Internet Engineering Task Force (IETF)
Request for Comments: 8317
Updates: 7385
Category: Standards Track
ISSN: 2070-1721

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

January 2018

Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)

Abstract

The MEF 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 described in RFC 7387, "A Framework for Ethernet-Tree (E-Tree) Service over a Multiprotocol Label Switching (MPLS) Network". This document discusses how those functional requirements can be met with a solution based on RFC 7432, "BGP MPLS Based Ethernet VPN (EVPN)", with some extensions and a description of how such a solution can offer a more efficient implementation of these functions than that of RFC 7796, "Ethernet-Tree (E-Tree) Support in Virtual Private LAN Service (VPLS)". This document makes use of the most significant bit of the Tunnel Type field (in the P-Multicast Service Interface (PMSI) Tunnel attribute) governed by the IANA registry created by RFC 7385; hence, it updates RFC 7385 accordingly.

Status of This Memo

This is an Internet Standards Track document.

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

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc8317.

Table of Contents

1. Introduction

The MEF Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet-Tree (E-Tree) [MEF6.1]. In an E-Tree service, a customer site that is typically represented by an Attachment Circuit (AC) (e.g., an 802.1Q VLAN tag [IEEE.802.1Q]), is labeled as either a Root or a Leaf site. A customer site may also be represented by a Media Access Control (MAC) address along with a VLAN tag. Root sites can communicate with all other customer sites (both Root and Leaf sites). However, Leaf sites can communicate with Root sites but not with other Leaf sites. In this document, unless explicitly mentioned otherwise, a site is always represented by an AC.

[RFC7387] describes a solution framework for supporting E-Tree service in MPLS networks. This document identifies the functional components of an overall solution to emulate E-Tree services in MPLS networks and supplements the multipoint-to-multipoint Ethernet LAN (E-LAN) services specified in [RFC7432] and [RFC7623].

[RFC7432] defines EVPN, a solution for multipoint Layer 2 Virtual Private Network (L2VPN) services with advanced multihoming capabilities that uses BGP for distributing customer/client MAC address reachability information over the MPLS/IP network. [RFC7623] combines the functionality of EVPN with [IEEE.802.1ah] Provider Backbone Bridging (PBB) for MAC address scalability.

This document discusses how the functional requirements for E-Tree service can be met with a solution based on EVPN [RFC7432] and PBB-EVPN [RFC7623] with some extensions to their procedures and BGP attributes. Such a solution based on PBB-EVPN or EVPN can offer a more efficient implementation of these functions than that of [RFC7796], "Ethernet-Tree (E-Tree) Support in Virtual Private LAN Service (VPLS)". This efficiency is achieved by performing filtering of unicast traffic at the ingress Provider Edge (PE) nodes as opposed to egress filtering where the traffic is sent through the network and gets filtered and discarded at the egress PE nodes. The details of this ingress filtering are described in Section 4.1. Since this document specifies a solution based on [RFC7432], the knowledge of that document is a prerequisite. This document makes use of the most significant bit of the Tunnel Type field (in the PMSI Tunnel attribute) governed by the IANA registry created by [RFC7385]; hence, it updates [RFC7385] accordingly. Section 3 discusses E-Tree scenarios, Sections 4 and 5 describe E-Tree solutions for EVPN and PBB-EVPN (respectively), and Section 6 covers BGP encoding for E-Tree solutions.

2. Terminology

2.1. Specification of Requirements

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

2.2. Terms and Abbreviations

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.802.1ah].

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

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

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

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

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".

ESI: An Ethernet Segment Identifier is a unique non-zero identifier that identifies an ES.

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

P2MP: Point-to-Multipoint.

PE: Provider Edge device.

3. E-Tree Scenarios

This document categorizes E-Tree scenarios into the following three categories, depending on the nature of the Root/Leaf site association:

Scenario 1: either Leaf or Root site(s) per PE;

Scenario 2: either Leaf or Root site(s) per Attachment Circuit (AC); or,

Scenario 3: either Leaf or Root site(s) per MAC address.

3.1. Scenario 1: Leaf or Root Site(s) per PE

In this scenario, a PE may receive traffic from either Root ACs or Leaf ACs for a given MAC-VRF/bridge table, but not both. In other words, a given EVPN Instance (EVI) on a Provider Edge (PE) device is either associated with Root(s) or Leaf(s). The PE may have both Root and Leaf ACs, albeit for different EVIs.

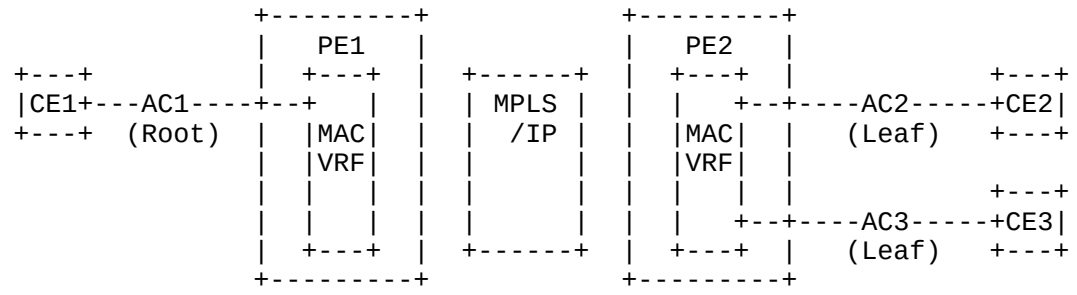


Figure 1: Scenario 1

In this scenario, tailored BGP Route Target (RT) import/export policies among the PEs belonging to the same EVI can be used to prevent communication among Leaf PEs. To prevent communication among Leaf ACs connected to the same PE and belonging to the same EVI, split-horizon filtering is used to block traffic from one Leaf AC to another Leaf AC on a MAC-VRF for 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 a topology constraint in EVPN, two BGP RTs are used for every EVI: one RT is associated with the Root sites (Root ACs) and the other is associated with the Leaf sites (Leaf ACs). On a per-EVI basis, every PE exports the single RT associated with its type of site(s). Furthermore, a PE with a Root site(s) imports both Root and Leaf RTs, whereas a PE with a Leaf site(s) only imports the Root RT.

For this scenario, if it is desired to use only a single RT per EVI (just like E-LAN services in [RFC7432]), then approach B in Scenario 2 (described below) needs to be used.

3.2. Scenario 2: Leaf or Root Site(s) per AC

In this scenario, a PE can receive traffic from both Root ACs and Leaf ACs for a given EVI. In other words, a given EVI on a PE can be associated with both Root(s) and Leaf(s).

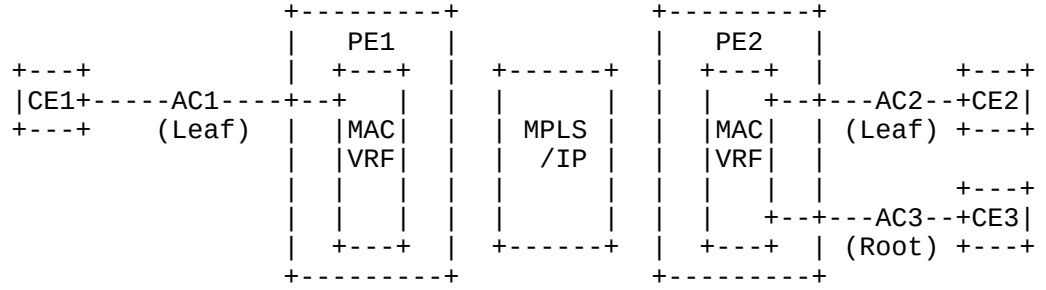


Figure 2: Scenario 2

In this scenario, (as in Scenario 1 Section 3.1), two RTs (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; thus, in order to apply the proper ingress filtering, 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 recognize the association of a destination MAC address to a Leaf or Root AC and, thus, support ingress filtering on the ingress PE with both Leaf and Root ACs, MAC addresses need to be colored with a Root or Leaf-Indication before advertising to other PEs. There are two approaches for such coloring:

(A) to always use two RTs (one to designate Leaf RT and another for Root RT), or

(B) to allow for a single RT to be used per EVI, just like [RFC7432], and, thus, color MAC addresses via a "color" flag in a new extended community as detailed in Section 6.1.

Approach A would require the same data-plane enhancements as approach B if MAC-VRF and bridge tables used per VLAN are to remain consistent with Section 6 of [RFC7432]. In order to avoid data-plane enhancements for approach A, multiple bridge tables per VLAN may be considered; however, this has major drawbacks (as described in Appendix A); thus, it is not recommended.

Given that both approaches A and B would require the same data-plane enhancements, approach B is chosen here in order to allow for RT usage consistent with baseline EVPN [RFC7432] and for better generality. It should be noted that if one wants to use RT constraints in order to avoid MAC advertisements associated with a Leaf AC to PEs with only Leaf ACs, then two RTs (one for Root and another for Leaf) can still be used with approach B; however, in such applications, Leaf/Root RTs will be used to constrain MAC advertisements and are not used to color the MAC routes for ingress filtering (i.e., in approach B, the coloring is always done via the new extended community).

If, for a given EVI, a significant number of PEs 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 should be used. The reason for such a recommendation is to alleviate the configuration overhead associated with using two RTs per EVI at the expense of having some unwanted MAC addresses on the Leaf-only PEs.

3.3. Scenario 3: Leaf or Root Site(s) per MAC Address

In this scenario, a customer Root or Leaf site is represented by a MAC address on an AC and a PE may receive traffic from both Root and Leaf sites on that AC for an EVI. This scenario is not covered in either [RFC7387] or [MEF6.1]; however, it is covered in this document for the sake of completeness. In this scenario, since an AC carries traffic from both Root and Leaf sites, the granularity at which Root or Leaf sites are identified is on a per-MAC-address basis. This scenario is considered in this document for EVPN service with only known unicast traffic because the Designated Forwarder (DF) filtering per [RFC7432] would not be compatible with the required egress filtering; that is, Broadcast, Unknown Unicast, and Multicast (BUM) traffic is not supported in this scenario; it is dropped by the ingress PE.

For this scenario, the approach B in Scenario 2 is used in order to allow for single RT usage by service providers.

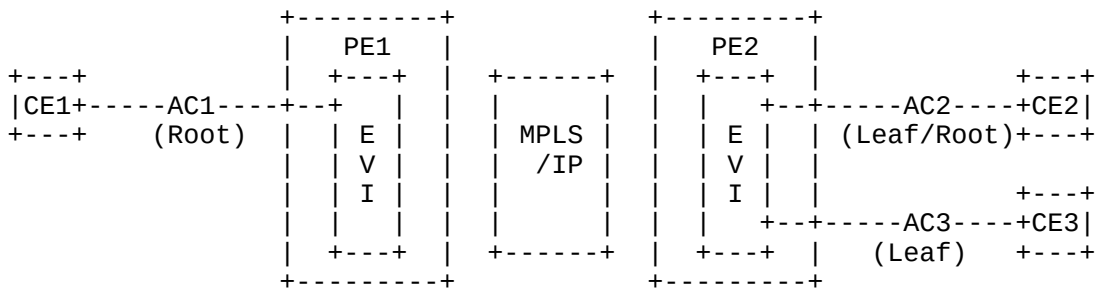


Figure 3: Scenario 3

In conclusion, the approach B in scenario 2 is the recommended approach across all the above three scenarios, and the corresponding solution is detailed in the following sections.

4. Operation for EVPN

[RFC7432] defines the notion of the Ethernet Segment Identifier (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 it will be seen shortly for BUM traffic. For known unicast traffic, additional extensions to [RFC7432] are needed (i.e., a new BGP extended community for Leaf-Indication described in Section 6.1) in order to enable ingress filtering as described in detail in the following sections.

4.1. Known Unicast Traffic

In EVPN, MAC learning is performed in the control plane via advertisement of BGP routes. Because of this, the filtering needed by an 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 the MPLS/IP core to be filtered at the egress PE, as is done in traditional E-Tree solutions (i.e., E-Tree for VPLS [RFC7796]).

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. This is done by advertising a Leaf-Indication flag (via an extended community) along with each of its MAC/IP Advertisement routes learned from a Leaf site. The lack of such a flag indicates that the MAC address is associated with a Root site. This scheme applies to all scenarios described in Section 3.

Tagging MAC addresses with a Leaf-Indication enables remote PEs to perform ingress filtering for known unicast traffic; that is, on the ingress PE, the MAC destination address lookup yields (in addition to the forwarding adjacency) a flag that indicates whether or not the target MAC is associated with a Leaf site. 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 a situation where MAC moves are allowed among Leaf and Root sites (e.g., non-static MAC), PEs can receive multiple MAC/IP Advertisement routes for the same MAC address with different Root or Leaf-Indications (and possibly different ESIs for multihoming scenarios). In such situations, MAC mobility procedures (see Section 15 of [RFC7432]) 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 (see Section 6.1). This new extended community MUST be advertised with MAC/IP Advertisement routes learned from a Leaf site. Besides MAC/IP Advertisement routes, no other EVPN routes are required to carry this new extended community for the purpose of known unicast traffic.

4.2. BUM Traffic

This specification does not provide support for filtering Broadcast, Unknown Unicast, and Multicast (BUM) traffic on the ingress PE; due to the multidestination nature of BUM traffic, it is not possible to perform filtering of the same on the ingress PE. 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 of when they originated from a Leaf AC (i.e., to be tagged with a Leaf label as specified in Section 6.1). This Leaf label allows for disposition PE (e.g., egress PE) to perform the necessary egress filtering function in a data plane similar to the ESI label in [RFC7432]. The allocation of the Leaf label is on a per-PE basis (e.g., independent of ESI and EVI) as described in the following sections.

The Leaf label can be upstream assigned for Point-to-Multipoint (P2MP) Label Switched Path (LSP) or downstream assigned for Ingress Replication tunnels. The main difference between a downstream- and upstream-assigned Leaf label is that, in the case of downstream-assigned Leaf labels, not all egress PE devices need to receive the label in MPLS-encapsulated BUM packets, just like the ESI label for Ingress Replication procedures defined in [RFC7432].

On the ingress PE, the PE needs to place all its Leaf ACs for a given bridge domain in a single split-horizon group in order to prevent intra-PE forwarding among its Leaf ACs. This intra-PE split-horizon filtering 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 [IEEE.802.1Q]). Other mechanisms for identifying Root or Leaf sites, such as the use of the source MAC address of the receiving frame, are optional. The scenarios below are described in context of a Root/Leaf AC, however, they can be extended to the Root/Leaf MAC address if needed.

4.2.1. BUM Traffic Originated from a Single-Homed Site on a Leaf AC

In this scenario, the ingress PE adds a Leaf label advertised using the E-Tree extended community (see Section 6.1), which indicates a Leaf site. This Leaf 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 ESs on that PE). This Leaf label is advertised to other PE devices using the E-Tree extended community (see Section 6.1) along with an Ethernet Auto-Discovery per ES (EAD-ES) route with an ESI of zero and a set of RTs corresponding to all EVIs on the PE where each EVI has at least one Leaf site. Multiple EAD-ES routes will need to be advertised if the number of RTs that need to be carried exceed the limit on a single route per [RFC7432]. The ESI for the EAD-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.

4.2.2. 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 labels and it operates per the procedures in [RFC7432].

4.2.3. BUM Traffic Originated from a Multihomed Site on a Leaf AC

In this scenario, it is assumed that while different ACs (VLANs) on the same ES could have a 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 (i.e., the service is EVPN L2 and not EVPN Integrated Routing and Bridging (IRB) [EVPN-INTEGRATED]). IRB use cases described in [EVPN-INTEGRATED] are outside the scope of this document.

In this scenario, if a multicast or broadcast packet is originated from a Leaf AC, then it only needs to carry a Leaf label as described in Section 4.2.1. 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 ES.

4.2.4. BUM Traffic Originated from a Multihomed Site on a Root AC

In this scenario, both the ingress and egress PE devices follow the procedure defined in [RFC7432] for adding and/or processing an ESI MPLS label; that is, existing procedures for BUM traffic in [RFC7432] are sufficient and there is no need to add a Leaf label.

4.3. E-Tree Traffic Flows for EVPN

Per [RFC7387], a generic E-Tree service supports all of the following traffic flows:

- known unicast traffic from Root to Roots & Leafs
- known unicast traffic from Leaf to Roots
- BUM traffic from Root to Roots & Leafs
- BUM traffic 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 only multicast and broadcast flows need to be supported, the L2VPN PEs can avoid performing any MAC learning function.

The following subsections will describe the operation of EVPN to support E-Tree service with and without MAC learning.

4.3.1. 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 ESs and advertises reachability in EVPN MAC/IP 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 ESs and advertise reachability in EVPN MAC/IP Advertisement routes. For scenarios where two different RTs are used per EVI (one to designate a Root site and another to designate a Leaf site), the MAC/IP Advertisement 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 Auto-Discovery per EVI (EAD-EVI) routes for aliasing (in the case of multihomed segments) and EAD-ES routes 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) (e.g., Root PEs) or Ingress Replication can be used (see Section 16 of [RFC7432]). 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, either Ingress Replication tunnels from each Leaf PE or a P2MP tree rooted at each Leaf PE can be used. The following two paragraphs describe when each of these tunneling schemes can be used and how to signal them.

When there are only a few Root PEs with small amount of multicast/broadcast traffic from Leaf PEs toward Root PEs, then Ingress Replication tunnels from Leaf PEs toward Root PEs should be sufficient. Therefore, if a Root PE needs to support a P2MP tunnel in the transmit direction from itself to Leaf PEs, and, at the same time, it wants to support Ingress Replication tunnels in the receive direction, the Root PE can signal it efficiently by using a new composite tunnel type defined in Section 6.2. This new composite tunnel type is advertised by the Root PE to simultaneously indicate a P2MP tunnel in the transmit direction and an Ingress Replication tunnel in the receive direction for the BUM traffic.

If the number of Root PEs is large, P2MP tunnels (e.g., Multipoint LDP (mLDP) or RSVP-TE) originated at the Leaf PEs may be used; thus, there will be no need to use the modified PMSI Tunnel attribute and the composite tunnel type values defined in Section 6.2.

4.3.2. 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 mainly multicast or broadcast. In this case, the PEs do not exchange EVPN MAC/IP Advertisement routes. Instead, the Inclusive Multicast Ethernet Tag route is used to support BUM traffic. In such scenarios, the small amount of unicast traffic (if any) is sent as part of 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 Root PEs are only a few and, thus, Ingress Replication is desired from Leaf PEs to these Root PEs, then the modified PMSI attribute and the composite tunnel type values defined in Section 6.2 should be used.

5. Operation for PBB-EVPN

In PBB-EVPN, the PE advertises a Root or Leaf-Indication along with each Backbone MAC (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 6.1 is advertised with each EVPN MAC/IP Advertisement route.

In the case where a multihomed ES 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 denotes Root, and the other B-MAC address is per PE and explicitly denotes 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 a "Leaf-indication" flag is set. In multihoming scenarios where an ES has both Root and Leaf ACs, it is assumed that while different ACs (VLANs) on the same ES could have a 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 (i.e., the service is L2 and not IRB). An 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 ES. 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 multihomed segments.

5.1. Known Unicast Traffic

For known unicast traffic, the PEs perform ingress filtering: on the ingress PE, the Customer/Client MAC (C-MAC) [RFC7623] destination address lookup yields, in addition to the target B-MAC address and forwarding adjacency, a flag that indicates whether the target B-MAC is associated with a Root or a Leaf site. The ingress PE also checks the status of the originating site; if both are Leafs, then the packet is not forwarded.

5.2. BUM Traffic

For BUM traffic, the PEs must perform egress filtering. When a PE receives an EVPN MAC/IP 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 EVPN MAC/IP Advertisement route indicates that the advertised B-MAC is a Leaf, and the local ES 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.

- If the EVPN MAC/IP Advertisement route indicates that the advertised B-MAC has changed its designation from a Leaf to a Root, and the local ES is a Leaf, then the source B-MAC address is removed from the B-MAC list corresponding to the local ES used for egress filtering (i.e., to unblock traffic from that B-MAC address).

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 the split-horizon filtering described in Section 6.2.1.3 of [RFC7623] and does not require any additional flags in the data plane.

Just as in Section 4.2, the PE places all Leaf ESs 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.

5.3. 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 5.2 without the need for C-MAC learning (in the data plane) in the I-component (C-bridge table) of PBB-EVPN PEs (at both ingress and egress PEs).

6. BGP Encoding

This document defines a new BGP extended community for EVPN.

6.1. E-Tree Extended Community

This extended community is a new transitive extended community [RFC4360] 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. It indicates that the frame is originated from a Leaf site.

The E-Tree extended community is encoded as an 8-octet value as follows:

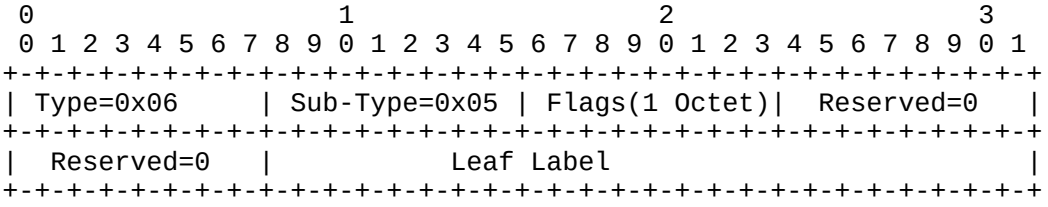
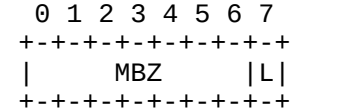


Figure 4: E-Tree Extended Community

The Flags field has the following format:



This document defines the following flags:

+ Leaf-Indication (L)

A value of one indicates a Leaf AC/site. The rest of the flag bits are reserved and should be set to zero.

When this extended community is advertised along with the MAC/IP Advertisement route (for known unicast traffic) per Section 4.1, the Leaf-Indication flag MUST be set to one and the Leaf label SHOULD be set to zero. The receiving PE MUST ignore Leaf label and only process the Leaf-Indication flag. A value of zero for the Leaf-Indication flag is invalid when sent along with a MAC/IP Advertisement route, and an error should be logged.

When this extended community is advertised along with the EAD-ES route (with an ESI of zero) for BUM traffic to enable egress filtering on disposition PEs per Sections 4.2.1 and 4.2.3, the Leaf label MUST be set to a valid MPLS label (i.e., a non-reserved, assigned MPLS label [RFC3032]) and the Leaf-Indication flag SHOULD be set to zero. The value of the 20-bit MPLS label is encoded in the high-order 20 bits of the Leaf label field. The receiving PE MUST ignore the Leaf-Indication flag. A non-valid MPLS label, when sent along with the EAD-ES route, should be ignored and logged as an error.

The reserved bits SHOULD be set to zero by the transmitter and MUST be ignored by the receiver.

6.2. PMSI Tunnel Attribute

[RFC6514] defines the PMSI Tunnel attribute, which is an optional transitive attribute with the following format:

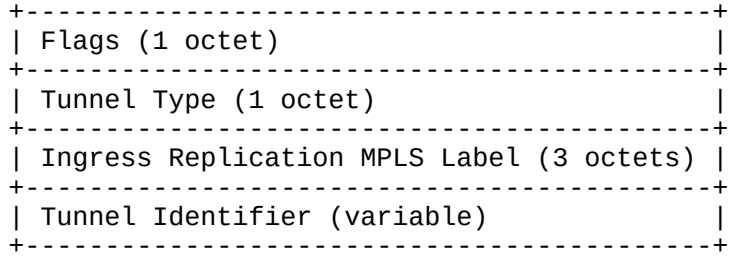


Table 1: PMSI Tunnel Attribute

This document defines a new composite tunnel type by introducing a new 'composite tunnel' bit in the Tunnel Type field and adding an MPLS label to the Tunnel Identifier field of the PMSI Tunnel attribute, as detailed below. All other fields remain as defined in [RFC6514]. Composite tunnel type is advertised by the Root PE to simultaneously indicate a non-Ingress-Replication tunnel (e.g., P2MP tunnel) in the transmit direction and an Ingress Replication tunnel in the receive direction for the BUM traffic.

When receiver Ingress Replication labels are needed, the high-order bit of the Tunnel Type field (composite tunnel bit) is set while the remaining low-order seven bits indicate the Tunnel Type as before (for the existing Tunnel Types). When this composite tunnel bit is set, the "tunnel identifier" field begins 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 treat the Tunnel Type as an undefined Tunnel Type and treat the PMSI Tunnel attribute as a malformed attribute [RFC6514]. That is why the composite tunnel bit is allocated in the Tunnel Type field rather than the Flags field. 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 label in the Tunnel Identifier field when sending its BUM traffic.

Using the composite tunnel bit for Tunnel Types 0x00 'no tunnel information present' and 0x06 'Ingress Replication' is invalid. A PE that receives a PMSI Tunnel attribute with such information considers it malformed, and it SHOULD treat this Update as though all the routes contained in this Update had been withdrawn per Section 6 of [RFC6514].

7. Security Considerations

Since this document uses the EVPN constructs of [RFC7432] and [RFC7623], the same security considerations in these documents are also applicable here. Furthermore, this document provides an additional security check by allowing sites (or ACs) of an EVPN instance to be designated as a "Root" or "Leaf" by the network operator / service provider and thus prevent any traffic exchange among "Leaf" sites of that VPN through ingress filtering for known unicast traffic and egress filtering for BUM traffic. Since (by default and for the purpose of backward compatibility) an AC that doesn't have a Leaf designation is considered a Root AC, in order to avoid any traffic exchange among Leaf ACs, the operator SHOULD configure the AC with a proper role (Leaf or Root) before activating the AC.

8. IANA Considerations

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

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

This document creates a one-octet registry called "E-Tree Flags". New registrations will be made through the "RFC Required" procedure defined in [RFC8126]. Initial registrations are as follows:

Bit	Name	Reference
0-6	Unassigned	
7	Leaf-Indication	This document

8.1. Considerations for PMSI Tunnel Types

The "P-Multicast Service Interface (PMSI) Tunnel Types" registry in the "Border Gateway Protocol (BGP) Parameters" registry has been updated to reflect the use of the most significant bit as the "composite tunnel" bit (see Section 6.2).

For this purpose, this document updates [RFC7385] by changing the previously unassigned values (i.e., 0x08 - 0xFA) as follows:

Value	Meaning	Reference
0x0C-0x7A	Unassigned	
0x7B-0x7E	Experimental	This Document
0x7F	Reserved	This Document
0x80-0xFA	Reserved for Composite Tunnel	This Document
0xFB-0xFE	Experimental	[RFC7385]
0xFF	Reserved	[RFC7385]

The allocation policy for values 0x08-0x7A is per IETF Review [RFC8126]. The range for "Experimental" has been expanded to include the previously assigned range of 0xFB-0xFE and the new range of 0x7B-0x7E. The values in these ranges are not to be assigned. The value 0x7F, which is the mirror image of (0xFF), is reserved in this document.

9. References

9.1. Normative References

[MEF6.1] MEF Forum, "Ethernet Services Definitions - Phase 2", MEF 6.1, April 2008, <https://mef.net/PDF_Documents/technical-specifications/MEF6-1.pdf>.

[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>.

[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>.

[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>.

[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>.

[RFC7385] Andersson, L. and G. Swallow, "IANA Registry for P-Multicast Service Interface (PMSI) Tunnel Type Code Points", RFC 7385, DOI 10.17487/RFC7385, October 2014, <https://www.rfc-editor.org/info/rfc7385>.

[RFC7387] Key, R., Ed., Yong, L., Ed., Delord, S., Jounay, F., and L. Jin, "A Framework for Ethernet Tree (E-Tree) Service over a Multiprotocol Label Switching (MPLS) Network", RFC 7387, DOI 10.17487/RFC7387, October 2014, <https://www.rfc-editor.org/info/rfc7387>.

[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>.

[RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <https://www.rfc-editor.org/info/rfc7623>.

[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>.

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

9.2. Informative References

[EVPN-INTEGRATED] Sajassi, A., Salam, S., Thoria, S., Drake, J., Rabadan, J., and L. Yong, "Integrated Routing and Bridging in EVPN", Work in Progress, draft-ietf-bess-evpn-inter-subnet-forwarding-03, February 2017.

[IEEE.802.1ah] IEEE, "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", Clauses 25 and 26, IEEE Std 802.1Q, DOI 10.1109/IEEESTD.2011.6009146.

[IEEE.802.1Q] IEEE, "IEEE Standard for Local and metropolitan area networks - Bridges and Bridged Networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q, DOI 10.1109/IEEESTD.2011.6009146.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <https://www.rfc-editor.org/info/rfc3032>.

[RFC7796] Jiang, Y., Ed., Yong, L., and M. Paul, "Ethernet-Tree (E-Tree) Support in Virtual Private LAN Service (VPLS)", RFC 7796, DOI 10.17487/RFC7796, March 2016, <https://www.rfc-editor.org/info/rfc7796>.

Appendix A. Multiple Bridge Tables per E-Tree Service Instance

When two MAC-VRFs (two bridge tables per VLAN) are used for an E-Tree service (one for Root ACs and another for Leaf ACs) on a given PE, then the following complications in a data-plane path can result.

Maintaining two MAC-VRFs (two bridge tables) per VLAN (when both Leaf and Root ACs exists for that VLAN) would require either that two lookups be performed per MAC address in each direction in case of a miss or that the duplication of many MAC addresses between the two bridge tables belonging to the same VLAN (same E-Tree instance) be made. 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 a Root bridge table, and locally learned MAC addresses from Root ACs need to be duplicated onto a 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, this option is not believed to be implementable without data-plane performance inefficiencies in some platforms; thus, this document introduces the coloring as described in Section 3.2 and detailed in Section 4.1.

Acknowledgements

We would like to thank Eric Rosen, Jeffrey Zhang, Wen Lin, Aldrin Issac, Wim Henderickx, Dennis Cai, and Antoni Przygienda for their valuable comments and contributions. The authors would also like to thank Thomas Morin for shepherding this document and providing valuable comments.

Authors' Addresses

Ali Sajassi (editor)
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