Internet Working Group
Internet Draft
Category: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Cisco

Florin Balus
Wim Henderickx
Alcatel-Lucent

Nabil Bitar
Verizon

Clarence Filsfils
Dennis Cai
Cisco

Aldrin Issac
Bloomberg

Lizhong Jin
ZTE

Expires: August 27, 2012   February 27, 2012

PBB E-VPN
draft-ietf-l2vpn-pbb-evpn-00.txt

Status of this Memo

Copyright Notice

Abstract

This document discusses how Ethernet Provider Backbone Bridging [802.1ah] can be combined with E-VPN in order to reduce the number of BGP MAC advertisement routes by aggregating Customer/Client MAC (C-MAC) addresses via Provider Backbone MAC address (B-MAC), provide client MAC address mobility using C-MAC aggregation and B-MAC sub-netting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes. The combined solution is referred to as PBB-EVPN.

Conventions

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

Table of Contents

1. Introduction

[E-VPN] introduces a solution for multipoint L2VPN services with advanced multi-homing capabilities using BGP for distributing customer/clinent MAC address reach-ability information over the core MPLS/IP network. [802.1ah] defines an architecture for Ethernet Provider Backbone Bridging (PBB), where MAC tunneling is employed to improve service instance and MAC address scalability in Ethernet networks and in VPLS networks [PBB-VPLS].

In this document, we discuss how PBB can be combined with E-VPN in order to reduce the number of BGP MAC advertisement routes by aggregating Customer/Client MAC (C-MAC) addresses via Provider Backbone MAC address (B-MAC), provide client MAC address mobility using C-MAC aggregation and B-MAC sub-netting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes. The combined solution is referred to as PBB-EVPN.

2. Contributors

In addition to the authors listed above, the following individuals also contributed to this document.

Keyur Patel
Cisco

3. Terminology

BEB: Backbone Edge Bridge
B-MAC: Backbone MAC Address
CE: Customer Edge
C-MAC: Customer/Client MAC Address
DHD: Dual-homed Device
DHN: Dual-homed Network
LACP: Link Aggregation Control Protocol
LSM: Label Switched Multicast
MDT: Multicast Delivery Tree
MES: MPLS Edge Switch
MP2MP: Multipoint to Multipoint
P2MP: Point to Multipoint
P2P: Point to Point
PoA: Point of Attachment
PW: Pseudowire
E-VPN: Ethernet VPN

4. Requirements

The requirements for PBB-EVPN include all the requirements for E-VPN that were described in [EVPN-REQ], in addition to the following:

4.1. MAC Advertisement Route Scalability

In typical operation, an [E-VPN] MES sends a BGP MAC Advertisement Route per customer/client MAC (C-MAC) address. In certain applications, this poses scalability challenges, as is the case in virtualized data center environments where the number of virtual machines (VMs), and hence the number of C-MAC addresses, can be in the millions. In such scenarios, it is required to reduce the number of BGP MAC Advertisement routes by relying on a MAC 'summarization' scheme, as is provided by PBB. Note that the MAC sub-netting capability already built into E-VPN is not sufficient in those environments, as will be discussed next.

4.2. C-MAC Mobility with MAC Sub-netting

Certain applications, such as virtual machine mobility, require support for fast C-MAC address mobility. For these applications, it is not possible to use MAC address sub-netting in E-VPN, i.e. advertise reach-ability to a MAC address prefix. Rather, the exact virtual machine MAC address needs to be transmitted in BGP MAC Advertisement route. Otherwise, traffic would be forwarded to the wrong segment when a virtual machine moves from one Ethernet segment to another. This hinders the scalability benefits of sub-netting.

It is required to support C-MAC address mobility, while retaining the scalability benefits of MAC sub-netting. This can be achieved by leveraging PBB technology, which defines a Backbone MAC (B-MAC) address space that is independent of the C-MAC address space, and aggregate C-MAC addresses via a B-MAC address and then apply sub-netting to B-MAC addresses.

4.3. C-MAC Address Learning and Confinement

In E-VPN, all the MES nodes participating in the same E-VPN instance are exposed to all the C-MAC addresses learnt by any one of these MES nodes because a C-MAC learned by one of the MES nodes is advertise in BGP to other MES nodes in that E-VPN instance. This is the case even if some of the MES nodes for that E-VPN instance are not involved in forwarding traffic to, or from, these C-MAC addresses. Even if an implementation does not install hardware forwarding entries for C-MAC addresses that are not part of active traffic flows on that MES, the device memory is still consumed by keeping record of the C-MAC addresses in the routing table (RIB). In network applications with millions of C-MAC addresses, this introduces a non-trivial waste of MES resources. As such, it is required to confine the scope of visibility of C-MAC addresses only to those MES nodes that are actively involved in forwarding traffic to, or from, these addresses.

4.4. Interworking with TRILL and 802.1aq Access Networks with C-MAC Address Transparency

[TRILL] and [802.1aq] define next generation Ethernet bridging technologies that offer optimal forwarding using IS-IS control plane, and C-MAC address transparency via Ethernet tunneling technologies. When access networks based on TRILL or 802.1aq are interconnected over an MPLS/IP network, it is required to guarantee C-MAC address transparency on the hand-off point and the edge (i.e. MES) of the MPLS network. As such, solutions that require termination of the access data-plane encapsulation (i.e. TRILL or

802.1aq) at the hand-off to the MPLS network do not meet this transparency requirement, and expose the MPLS edge devices to the MAC address scalability problem.

PBB-EVPN supports seamless interconnect with these next generation Ethernet solutions while guaranteeing C-MAC address transparency on the MES nodes.

4.5. Per Site Policy Support

In many applications, it is required to be able to enforce connectivity policy rules at the granularity of a site (or segment). This includes the ability to control which MES nodes in the network can forward traffic to, or from, a given site. PBB-EVPN is capable of providing this granularity of policy control. In the case where per C-MAC address granularity is required, the EVI can always continue to operate in E-VPN mode.

4.6. Avoiding C-MAC Address Flushing

It is required to avoid C-MAC address flushing upon link, port or node failure for multi-homed devices and networks. This is in order to speed up re-convergence upon failure.

5. Solution Overview

The solution involves incorporating IEEE 802.1ah Backbone Edge Bridge (BEB) functionality on the E-VPN MES nodes similar to PBB-VPLS PEs (PBB-VPLS) where BEB functionality is incorporated in PE nodes. The MES devices would then receive 802.1Q Ethernet frames from their attachment circuits, encapsulate them in the PBB header and forward the frames over the IP/MPLS core. On the egress E-VPN MES, the PBB header is removed following the MPLS disposition, and the original 802.1Q Ethernet frame is delivered to the customer equipment.

```
                 BEB    +--------------+  BEB
                 ||     |              |  ||
                 \/     |              |  \/
    +-----+ AC1 +-----+ |              | +-----+   +-----+
    | CE1|-----|    |   |              | |    |---| CE2|
    +-----+\    |MES1|  |   IP/MPLS    | |MES3|   +-----+
            \   +-----+ |   Network    | +-----+
             \          |              |
           AC2\ +-----+ |              |
               \|    |  |              |
               |MES2| |              |
               +----+ |              |
                 /\   +--------------+
                 ||
                 BEB
     <-802.1Q-> <------PBB over MPLS------> <-802.1Q->
```

Figure 1: PBB-EVPN Network

The MES nodes perform the following functions:

- Learn customer/client MAC addresses (C-MACs) over the attachment circuits in the data-plane, per normal bridge operation.

- Learn remote C-MAC to B-MAC bindings in the data-plane from traffic ingress from the core per [802.1ah] bridging operation.

- Advertise local B-MAC address reach-ability information in BGP to all other MES nodes in the same set of service instances. Note that every MES has a set of local B-MAC addresses that uniquely identify the device. More on the MES addressing in section 5.

- Build a forwarding table from remote BGP advertisements received associating remote B-MAC addresses with remote MES IP addresses and the associated MPLS label(s).

6. BGP Encoding

PBB-EVPN leverages the same BGP Routes and Attributes defined in [E-VPN], adapted as follows:

6.1. BGP MAC Advertisement Route

The E-VPN MAC Advertisement Route is used to distribute B-MAC addresses of the MES nodes instead of the C-MAC addresses of end-stations/hosts. This is because the C-MAC addresses are learnt in the data-plane for traffic arriving from the core. The MAC Advertisement Route is encoded as follows:

- The RD is set to a Type 1 RD [RFC4364]. The value field encodes the IP address of the MES (typically, the loopback address) followed by 0. The reason for such encoding is that the RD cannot be that of a single EVI since the same B-MAC address can span across multiple EVIs.

- The MAC address field contains the B-MAC address.
- The Ethernet Tag field is set to 0.

The route is tagged with the set of RTs corresponding to all EVIs associated with the B-MAC address.

All other fields are set as defined in [E-VPN].

6.2. Ethernet Auto-Discovery Route

This route and any of its associated modes is not needed in PBB-EVPN.

6.3. Per VPN Route Targets

PBB-EVPN uses the same set of route targets defined in [E-VPN]. More specifically, the RT associated with a VPN is set to the value of the I-SID associated with the service instance. This eliminates the need for manually configuring the VPN-RT.

6.4. MAC Mobility Extended Community

This extended community is a new transitive extended community. It may be advertised along with MAC Advertisement routes. When used in PBB-EVPN, it indicates that the C-MAC forwarding tables for the I-SIDs associated with the RTs tagging the MAC Advertisement routes must be flushed. This extended community is encoded in 8-bytes as follows:

- Type (1 byte) = Pending IANA assignment.
- Sub-Type (1 byte) = Pending IANA assignment.
- Reserved (2 bytes)
- Counter (4 bytes)

Note that all other BGP messages and/or attributes are used as defined in [E-VPN].

7. Operation

This section discusses the operation of PBB-EVPN, specifically in areas where it differs from [E-VPN].

7.1. MAC Address Distribution over Core

In PBB-EVPN, host MAC addresses (i.e. C-MAC addresses) need not be distributed in BGP. Rather, every MES independently learns the C-MAC addresses in the data-plane via normal bridging operation. Every MES has a set of one or more unicast B-MAC addresses associated with it, and those are the addresses distributed over the core in MAC Advertisement routes. Given that these B-MAC addresses are global within the provider's network, there's no need to advertise them on a per service instance basis.

7.2. Device Multi-homing

7.2.1. MES MAC Layer Addressing & Multi-homing

In [802.1ah] every BEB is uniquely identified by one or more B-MAC addresses. These addresses are usually locally administered by the

Service Provider. For PBB-EVPN, the choice of B-MAC address(es) for the MES nodes must be examined carefully as it has implications on the proper operation of multi-homing. In particular, for the scenario where a CE is multi-homed to a number of MES nodes with all-active redundancy and flow-based load-balancing, a given C-MAC address would be reachable via multiple MES nodes concurrently. Given that any given remote MES will bind the C-MAC address to a single B-MAC address, then the various MES nodes connected to the same CE must share the same B-MAC address. Otherwise, the MAC address table of the remote MES nodes will keep flip-flopping between the B-MAC addresses of the various MES devices. For example, consider the network of Figure 1, and assume that MES1 has B-MAC BM1 and MES2 has B-MAC BM2. Also, assume that both links from CE1 to the MES nodes are part of an all-active multi-chassis Ethernet link aggregation group. If BM1 is not equal to BM2, the consequence is that the MAC address table on MES3 will keep oscillating such that the C-MAC address CM of CE1 would flip-flop between BM1 or BM2, depending on the load-balancing decision on CE1 for traffic destined to the core.

Considering that there could be multiple sites (e.g. CEs) that are multi-homed to the same set of MES nodes, then it is required for all the MES devices in a Redundancy Group to have a unique B-MAC address per site. This way, it is possible to achieve fast convergence in the case where a link or port failure impacts the attachment circuit connecting a single site to a given MES.

```
                            +---------+
             +-------+ MES1 | IP/MPLS |
            /               |         |
         CE1                | Network |    MESr
        M1  \               |         |
             +-------+ MES2 |         |
             /-------+      |         |
            /               |         |
         CE2                |         |
       M2   \               |         |
             \              |         |
              +------+ MES3 +---------+
```

Figure 2: B-MAC Address Assignment

In the example network shown in Figure 2 above, two sites corresponding to CE1 and CE2 are dual-homed to MES1/MES2 and MES2/MES3, respectively. Assume that BM1 is the B-MAC used for the site corresponding to CE1. Similarly, BM2 is the B-MAC used for the site corresponding to CE2. On MES1, a single B-MAC address (BM1) is required for the site corresponding to CE1. On MES2, two B-MAC addresses (BM1 and BM2) are required, one per site. Whereas on MES3, a single B-MAC address (BM2) is required for the site corresponding to CE2. All three MES nodes would advertise their respective B-MAC addresses in BGP using the MAC Advertisement routes defined in [E-VPN]. The remote MES, MESr, would learn via BGP that BM1 is reachable via MES1 and MES2, whereas BM2 is reachable via both MES2 and MES3. Furthermore, MESr establishes via the normal bridge learning that C-MAC M1 is reachable via BM1, and C-MAC M2 is reachable via BM2. As a result, MESr can load-balance traffic destined to M1 between MES1 and MES2, as well as traffic destined to M2 between both MES2 and MES3. In the case of a failure that causes, for example, CE1 to be isolated from MES1, the latter can withdraw the route it has advertised for BM1. This way, MESr would update its path list for BM1, and will send all traffic destined to M1 over to MES2 only.

For single-homed sites, it is possible to assign a unique B-MAC address per site, or have all the single-homed sites connected to a given MES share a single B-MAC address. The advantage of the first model over the second model is the ability to avoid C-MAC destination address lookup on the disposition PE (even though source C-MAC learning is still required in the data-plane). Also, by assigning the B-MAC addresses from a contiguous range, it is possible to advertise a single B-MAC subnet for all single-homed sites, thereby rendering the number of MAC advertisement routes required at par with the second model.

In summary, every MES may use a unicast B-MAC address shared by all single-homed CEs or a unicast B-MAC address per single-homed CE, and in addition a unicast B-MAC address per dual-homed CE. In the latter case, the B-MAC address MUST be the same for all MES nodes in a Redundancy Group connected to the same CE.

7.2.1.1. Automating B-MAC Address Assignment

The MES B-MAC address used for single-homed sites can be automatically derived from the hardware (using for e.g. the backplane's address). However, the B-MAC address used for multi-homed sites must be coordinated among the RG members. To automate the assignment of this latter address, the MES can derive this B-MAC address from the MAC Address portion of the CE's LACP System Identifier by flipping the 'Locally Administered' bit of the CE's address. This guarantees the uniqueness of the B-MAC address within the network, and ensures that all MES nodes connected to the same multi-homed CE use the same value for the B-MAC address.

Note that with this automatic provisioning of the B-MAC address associated with mult-homed CEs, it is not possible to support the uncommon scenario where a CE has multiple bundles towards the MES nodes, and the service involves hair-pinning traffic from one bundle to another. This is because the split-horizon filtering relies on B-MAC addresses rather than Site-ID Labels (as will be described in the next section). The operator must explicitly configure the B-MAC address for this fairly uncommon service scenario.

Whenever a B-MAC address is provisioned on the MES, either manually or automatically (as an outcome of CE auto-discovery), the MES MUST transmit an MAC Advertisement Route for the B-MAC address with a downstream assigned MPLS label that uniquely identifies that address on the advertising MES. The route is tagged with the RTs of the associated EVIs as described above.

7.2.2. Split Horizon and Designated Forwarder Election

[E-VPN] relies on access split horizon, where the Ethernet Segment Label is used for egress filtering on the attachment circuit in order to prevent forwarding loops. In PBB-EVPN, the B-MAC source address can be used for the same purpose, as it uniquely identifies the originating site of a given frame. As such, Segment Labels are not used in PBB-EVPN, and the egress filtering is done based on the B-MAC source address. It is worth noting here that [802.1ah] defines this B-MAC address based filtering function as part of the I-Component options, hence no new functions are required to support split-horizon beyond what is already defined in [802.1ah].
Given that the Segment label is not used in PBB-EVPN, the MES sets the Label field in the Ethernet Segment Route to 0.

The Designated Forwarder election procedures remain unchanged from [E-VPN].

7.3. Network Multi-homing

When an Ethernet network is multi-homed to a set of MES nodes running PBB-EVPN, an all-active redundancy model can be supported with per service instance (i.e. I-SID) load-balancing. In this model, DF election is performed to ensure that a single MES node in the redundancy group is responsible for forwarding traffic associated with a given I-SID. This guarantees that no forwarding loops are created. Filtering based on DF state applies to both unicast and multicast traffic, and in both access-to-core as well as core-to-access directions (unlike the multi-homed device scenario where DF filtering is limited to multi-destination frames in the core-to-access direction).
Similar to the multi-homed device scenario, a unique B-MAC address is used on the MES per multi-homed network (Segment). This helps eliminate the need for C-MAC address flushing in all but one failure scenario (more details on this in the Failure Handling section below). The B-MAC address may be auto-provisioned by snooping on the

BPDUs of the multi-homed network: the B-MAC address is set to the root bridge ID of the CIST albeit with the 'Locally Administered' bit set.

7.3.1. B-MAC Address Advertisement

For every multi-homed network, the MES advertises two MAC Advertisement routes with different RDs and identical MAC addresses and ESIs. One of these routes will be tagged with a lower Local Pref attribute than the other. The route with the higher Local Pref will be tagged with the RTs corresponding to the I-SIDs for which the advertising MES is the DF. Whereas, the route with the lower Local Pref will be tagged with the RTs corresponding to the I-SIDs for which the advertising MES is the backup DF. Consider the example network of the figure below, where a multi-homed network (MHN1) is connected to two MES nodes (MES1 and MES2).

```
                            +---------+
             +-------+ MES1 | IP/MPLS |
    +------+         BM1    |         |
    |      |                | Network |    MESr
    | MHN1 |         BM1    |         |
    +------+ +-------+ MES2 |         |
                            +---------+
```

Figure 3: Multi-homed Network

Both MES nodes use the same B-MAC address (BM1) for the Ethernet Segment (ESI1) associated with MHN1. Assume, for instance, that MES1 is the DF for the even I-SIDs whereas MES2 is the DF for the odd I-SIDs. In this example, the routes advertised by MES1 and MES2 would be as follows:

MES1:

Route 1: RD11, BM1, ESI1, Local Pref = 120, RT2, RT4, RT6...
Route 2: RD12, BM1, ESI1, Local Pref = 80, RT1, RT3, RT5...

MES2:

Route 1: RD21, BM1, ESI1, Local Pref = 120, RT1, RT3, RT5...
Route 2: RD22, BM1, ESI1, Local Pref = 80, RT2, RT4, RT6

Upon receiving the above MAC Advertisement routes, the remote MES nodes (e.g. MESr) would install forwarding entries for BM1 towards MES1 for the even I-SIDs, and towards MES2 for the odd I-SIDs.

It is worth noting that the procedures of this section can also be used for a multi-homed device in order to support all-active redundancy with per I-SID load-balancing.

7.3.2. Failure Handling

In the case of an MES node failure, or when the MES is isolated from the multi-homed network due to a port or link failure, the affected

MES withdraws its MAC Advertisement routes for the associated B-MAC. This serves as a trigger for the remote MES nodes to adjust their forwarding entries to point to the backup DF. Because the same B-MAC address is used on both the DF and backup DF nodes, then there is no need to flush the C-MAC address table upon the occurrence of these failures.

In the case where the multi-homed network is partitioned, the MES nodes can detect this condition by snooping on the network's BPDUs. When a MES detects that the root bridge ID has changed, it must change the value of the B-MAC address associated with the Ethernet Segment. This is done by the MES withdrawing the previous MAC Advertisement route, and advertising a new route for the updated B-MAC. The MES, which detects the failure, must inform the remote MES nodes to flush their C-MAC address tables for the affected I-SIDs. This is required because when the multi-homed network is partitioned, certain C-MAC addresses will move from being associated with the old B-MAC address to the new B-MAC addresses. Other C-MAC addresses will have their reachability remaining intact. Given that the MES node has no means of identifying which C-MACs have moved and which have not, the entire C-MAC forwarding table for the affected I-SIDs must be flushed. The affected MES signals the need for the C-MAC flushing by sending the MAC Mobility Extended Community in the MP_UNREACH_NLRI attribute containing the E-VPN NLRI for the withdrawn MAC Advertisement route.

7.4. Frame Forwarding

The frame forwarding functions are divided in between the Bridge Module, which hosts the [802.1ah] Backbone Edge Bridge (BEB) functionality, and the MPLS Forwarder which handles the MPLS imposition/disposition. The details of frame forwarding for unicast and multi-destination frames are discussed next.

7.4.1. Unicast

Known unicast traffic received from the AC will be PBB-encapsulated by the MES using the B-MAC source address corresponding to the originating site. The unicast B-MAC destination address is determined based on a lookup of the C-MAC destination address (the binding of the two is done via transparent learning of reverse traffic). The resulting frame is then encapsulated with an LSP tunnel label and the MPLS label which uniquely identifies the B-MAC destination address on the egress MES. If per flow load-balancing over ECMPs in the MPLS core is required, then a flow label is added as the end of stack label.

For unknown unicast traffic, the MES forwards these frames over MPLS core. When these frames are to be forwarded, then the same set of options used for forwarding multicast/broadcast frames (as described in next section) are used.

7.4.2. Multicast/Broadcast

Multi-destination frames received from the AC will be PBB-encapsulated by the MES using the B-MAC source address corresponding to the originating site. The multicast B-MAC destination address is selected based on the value of the I-SID as defined in [802.1ah]. The resulting frame is then forwarded over the MPLS core using one out of the following two options:

Option 1: the MPLS Forwarder can perform ingress replication over a set of MP2P tunnel LSPs. The frame is encapsulated with a tunnel LSP label and the E-VPN ingress replication label advertised in the Inclusive Multicast Route.

Option 2: the MPLS Forwarder can use P2MP tunnel LSP per the procedures defined in [E-VPN]. This includes either the use of Inclusive or Aggregate Inclusive trees.

Note that the same procedures for advertising and handling the Inclusive Multicast Route defined in [E-VPN] apply here.

8. Minimizing ARP Broadcast

The MES nodes implement an ARP-proxy function in order to minimize the volume of ARP traffic that is broadcasted over the MPLS network. This is achieved by having each MES node snoop on ARP request and response messages received over the access interfaces or the MPLS core. The MES builds a cache of IP / MAC address bindings from these snooped messages. The MES then uses this cache to respond to ARP requests ingress on access ports and targeting hosts that are in remote sites. If the MES finds a match for the IP address in its ARP cache, it responds back to the requesting host and drops the request. Otherwise, if it does not find a match, then the request is flooded over the MPLS network using either ingress replication or LSM.

9. Seamless Interworking with TRILL and IEEE 802.1aq/802.1Qbp

PBB-EVPN enables seamless connectivity of TRILL or 802.1aq/802.1Qbp networks over an MPLS/IP core while maintaining control-plane separation among these networks. We will refer to one or any of TRILL, 802.1aq or 802.1Qbp networks collectively as 'NG-Ethernet networks' thereafter.
Every NG-Ethernet network that is connected to the MPLS core runs an

independent instance of the corresponding IS-IS control-plane. Each MES participates in the NG-Ethernet network control plane of its local site. The MES peers, in IS-IS protocol, with the switches internal to the site, but does not terminate the TRILL / PBB data-

plane encapsulation. So, from a control-plane viewpoint, the MES appears as an edge switch; whereas, from a data-plane viewpoint, the MES appears as a core switch to the NG-Ethernet network.
The MES nodes encapsulate TRILL / PBB frames with MPLS in the imposition path, and de-capsulate them in the disposition path.

9.1. TRILL Nickname Advertisement Route

A new BGP route is defined to support the interconnection of TRILL networks over PBB-EVPN: the TRILL Nickname Advertisement' route, encoded as follows:

```
+---------------------------------------+
| RD (8 octets)                         |
+---------------------------------------+
|Ethernet Segment Identifier (10 octets)|
+---------------------------------------+
| Ethernet Tag ID (4 octets)            |
+---------------------------------------+
| Nickname Length (1 octet)             |
+---------------------------------------+
| RBridge Nickname (2 octets)           |
+---------------------------------------+
| MPLS Label (n * 3 octets)             |
+---------------------------------------+
```

Figure 4: TRILL Nickname Advertisement Route

The MES uses this route to advertise the reachability of TRILL RBridge nicknames to other MES nodes in the VPN instance. The MPLS label advertised in this route can be allocated on a per VPN basis and serves the purpose of identifying to the disposition MES that the MPLS-encapsulated packet holds an MPLS encapsulated TRILL frame.

The encapsulation for the transport of TRILL frames over MPLS is encoded as shown in the figure below:

```
+------------------+
| IP/MPLS Header   |
+------------------+
| TRILL Header     |
+------------------+
| Ethernet Header  |
+------------------+
| Ethernet Payload |
+------------------+
| Ethernet FCS     |
+------------------+
```

Figure 5: TRILL over MPLS Encapsulation

It is worth noting here that while it is possible to transport Ethernet encapsulated TRILL frames over MPLS, that approach unnecessarily wastes 16 bytes per packet. That approach further requires either the use of well-known MAC addresses or having the MES nodes advertise in BGP their device MAC addresses, in order to resolve the TRILL next-hop L2 adjacency. To that end, it is simpler and more efficient to transport TRILL natively over MPLS and that is why we are defining the above BGP route for TRILL Nickname advertisement.

9.2. IEEE 802.1aq / 802.1Qbp B-MAC Advertisement Route

B-MAC addresses associated with 802.1aq / 802.1Qbp switches are advertised using the BGP MAC Advertisement route already defined in [E-VPN].

The encapsulation for the transport of PBB frames over MPLS is similar to that of classical Ethernet, albeit with the additional PBB header, as shown in the figure below:

```
+------------------+
| IP/MPLS Header   |
+------------------+
| PBB Header       |
+------------------+
| Ethernet Header  |
+------------------+
| Ethernet Payload |
+------------------+
| Ethernet FCS     |
+------------------+
```

Figure 6: PBB over MPLS Encapsulation

9.3. Operation

For correct connectivity, the TRILL Nicknames or 802.1aq/802.1Qbp B-MACs must be globally unique in the network. This can be achieved, for instance, by using a hierarchical Nickname (or B-MAC) assignment paradigm, and encoding a Site ID in the high-order bits of the Nickname (or B-MAC):

Nickname (or B-MAC) = [Site ID : Rbridge ID (or MAC)]

The only practical difference between TRILL Nicknames and B-MACs, in this regards, is with respect to the size of the address space:

Nicknames are 16-bits wide whereas B-MACs are 48-bits wide.

Every MES then advertises (in BGP) the Nicknames (or B-MACs) of all switches local to its site in the TRILL Nickname Advertisement routes (or MAC Advertisement routes).
Furthermore, the MES advertises in IS-IS (to the local island) the Rbridge nicknames (or B-MACs) of all remote switches in all the other TRILL (or IEEE 802.1aq/802.1Qbp) islands that the MES has learned via BGP.

Note that by having multiple MES nodes (connected to the same TRILL or 802.1aq /802.1Qbp island) advertise routes to the same RBridge nickname (or B-MAC), with equal BGP Local_Pref attribute, it is possible to perform active/active load-balancing to/from the MPLS core.

When a MES receives an Ethernet-encapsulated TRILL frame from the access side, it removes the Ethernet encapsulation (i.e. outer MAC header), and performs a lookup on the egress RBridge nickname in the TRILL header to identify the next-hop. If the lookup yields that the next hop is a remote MES, the local MES would then encapsulate the TRILL frame in MPLS. The label stack comprises of the VPN label (advertised by the remote MES), followed by an LSP/IGP label. From that point onwards, regular MPLS forwarding is applied.

On the disposition MES, assuming penultimate-hop-popping is employed, the MES receives the MPLS-encapsulated TRILL frame with a single label: the VPN label. The value of the label indicates to the disposition MES that this is a TRILL packet, so the label is popped, the TTL field (in the TRILL header) is reinitialized and normal TRILL processing is employed from this point onwards.

By the same token, when a MES receives a PBB-encapsulated Ethernet frame from the access side, it performs a lookup on the B-MAC destination address to identify the next hop. If the lookup yields that the next hop is a remote MES, the local MES would then encapsulate the PBB frame in MPLS. The label stack comprises of the VPN label (advertised by the remote PE), followed by an LSP/IGP label. From that point onwards, regular MPLS forwarding is applied.

On the disposition MES, assuming penultimate-hop-popping is employed, the MES receives the MPLS-encapsulated PBB frame with a single label: the VPN label. The value of the label indicates to the disposition MES that this is a PBB frame, so the label is popped, the TTL field (in the 802.1Qbp F-Tag) is reinitialized and normal PBB processing is employed from this point onwards.

10. Solution Advantages

In this section, we discuss the advantages of the PBB-EVPN solution in the context of the requirements set forth in section 3 above.

10.1. MAC Advertisement Route Scalability

In PBB-EVPN the number of MAC Advertisement Routes is a function of the number of segments (sites), rather than the number of hosts/servers. This is because the B-MAC addresses of the MESes, rather than C-MAC addresses (of hosts/servers) are being advertised in BGP. And, as discussed above, there's a one-to-one mapping between multi-homed segments and B-MAC addresses, whereas there's a one-to-one or many-to-one mapping between single-homed segments and B-MAC addresses for a given MES. As a result, the volume of MAC Advertisement Routes in PBB-EVPN is multiple orders of magnitude less than E-VPN.

10.2. C-MAC Mobility with MAC Sub-netting

In PBB-EVPN, if a MES allocates its B-MAC addresses from a contiguous range, then it can advertise a MAC prefix rather than individual 48-bit addresses. It should be noted that B-MAC addresses can easily be assigned from a contiguous range because MES nodes are within the provider administrative domain; however, CE devices and hosts are typically not within the provider administrative domain. The advantage of such MAC address sub-netting can be maintained even as C-MAC addresses move from one Ethernet segment to another. This is because the C-MAC address to B-MAC address association is learnt in the data-plane and C-MAC addresses are not advertised in BGP. To illustrate how this compares to E-VPN, consider the following example:

If a MES running E-VPN advertises reachability for a MAC subnet that spans N addresses via a particular segment, and then 50% of the MAC addresses in that subnet move to other segments (e.g. due to virtual machine mobility), then in the worst case, N/2 additional MAC Advertisement routes need to be sent for the MAC addresses that have moved. This defeats the purpose of the sub-netting. With PBB-EVPN, on the other hand, the sub-netting applies to the B-MAC addresses which are statically associated with MES nodes and are not subject to mobility. As C-MAC addresses move from one segment to another, the binding of C-MAC to B-MAC addresses is updated via data-plane learning.

10.3. C-MAC Address Learning and Confinement

In PBB-EVPN, C-MAC address reachability information is built via data-plane learning. As such, MES nodes not participating in active conversations involving a particular C-MAC address will purge that address from their forwarding tables. Furthermore, since C-MAC addresses are not distributed in BGP, MES nodes will not maintain any record of them in control-plane routing table.

10.4. Seamless Interworking with TRILL and 802.1aq Access Networks

Consider the scenario where two access networks, one running MPLS and the other running 802.1aq, are interconnected via an MPLS backbone network. The figure below shows such an example network.

```
                            +--------------+
                            |              |
            +---------+     |     MPLS     |    +---------+
    +----+  |         |   +----+        +----+  |         |  +----+
    | CE |--|         |   |MES1|        |MES2|  |         |--| CE |
    +----+  | 802.1aq |---|    |        |    |--|  MPLS   |  +----+
    +----+  |         |   +----+        +----+  |         |  +----+
    | CE |--|         |     |   Backbone   |    |         |--| CE |
    +----+  +---------+     +--------------+    +---------+  +----+
```

Figure 7: Interoperability with 802.1aq

If the MPLS backbone network employs E-VPN, then the 802.1aq data-plane encapsulation must be terminated on MES1 or the edge device connecting to MES1. Either way, all the MES nodes that are part of the associated service instances will be exposed to all the C-MAC addresses of all hosts/servers connected to the access networks. However, if the MPLS backbone network employs PBB-EVPN, then the 802.1aq encapsulation can be extended over the MPLS backbone, thereby maintaining C-MAC address transparency on MES1. If PBB-EVPN is also extended over the MPLS access network on the right, then C-MAC addresses would be transparent to MES2 as well.

Interoperability with TRILL access network will be described in future revision of this draft.

10.5. Per Site Policy Support

In PBB-EVPN, a unique B-MAC address can be associated with every site (single-homed or multi-homed). Given that the B-MAC addresses are sent in BGP MAC Advertisement routes, it is possible to define per site (i.e. B-MAC) forwarding policies including policies for E-TREE service.

10.6. Avoiding C-MAC Address Flushing

With PBB-EVPN, it is possible to avoid C-MAC address flushing upon topology change affecting a multi-homed device. To illustrate this, consider the example network of Figure 1. Both MES1 and MES2 advertize the same B-MAC address (BM1) to MES3. MES3 then learns the C-MAC addresses of the servers/hosts behind CE1 via data-plane learning. If AC1 fails, then MES3 does not need to flush any of the C-MAC addresses learnt and associated with BM1. This is because MES1 will withdraw the MAC Advertisement routes associated with BM1, thereby leading MES3 to have a single adjacency (to MES2) for this B-MAC address. Therefore, the topology change is communicated to MES3 and no C-MAC address flushing is required.

11. Acknowledgements

TBD.

12. Security Considerations

There are no additional security aspects beyond those of VPLS/H-VPLS that need to be discussed here.

13. IANA Considerations

This document requires IANA to assign a new SAFI value for L2VPN_MAC SAFI.

14. Intellectual Property Considerations

This document is being submitted for use in IETF standards discussions.

15. Normative References

[802.1ah] "Virtual Bridged Local Area Networks Amendment 7: Provider Backbone Bridges", IEEE Std. 802.1ah-2008, August 2008.

16. Informative References

[PBB-VPLS] Sajassi et al., "VPLS Interoperability with Provider Backbone Bridges", draft-ietf-l2vpn-vpls-pbb-interop-00.txt, work in progress, September, 2011.

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (E-VPN)", draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt, work in progress, October, 2010.

[E-VPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-raggarwa-sajassi-l2vpn-evpn-01.txt, November, 2010.
, work in progress, June, 2010.

17. Authors' Addresses

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sajassi@cisco.com

Samer Salam

Cisco
595 Burrard Street, Suite 2123
Vancouver, BC V7X 1J1, Canada
Email: ssalam@cisco.com

Sami Boutros
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sboutros@cisco.com

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

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Florin Balus
Alcatel-Lucent
701 E. Middlefield Road
Mountain View, CA, USA 94043
Email: florin.balus@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.be

Clarence Filsfils
Cisco
Email: cfilsfil@cisco.com

Dennis Cai
Cisco
Email: dcai@cisco.com

Lizhong Jin
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China
Email: lizhong.jin@zte.com.cn