-plane implementation of additional MAC lookup or duplication of MAC entries, option (A) is not believed to be implementable without dataplane performance inefficiencies in some platforms and thus this draft introduces the coloring option (B) as detailed in section 3.1.

For this scenario, if for a given EVI, the vast majority of PEs will have both Leaf and Root sites attached, even though they may start as Root-only or Leaf-only PEs, then a single RT per EVI MAY be used in order to alleviate the configuration overhead associated with using two RTs per EVI at the expense of having unwanted MAC addresses on the Leaf-only PEs.

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 single 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 identified is on a per MAC address. This scenario is considered in this draft for EVPN service with only known unicast traffic because the DF filtering per [RFC7432] would not be compatible with the required egress filtering - i.e., BUM traffic is not supported in this scenario and it is dropped by the ingress PE.

```
                     +---------+            +---------+
                     |   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

[RFC7432] 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, [RFC7432] 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/IP Advertisement route, then it notifies 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.

In situation where MAC moves are allowed among Leaf and Root sites (e.g., non-static MAC), PEs can receive multiple MAC/IP advertisements routes for the same MAC address with different Leaf/Root indications (and possibly different ESIs for multi-homing scenarios). In such situations, MAC mobility procedures take

precedence to first identify the location of the MAC before associating that MAC with a Root or a Leaf site.

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

This specification does not provide support for filtering BUM traffic on the ingress PE because it is not possible to perform filtering of BUM traffic on the ingress PE, as is the case with known unicast described above, due to the multi-destination nature of BUM 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 [RFC7432] (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].

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 as well as known-unicast traffic.

There are four scenarios to consider as follows. In all these scenarios, the ingress 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 label 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 AC identifier for that segment

(e.g., Ethernet Tag of the frame for 802.1Q frames). Other mechanisms for identifying root or leaf (e.g., on a per MAC address basis) 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 [RFC7432]. 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 label 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 leafs), the same AC (e.g., 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 or broadcast 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 [RFC7432] for adding and/or processing an ESI MPLS label.

3.3 E-TREE Traffic Flows for EVPN

Per [RFC7387], 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(s) 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 [RFC7432].

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 route is used to support BUM traffic.

The fields of this 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 advertises 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. Just like the EVPN case, the new E-TREE Extended Community defined in section [5.1] is 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 advertised: one B-MAC address is per ES as specified in [RFC7623] and implicitly denoting Root, and the other B-MAC address is per PE and 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 where "Leaf-indication" flag is set. In such multi-homing scenarios where and Ethernet Segment has

both Root and Leaf ACs, it is assumed that While different ACs (VLANs) on the same ES could have different root/leaf designation (some being roots and some being leafs), 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 L2 and not IRB. IRB use case is outside the scope of this document.

The ingress PE uses the right B-MAC source address depending on whether the Ethernet frame originated from the Root or Leaf AC 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.

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. The B-MAC address used for indicating Leaf sites can be the same for both single-homed and multi-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 (which will be used as a source B-MAC for BUM traffic), it updates its egress filtering (based on the source B-MAC 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 its B-MAC list used for egress filtering - i.e., to block traffic from that B-MAC address.

- 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 [RFC7623] and does not require any additional flags in the data-plane.

Just as in section 3.2, 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 as well as known-unicast traffic.

4.3 E-Tree without MAC Learning

In scenarios where the traffic of interest is only Multicast and/or broadcast, the PEs implementing an E-Tree service do not need to do any MAC learning. In such scenarios the filtering must be performed on egress PEs. For PBB-EVPN, the handling of such traffic is per section 4.2 without C-MAC learning part of it at both ingress and egress PEs.

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 0x05. 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. There is no need to send ESI Label Extended Community when sending Ethernet A-D per ES route with an ESI of zero. 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=0x05 | 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 (for known unicast traffic), 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) for BUM traffic, 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

[RFC6514] 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.

Using the Composite flag for Tunnel Types 0x00 'no tunnel information present' and 0x06 'Ingress Replication' is invalid, and should be treated as an invalid tunnel type on reception.

6 Acknowledgement

We would like to thank Dennis Cai, Antoni Przygienda, and Jeffrey Zhang for their valuable comments. The authors would also like to thank Thomas Morin for shepherding this document and providing valuable comments.

7 Security Considerations

Since this draft uses the EVPN constructs of [RFC7432] and [RFC7623], the same security considerations in these drafts are also applicable here. Furthermore, this draft provides additional security check by allowing sites (or ACs) of an EVPN instance to be designated as "Root" or "Leaf" and preventing any traffic exchange among "Leaf" sites of that VPN through ingress filtering for known unicast traffic and egress filtering for BUM traffic.

8 IANA Considerations

IANA has allocated value 5 in the "EVPN Extended Community Sub-Types" registry defined in [RFC7153] as follow:

SUB-TYPE VALUE	NAME	Reference
0x05	E-TREE Extended Community	This document

8.1 Considerations for PMSI Tunnel Types

The "P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types" registry in the "Border Gateway Protocol (BGP) Parameters" registry needs to be updated to reflect the use of the most significant bit to advertise the use of "composite tunnels" (section 5.2).

For this purpose, this document updates RFC7385.

The registry is to be updated, by removing the entries for 0xFB-0xFE and 0x0F, and replacing them by:

- 0x7B-0x7E Reserved for Experimental Use [this document]
- 0x7F Reserved [this document]
- 0x80-0xFF Not Allocatable, corresponds to Composite tunnel types [this document]

The allocation policy for values 0x00 to 0x7A is IETF Review [RFC5226]. The range for experimental use is now 0x7B-0x7E, and value in this range are not to be assigned. The status of 0x7F may only be changed through Standards Action [RFC5226].

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.

[RFC7623] Sajassi et al., "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", September, 2015.

[RFC7385] Andersson et al., "IANA Registry for P-Multicast Service Interface (PMSI) Tunnel Type Code Points", October, 2014.

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

[RFC6514] Aggarwal et al., "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", February, 2012.

9.2 Informative References

[RFC7387] Key et al., "A Framework for E-Tree Service over MPLS

Network", October 2014.

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

Contributors

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

Wim Henderickx
Nokia

Aldrin Isaac
Wen Lin
Juniper

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

John Drake
Juniper
Email: jdrake@juniper.net

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

Sami Boutros
VMware
Email: sboutros@vmware.com

Jorge Rabadan
Nokia
Email: jorge.rabadan@nokia.com