Internet Working Group
Internet Draft
Category: Standards Track

A. Sajassi
P. Brissette
Cisco
R. Schell
Verizon
J. Drake
Juniper
J. Rabadan
Nokia

Expires: July 18, 2019

January 18, 2019

EVPN Virtual Ethernet Segment
draft-ietf-bess-evpn-virtual-eth-segment-04

Abstract

EVPN and PBB-EVPN introduce a family of solutions for multipoint Ethernet services over MPLS/IP network with many advanced features among which their multi-homing capabilities. These solutions define two types of multi-homing for an Ethernet Segment (ES): 1) Single-Active and 2) All-Active, where an Ethernet Segment is defined as a set of links between the multi-homed device/network and a set of PE devices that they are connected to.

Some Service Providers want to extend the concept of the physical links in an ES to Ethernet Virtual Circuits (EVCs) where many of such EVCs can be aggregated on a single physical External Network-to-Network Interface (ENNI). An ES that consists of a set of EVCs instead of physical links is referred to as a virtual ES (vES). This draft describes the requirements and the extensions needed to support vES in EVPN and PBB-EVPN.

Copyright and License Notice

Conventions

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.

Table of Contents

1. Introduction

[RFC7432] and [RFC7623] introduce a family of solutions for multipoint Ethernet services over MPLS/IP network with many advanced features among which their multi-homing capabilities. These solutions define two types of multi-homing for an Ethernet Segment (ES): 1) Single-Active and 2) All-Active, where an Ethernet Segment is defined as a set of links between the multi-homed device/network and a set of PE devices that they are connected to.

This document extends the Ethernet Segment concept so that an ES can be associated to a set of EVCs (e.g., VLANs) or other objects such as MPLS Label Switch Paths (LSPs) or Pseudowires (PWs).

1.1 Virtual Ethernet Segments in Access Ethernet Networks

Some Service Providers (SPs) want to extend the concept of the physical links in an ES to Ethernet Virtual Circuits (EVCs) where many of such EVCs (e.g., VLANs) can be aggregated on a single physical External Network-to-Network Interface (ENNI). An ES that consists of a set of EVCs instead of physical links is referred to as a virtual ES (vES). Figure 1 depicts two PE devices (PE1 and PE2) each with an ENNI where a number of vES's are aggregated on - each of which through its associated EVC.

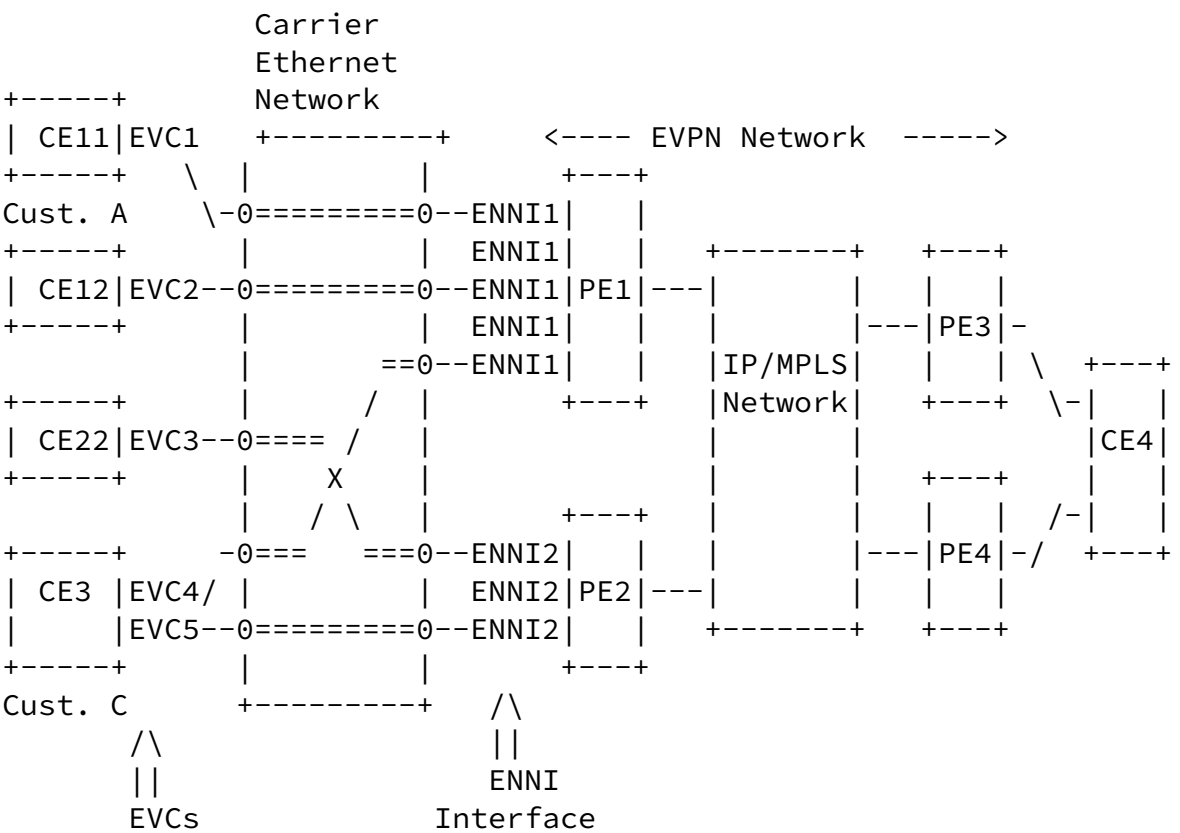

Carrier
Ethernet
Network
<---- EVPN Network ----->
CE11
EVC1
Cust. A
CE12
EVC2
CE22
EVC3
CE3
EVC4/
EVC5
Cust. C
ENNI1
ENNI1
ENNI1
ENNI1
ENNI2
ENNI2
ENNI2
PE1
PE2
IP/MPLS
Network
PE3
PE4
CE4
EVCs
ENNI
Interface

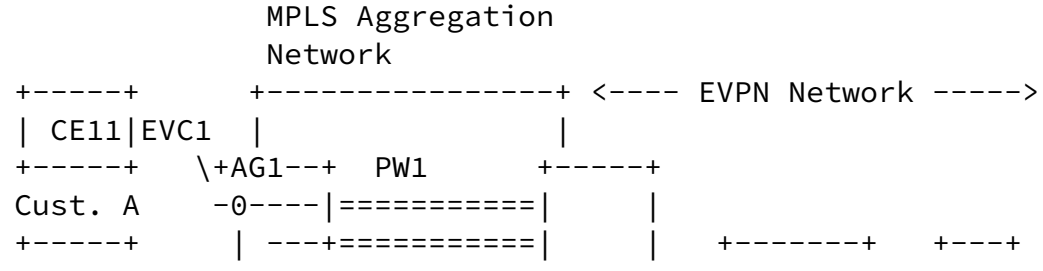
Figure 1: DHD/DHN (both SA/AA) and SH on same ENNI

ENNIs are commonly used to reach off-network / out-of-franchise customer sites via independent Ethernet access networks or third-party Ethernet Access Providers (EAP) (see Figure 1). ENNIs can aggregate traffic from hundreds to thousands of vES's; where, each vES is represented by its associated EVC on that ENNI. As a result, ENNIs and their associated EVCs are a key element of SP off-networks that are carefully designed and closely monitored.

In order to meet customer's Service Level Agreements (SLA), SPs build redundancy via multiple EVPN PEs and across multiple ENNIs (as shown in Figure 1) where a given vES can be multi-homed to two or more EVPN PE devices (on two or more ENNIs) via their associated EVCs. Just like physical ES's in [RFC7432] and [RFC7623] solutions, these vES's can be single-homed or multi-homed ES's and when multi-homed, then can operate in either Single-Active or All-Active redundancy modes. In a typical SP off-network scenario, an ENNI can be associated with several thousands of single-homed vES's, several hundreds of Single-Active vES's and it may also be associated with tens or hundreds of All-Active vES's.

1.2 Virtual Ethernet Segments in Access MPLS Networks

Other Service Providers (SPs) want to extend the concept of the physical links in an ES to individual Pseudowires (PWs) or to MPLS Label Switched Paths (LSPs) in Access MPLS networks - i.e., a vES consisting of a set of PWs or a set of LSPs. Figure 2 illustrates this concept.



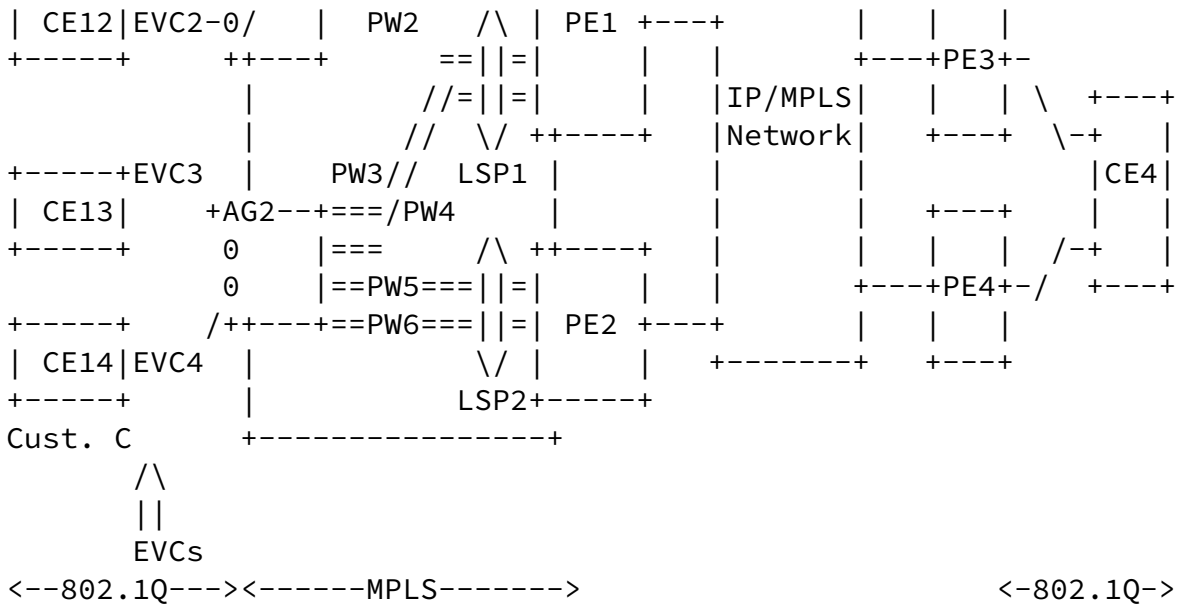


Figure 2: DHN and SH on Access MPLS networks

In some cases, Service Providers use Access MPLS Networks that belong to separate administrative entities or third parties as a way to get access to the their own IP/MPLS network infrastructure. This is the case illustrated in Figure 2.

In such scenarios, a virtual ES (vES) is defined as a set of individual PWs if they cannot be aggregated into a common LSP. If the aggregation of PWs is possible, the vES can be associated to an LSP in a given PE. In the example of Figure 2, EVC3 is connected to a VPWS instance in AG2 that is connected to PE1 and PE2 via PW3 and PW5

respectively. EVC4 is connected to a separate VPWS instance on AG2 that gets connected to an EVI on PE1 and PE2 via PW4 and PW6, respectively. Since the PWs for the two VPWS instances can be aggregated into the same LSPs going to the EVPN network, a common virtual ES can be defined for LSP1 and LSP2. This vES will be shared by two separate EVIs in the EVPN network.

In some cases, this aggregation of PWs into common LSPs may not be

possible. For instance, if PW3 were terminated into a third PE, e.g. PE3, instead of PE1, the vES would need to be defined on a per individual PW on each PE, i.e. PW3 and PW5 would belong to ES-1, whereas PW4 and PW6 would be associated to ES-2.

For MPLS/IP access networks where a vES represents a set of PWs or LSPs, this document extends Single-Active multi-homing procedures of [RFC7432] and [7623] to vES. The vES extension to All-Active multi-homing is outside of the scope of this document for MPLS/IP access networks.

This draft describes requirements and the extensions needed to support vES in [RFC7432] and [RFC7623]. Section 3 lists the set of requirements for vES's. Section 4 describes extensions for vES that are applicable to EVPN solutions including [RFC7432], [RFC7623], and [RFC8214]. Furthermore, these extensions meet the requirements described in section 3. Section 5 describes the failure handling and recovery for vES's in [RFC7432] and [RFC7623]. Section 6 covers scalability and fast convergence required for vES's in [RFC7432] and [RFC7623].

2. Terminology

AC: Attachment Circuit
BEB: Backbone Edge Bridge
B-MAC: Backbone MAC Address
CE: Customer Edge
CFM: Connectivity Fault Management
C-MAC: Customer/Client MAC Address
DHD: Dual-homed Device
DHN: Dual-homed Network
ENNI: External Network-Network Interface
ES: Ethernet Segment
ESI: Ethernet-Segment Identifier
EVC: Ethernet Virtual Circuit
EVPN: Ethernet VPN
I-SID: Service Instance Identifier (24 bits and global within a PBB network see [RFC7080])
LACP: Link Aggregation Control Protocol

PBB: Provider Backbone Bridge

PBB-EVPN: Provider Backbone Bridge EVPN
PE: Provider Edge
SH: Single-Homed

Single-Active Redundancy Mode (SA): When only a single PE, among a group of PEs attached to an Ethernet-Segment, is allowed to forward traffic to/from that Ethernet Segment, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode (AA): When all PEs attached to an Ethernet segment are allowed to forward traffic to/from that Ethernet-Segment, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

3. Requirements

This section describes the requirements specific to virtual Ethernet Segment (vES) for (PBB-)EVPN solutions. These requirements are in addition to the ones described in [RFC7209], [RFC7432], and [RFC7623].

3.1. Single-Homed & Multi-Homed Virtual Ethernet Segments

A PE needs to support the following types of vES's:

(R1a) A PE MUST handle single-homed vES's on a single physical port (e.g., single ENNI)

(R1b) A PE MUST handle a mix of Single-Homed vES's and Single-Active multi-homed vES's simultaneously on a single physical port (e.g., single ENNI). Single-Active multi-homed vES's will be simply referred to as Single-Active vES's through the rest of this document.

(R1c) A PE MAY handle All-Active multi-homed vES's on a single physical port. All-Active multi-homed vES's will be simply referred to as All-Active vES's through the rest of this document.

(R1d) A PE MAY handle a mixed of All-Active vES's along with other types of vES's on a single physical port

(R1e) A Multi-Homed vES (Single-Active or All-Active) can be spread across any two or more PEs (on two or more ENNIs)

3.2. Scalability

A single physical port (e.g., ENNI) can be associated with many vES's. The following requirements give a quantitative measure for each vES type.

(R2a) A PE SHOULD handle very large number of Single-Homed vES's on a single physical port (e.g., thousands of vES's on a single ENNI)

(R2b) A PE SHOULD handle large number of Single-Active vES's on a single physical port (e.g., hundreds of vES's on a single ENNI)

(R2c) A PE MAY handle large number of All-Active Multi-Homed vES's on a single physical port (e.g., hundreds of vES's on a single ENNI)

(R2d) A PE SHOULD handle the above scale for a mix of Single-homed vES's and Single-Active vES's simultaneously on a single physical port (e.g., single ENNI)

(R4e) A PE MAY handle the above sale for a mixed of All-Active Multi-Homed vES's along with other types of vES's on a single physical port

3.3. Local Switching

Many vES's of different types can be aggregated on a single physical port on a PE device and some of these vES can belong to the same service instance (or customer). This translates into the need for supporting local switching among the vES's of the same service instance on the same physical port (e.g., ENNI) of the PE.

(R3a) A PE MUST support local switching among different vES's belonging to the same service instance (or customer) on a single physical port. For example, in Figure 1, PE1 MUST support local switching between CE11 and CE12 (both belonging to customer A) that are mapped to two Single-homed vES's on ENNI1.

In case of Single-Active vES's, the local switching is performed among active EVCs belonging to the same service instance on the same ENNI.

3.4. EVC Service Types

A physical port (e.g., ENNI) of a PE can aggregate many EVCs each of which is associated with a vES. Furthermore, an EVC may carry one or more VLANs. Typically, an EVC carries a single VLAN and thus it is associated with a single broadcast domain. However, there is no restriction on an EVC to carry more than one VLAN.

(R4a) An EVC can be associated with a single broadcast domain - e.g., VLAN-based service or VLAN bundle service

(R4b) An EVC MAY be associated with several broadcast domains - e.g., VLAN-aware bundle service

In the same way, a PE can aggregate many LSPs and PWs. In the case of individual PWs per vES, typically a PW is associated with a single broadcast domain, but there is no restriction on the PW to carry more than one VLAN if the PW is of type Raw mode.

(R4c) A PW can be associated with a single broadcast domain - e.g., VLAN-based service or VLAN bundle service.

(R4d) An PW MAY be associated with several broadcast domains - e.g., VLAN-aware bundle service."

3.5. Designated Forwarder (DF) Election

Section 8.5 of [RFC7432] describes the default procedure for DF election in EVPN which is also used in [RFC7623] and [RFC8214]. This default DF election procedure is performed at the granularity of <ESI, Ethernet Tag>. In case of a vES, the same EVPN default procedure for DF election also applies; however, at the granularity of <vESI, Ethernet Tag>; where vESI is the virtual Ethernet Segment Identifier. As in [RFC7432], this default procedure for DF election at the granularity of <vESI, Ethernet Tag> is also referred to as "service carving"; where, Ethernet Tag is represented by an I-SID in PBB-EVPN and by a VLAN ID (VID) in EVPN. With service carving, it is possible to evenly distribute the DFs for different vES's among different PEs, thus distributing the traffic among different PEs. The following list the requirements apply to DF election of vES's for EVPN.

(R5a) A vES with m EVCs can be distributed among n ENNIs belonging to p PEs in any arbitrary oder; where n >= p >= m. For example, if there is an vES with 2 EVCs and there are 5 ENNIs on 5 PEs (PE1 through PE5), then vES can be dual-homed to PE2 and PE4 and the DF election must be performed between PE2 and PE4.

(R5b) Each vES MUST be identified by its own virtual ESI (vESI)

3.6. OAM

In order to detect the failure of individual EVC and perform DF election for its associated vES as the result of this failure, each EVC should be monitored independently.

(R6a) Each EVC SHOULD be monitored for its health independently

(R6b) A single EVC failure (among many aggregated on a single physical port/ENNI) MUST trigger DF election for its associated vES.

3.7. Failure & Recovery

(R7a) Failure and failure recovery of an EVC for a Single-homed vES SHALL NOT impact any other EVCs for its own service instance or any other service instances. In other words, for PBB-EVPN, it SHALL NOT trigger any MAC flushing both within its own I-SID as well as other I-SIDs.

(R7b) In case of All-Active Multi-Homed vES, failure and failure recovery of an EVC for that vES SHALL NOT impact any other EVCs for its own service instance or any other service instances. In other words, for PBB-EVPN, it SHALL NOT trigger any MAC flushing both within its own I-SID as well as other I-SIDs.

(R7c) Failure & failure recovery of an EVC for a Single-Active vES SHALL only impact its own service instance. In other words, for PBB-EVPN, MAC flushing SHALL be limited to the associated I-SID only and SHALL NOT impact any other I-SIDs.

(R7d) Failure & failure recovery of an EVC for a Single-Active vES MAY only impact C-MACs associated with MHD/MHNs for that service instance. In other words, MAC flushing SHOULD be limited to single service instance (I-SID in the case of PBB-EVPN) and only CMACs for Single-Active MHD/MHNs.

3.8. Fast Convergence

Since large number of EVCs (and their associated vES's) are aggregated via a single physical port (e.g., ENNI), then the failure

of that physical port impacts large number of vES's and triggers large number of ES route withdrawals. Formulating, sending, receiving, and processing such large number of BGP messages can introduce delay in DF election and convergence time. As such, it is highly desirable to have a mass-withdraw mechanism similar to the one in the [RFC7432] for withdrawing large number of Ethernet A-D routes.

(R8a) There SHOULD be a mechanism equivalent to EVPN mass-withdraw such that upon an ENNI failure, only a single BGP message is needed to indicate to the remote PEs to trigger DF election for all impacted vES associated with that ENNI.

4. Solution Overview

The solutions described in [RFC7432] and [RFC7623] are leveraged as

is with one simple modification and that is the ESI assignment is performed for a group of EVCs or LSPs/PWs instead of a group of physical links. In other words, the ESI is associated with a virtual ES (vES) and that's why it will be referred to as vESI.

For the EVPN solution, everything basically remains the same except for the handling of physical port failure where many vES's can be impacted. Section 5.1 and 5.3 below describe the handling of physical port/link failure for EVPN. In a typical multi-homed operation, MAC addresses are learned behind a vES are advertised with the ESI corresponding to the vES (i.e., vESI). EVPN aliasing and mass-withdraw operations are performed with respect to vES. In other words, the Ethernet A-D routes for these operations are advertised with vESI instead of ESI.

For PBB-EVPN solution, the main change is with respect to the BMAC address assignment which is performed similar to what is described in section 7.2.1.1 of [RFC7623] with the following refinements:

- One shared BMAC address SHOULD used per PE for the single-homed vES's. In other words, a single BMAC is shared for all single-homed vES's on that PE.

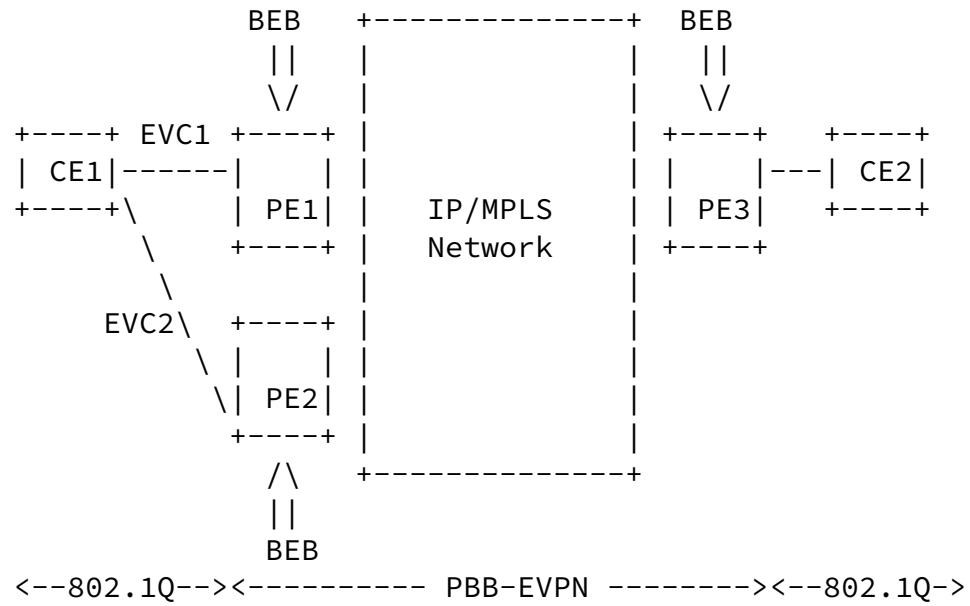
- One shared BMAC address SHOULD be used per PE per physical port (e.g., ENNI) for the Single-Active vES's. In other words, a single

BMAC is shared for all Single-Active vES's that share the same ENNI.

- One shared BMAC address MAY be used for all Single-Active vES's on that PE.

- One BMAC address SHOULD be used per set of EVCs representing an All-Active multi-homed vES. In other words, a single BMAC address is used per vES for All-Active multi-homing scenarios.

- A single BMAC address MAY also be used per vES per PE for Single-Active multi-homing scenarios.



4.1. EVPN DF Election for vES

The procedure for service carving for virtual Ethernet Segments is the same as the one outlined in section 8.5 of [RFC7432] except for the fact that ES is replaced with vES. For the sake of clarity and completeness, this procedure is repeated below:

1. When a PE discovers the vESI or is configured with the vESI associated with its attached vES, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute.

2. The PE then starts a timer (default value = 3 seconds) to allow the reception of Ethernet Segment routes from other PE nodes connected to the same vES. This timer value MUST be same across all PEs connected to the same vES.

3. When the timer expires, each PE builds an ordered list of the IP addresses of all the PE nodes connected to the vES (including itself), in increasing numeric value. Each IP address in this list is extracted from the "Originator Router's IP address" field of the advertised Ethernet Segment route. Every PE is then given an ordinal indicating its position in the ordered list, starting with 0 as the ordinal for the PE with the numerically lowest IP address. The ordinals are used to determine which PE node will be the DF for a given EVPN instance on the vES using the following rule: Assuming a redundancy group of N PE nodes, the PE with ordinal i is the DF for an EVPN instance with an associated Ethernet Tag value of V when (V mod N) = i.

It should be noted that using "Originator Router's IP address" field in the Ethernet Segment route to get the PE IP address needed for the ordered list, allows for a CE to be multi-homed across different ASes if such need ever arises.

4. The PE that is elected as a DF for a given EVPN instance will unblock traffic for that EVPN instance. Note that the DF PE unblocks all traffic in both ingress and egress directions for Single-Active vES and unblocks multi-destination in egress direction for All-Active Multi-homed vES. All non-DF PEs block all traffic in both ingress and egress directions for Single-Active vES and block multi-destination

traffic in the egress direction for All-Active multi-homed vES.

In the case of an EVC failure, the affected PE withdraws its Ethernet Segment route if there are no more EVCs associated to the vES in the PE. This will re-trigger the DF Election procedure on all the PEs in the Redundancy Group. For PE node failure, or upon PE commissioning or decommissioning, the PEs re-trigger the DF Election Procedure across all affected vES's. In case of a Single-Active multi-homing, when a service moves from one PE in the Redundancy Group to another PE as a result of DF re-election, the PE, which ends up being the elected DF for the service, SHOULD trigger a MAC address flush notification towards the associated vES. This can be done, for e.g. using IEEE 802.1ak MVRP 'new' declaration.

For LSP and PW based vES, the non-DF PE SHOULD signal PW-status 'standby' signaling to the Aggregation PE (e.g., AG PE in Figure 2), and the new DF PE MAY send an LDP MAC withdraw message as a MAC address flush notification. It should be noted that the PW-status is signaled for the scenarios where there is a one-to-one mapping between EVI/BD and the PW.

5. Failure Handling & Recovery

There are a number of failure scenarios to consider such as:

A: CE Uplink Port Failure
B: Ethernet Access Network Failure
C: PE Access-facing Port or link Failure
D: PE Node Failure
E: PE isolation from IP/MPLS network

[RFC7432], [RFC7623], and [RFC8214] solutions provide protection against such failures as described in the corresponding references. In the presence of virtual Ethernet Segments (vES's) in these solutions, besides the above failure scenarios, there is one more scenario to consider and that is EVC failure. This implies that individual EVCs need to be monitored and upon their failure

detection, appropriate DF election procedures and failure recovery mechanism need to be executed.

[ETH-OAM] is used for monitoring EVCs and upon failure detection of a

given EVC, DF election procedure per section [4.1] is executed. For PBB-EVPN, some extensions are needed to handle the failure and recovery procedures of [RFC7623] in order to meet the above requirements. These extensions are describe in the next section.

[MPLS-OAM] and [PW-OAM] are used for monitoring the status of LSPs and/or PWs associated to vES.

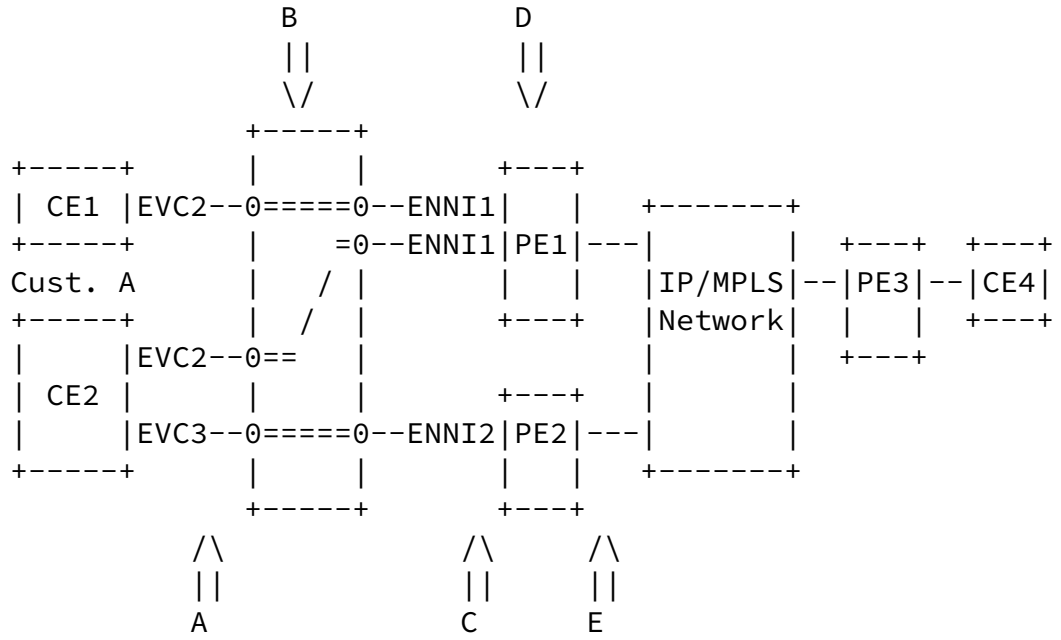


Figure 4: Failure Scenarios A,B,C,D and E

5.1. Failure Handling for Single-Active vES in EVPN

When a DF PE connected to a Single-Active multi-homed Ethernet Segment loses connectivity to the segment, due to link or port failure, it signals to the remote PEs to withdraw all MAC addresses associated with that Ethernet Segment. This is done by advertising a mass-withdraw message using Ethernet A-D per-ES route. It should be noted that for dual-homing use cases where there is only a single backup path, MAC withdraw can be avoided by the remote PEs as they can simply update their nexthop associated with the affected MAC entries to the backup path per procedure described in section 8.2 of [RFC7432].

In case of an EVC failure which impacts a single vES, the exact same EVPN procedure is used. In this case, the message using Ethernet A-D

per ES route carries the vESI representing the vES which in turn is associated with the failed EVC. The remote PEs upon receiving this message perform the same procedures outlined in section 8.2 of [RFC7432].

5.2. EVC Failure Handling for Single-Active vES in PBB-EVPN

When a PE connected to a Single-Active multi-homed Ethernet Segment loses connectivity to the segment, due to link or port failure, it signals the remote PE to flush all CMAC addresses associated with that Ethernet Segment. This is done by advertising a BMAC route along with MAC Mobility Extended community.

In case of an EVC failure that impacts a single vES, if the above PBB-EVPN procedure is used, it results in excessive CMAC flushing because a single physical port can support large number of EVCs (and their associated vES's) and thus advertising a BMAC corresponding to the physical port with MAC mobility Extended community will result in flushing CMAC addresses not just for the impacted EVC but for all other EVCs on that port.

In order to reduce the scope of CMAC flushing to only the impacted service instances (the service instance(s) impacted by the EVC failure), the BGP flush message is sent along with a list of impacted I-SID(s) represented by the new EVPN I-SID Extended Community as defined in section 6. Since typically an EVC maps to a single broadcast domain and thus a single service instance, the list only contains a single I-SID. However, if the failed EVC carries multiple VLANs each with its own broadcast domain, then the list contains several I-SIDs - one for each broadcast domain. This new BGP flush message basically instructs the remote PE to perform flushing for CMACs corresponding to the advertised BMAC only across the advertised list of I-ISIDs.

The new I-SID Extended Community provides a way to encode upto 24 I-SIDs in each Extended Community if the impacted I-SIDs are sequential (the base I-SID value plus the next 23 I-SID values). If the number of I-SIDs associated with a failed EVC is large or if the affected I-SIDs are not sequential, then multiple I-SID Extended Communities can be sent along with the flush message. However, if the number of affected I-SIDs is very large such that the corresponding I-SID Extended Communities cannot be fitted in a single BGP attribute, then the EVC failure can be treated as a port failure and the procedures of section 5.4 can be exercised (i.e., a single BGP flush message without the I-SID list can be transmitted). When the BGP flush message is transmitted without the I-SID list, then it instructs the receiving PEs to flush CMACs associated with that BMAC across all I-

SIDs.

There can be scenarios (although unlikely) where multiple EVCs within the same physical port can fail within a short time resulting in the PE advertising multiple BGP flush messages each with their own list of I-SIDs; however, the route reflector receiving these messages will only send the last flush message. This results in PEs receiving such flush messages not to properly flush all the affected I-SIDs. In order to address such scenarios, a timer T1 is started upon an EVC1 failure on the advertising PE. If there is another EVC2 failure within T1, affected I-SIDs are aggregated for both EVC1 and EVC2 to be sent along the new flush message. Furthermore when EVC2 failure occurs, another timer T2 (with the same value as T1) is started to keep track of the affected I-SIDs for EVC2. Such I-SID aggregation may result in multiple flushing for the same I-SID(s) on the receiving PEs. The default value for this timer T is 10 seconds.

The I-SID dependent flushing mechanism described in this section is also backward compatible for the PEs supporting [RFC7623] such that the PEs that don't understand the I-SID list (i.e., the new I-SID Extended Community) simply ignore it and default to flushing all the I-SIDs for the B-MAC - i.e., the PEs default to per-port flushing described in section 5.4.

The above BMAC route that is advertised with the MAC Mobility Extended Community, can either represent the MAC address of the physical port that the failed EVC is associated with, or it can represent the MAC address of the PE. In the latter case, this is the dedicated per-PE MAC address used for all Single-Active vES's on that PE. The former one performs better than the latter one in terms of reducing the scope of flushing and thus it is the recommended approach because only CMAC addresses for the impacted service instances on the failed EVC are flushed.

5.3. Port Failure Handling for Single-Active vES's in EVPN

When a large number of EVCs are aggregated via a single physical port on a PE; where each EVC corresponds to a vES, then the port failure impacts all the associated EVCs and their corresponding vES's. If the number of EVCs corresponding to the Single-Active vES's for that physical port is in thousands, then thousands of service instances

are impacted. Therefore, the BGP flush message need to be inclusive of all these impacted service instances. In order to achieve this, the following extensions are added to the baseline EVPN mechanism:

1) When a PE advertises an Ether-AD per ES route for a given vES, it colors it with the MAC address of the physical port which is associated with that vES using EVPN Router's MAC Extended Community per [EVPN-IRB]. The receiving PEs take note of this color and create

a list of vES's for this color.

2) Upon a port failure (e.g., ENNI failure), the PE advertise a special mass-withdraw message with the MAC address of the failed port (i.e., the color of the port) encoded in the ESI field. For this encoding, type 3 ESI is used with the MAC field set to the MAC address of the port and the 3-octet local discriminator field set to 0xFFFFFF. This mass-withdraw route is advertised with a list of Route Targets corresponding to the impacted service instances. If the number of Route Targets is more than they can fit into a single attribute, then a set of Ethernet A-D per ES routes are advertised. The remote PEs upon receiving this message, realize that this is a special mass-withdraw message and they access the list of the vES's for the specified color. Next, they initiate mass-withdraw procedure for each of the vES's in the list.

In scenarios where a logical ENNI is used the above procedure equally applies. The logical ENNI is represented by type 3 ESI and the MAC address used in the ENNI's ESI is used as a color for vES's as described above.

5.4. Port Failure Handling for Single-Active vES's in PBB-EVPN

When a large number of EVCs are aggregated via a single physical port on a PE; where each EVC corresponds to a vES, then the port failure impacts all the associated EVCs and their corresponding vES's. If the number of EVCs corresponding to the Single-Active vES's for that physical port is in thousands, then thousands of service instances (I-SIDs) are impacted. In such failure scenarios, the following two MAC flushing mechanisms per [RFC7623] can be performed.

1) If the MAC address of the physical port is used for PBB

encapsulation as BMAC SA, then upon the port failure, the PE MUST use the EVPN MAC route withdrawal message to signal the flush.

2) If the PE shared MAC address is used for PBB encapsulation as BMAC SA, then upon the port failure, the PE MUST re-advertise this MAC route with the MAC Mobility Extended Community to signal the flush.

The first method is recommended because it reduces the scope of flushing the most.

5.5. Fast Convergence in PBB-EVPN

As described above, when a large number of EVCs are aggregated via a physical port on a PE; where each EVC corresponds to a vES, then the

port failure impacts all the associated EVCs and their corresponding vES's. Two actions must be taken as the result of such port failure:

- Flushing of all CMACs associated with the BMAC of the failed port for the impacted I-SIDs

- DF election for all impacted vES's associated with the failed port

Section 5.4 describes how to flush CMAC address in the most optimum way - e.g., to flush least number of CMAC addresses for the impacted I-SIDs. This section describes how to perform DF election in the most optimum way - e.g., to trigger DF election for all impacted vES's (which can be in thousands) among the participating PEs via a single BGP message as opposed to sending thousands of BGP messages - one per vES.

In order to devise such fast convergence mechanism that can be triggered via a single BGP message, all vES's associated with a given physical port (e.g., ENNI) are colored with the same color representing that physical port. The MAC address of the physical port is used for this coloring purposes and when the PE advertises an ES route for a vES associated with that physical port, it advertises it with an EVPN Router's MAC Extended Community indicating the color of that port.

The receiving PEs take note of this color and for each such color, they create a list of vES's associated with this color (i.e., associated with this MAC address). Now, when a port failure occurs, the impacted PE needs to notify the other PEs of this color so that these PEs can identify all the impacted vES's associated with this color (from the above list) and re-execute DF election procedures for all the impacted vES's. This is done by withdrawing the BMAC address associated with the failed port.

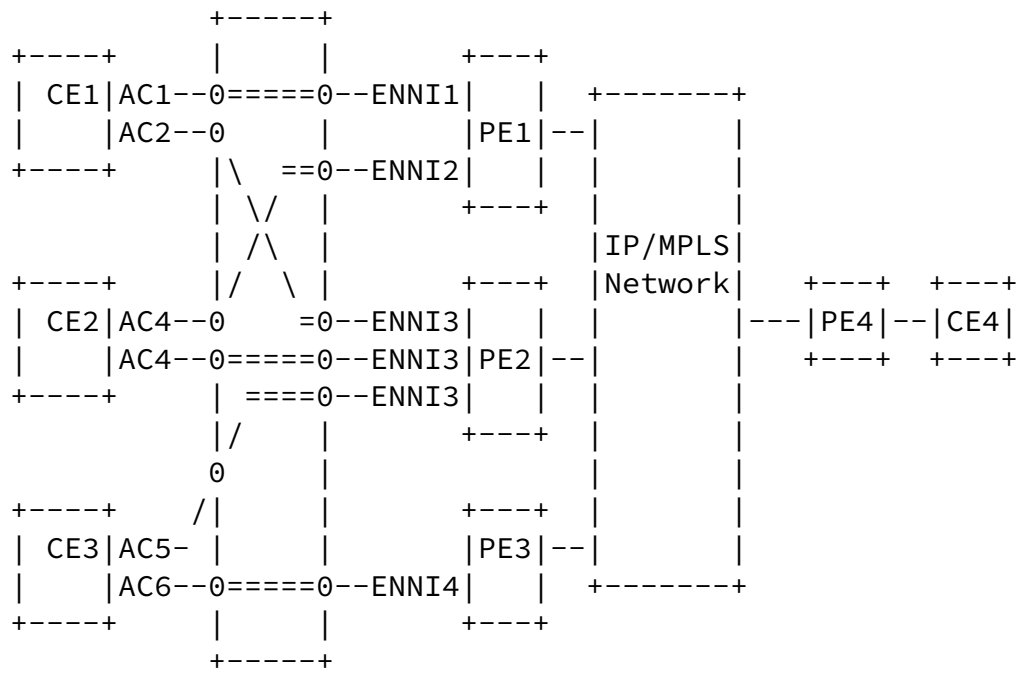


Figure 5: Fast Convergence Upon ENNI Failure

The following describes the procedure for coloring vES's and fast convergence using this color in more details:

1- When a vES is configured, the PE colors the vES with the MAC address of the corresponding physical port and advertises the Ethernet Segment route for this vES with this color.

2- All other PEs (in the redundancy group) take note of this color and add the vES to the list for this color.

3- Upon the occurrence of a port failure (e.g., an ENNI failure), the PE sends the flush message by withdrawing the BMAC address associated with the failed port. The PE should prioritize sending this flush message over ES route withdrawal messages of impacted vES's.

4- On reception of the flush message, other PEs use this info to flush their impacted CMACs and to initiate DF election procedures across all their affected vES's.

5- The PE with the physical port failure (ENNI failure), also sends ES route withdrawal for every impacted vES's. The other PEs upon receiving these messages, clear up their BGP tables. It should be noted the ES route withdrawal messages are not used for executing DF election procedures by the receiving PEs.

6. BGP Encoding

This document defines one new BGP Extended Community for EVPN.

6.1. I-SID Extended Community

A new EVPN BGP Extended Community called I-SID is introduced. This new extended community is a transitive extended community with the Type field of 0x06 (EVPN) and the Sub-Type of 0x07.

The I-SID Extended Community is encoded as an 8-octet value as follows:

0 1 2 3

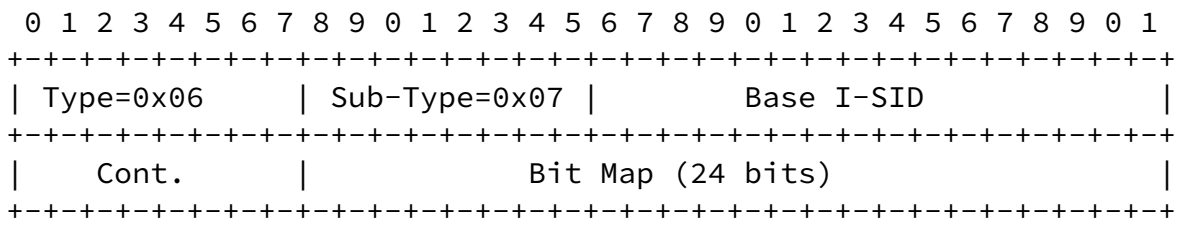


Figure 6: I-SID Extended Community

This extended community is used to indicate the list of I-SIDs associated with a given Ethernet Segment.

24-bit map represents the next 24 I-SID after the base I-SID. For example based I-SID of 10025 with 24-bit map of zero means, only a single I-SID of 10025. I-SID of 10025 with bit map of 0x000001 means there are two I-SIDs, 10025 and 10026.

7. Acknowledgements

The authors would like to thanks Mei Zhang and Jose Liste for their reviews and feedbacks of this document.

8. Security Considerations

All the security considerations in [RFC7432] and [RFC7623] apply directly to this document because this document leverages the control and data plane procedures described in those documents.

This document does not introduce any new security considerations beyond that of [RFC7432] and [RFC7623] because advertisements and processing of Ethernet Segment route for vES in this document follows that of physical ES in those RFCs.

9. IANA Considerations

IANA has allocated sub-type value 7 in the "EVPN Extended Community Sub-Types" registry defined in "https://www.iana.org/assignments/bgp-extended-communities/bgp-extended-communities.xhtml#evpn" as follows:

SUB-TYPE	NAME	Reference

0x07	I-SID Ext Comm	[draft-sajassi-bess-evpn-virtual-eth-segment]

It is requested from IANA to update the reference to this document.

10. Intellectual Property Considerations

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

11. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017.

[RFC7432] Sajassi, et al., "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015.

[RFC7623] Sajassi, et al., "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, September 2015.

[RFC8214] Boutrus, et al., "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, August 2017.

[EVPN-IRB] Sajassi, et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-05, July 2018.

12. Informative References

[RFC7209] Sajassi, et al., "Requirements for Ethernet VPN (EVPN)", RFC 7209, May 2014.

[RFC7080] Sajassi, A., Salam, S., Bitar, N., and F. Balus, "Virtual Private LAN Service (VPLS) Interoperability with Provider Backbone Bridges", RFC 7080, December 2013.

13. Authors' Addresses

Ali Sajassi
Cisco Systems

Email: sajassi@cisco.com

Patrice Brissette
Cisco Systems
Email: pbrisset@cisco.com

Rick Schell
Verizon
Email: richard.schell@verizon.com

John E Drake
Juniper
Email: jdrake@juniper.net

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