L2VPN Working Group

Pranjal Kumar Dutta
Florin Balus
Alcatel-Lucent

Internet Draft
Intended status: Standard
Expires: December 16, 2011

Olen Stokes
Extreme Networks

Geraldine Calvignac
France Telecom

June 23, 2011

LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS
draft-ietf-l2vpn-vpls-ldp-mac-opt-04.txt

Status of this Memo

Copyright Notice

Abstract

[RFC4762] describes a mechanism to remove or unlearn MAC addresses that have been dynamically learned in a VPLS Instance for faster convergence on topology change. The procedure also removes MAC addresses in the VPLS that do not require relearning due to such topology change.

This document defines an enhancement to the MAC Address Withdrawal procedure with empty MAC List [RFC4762], which enables a Provider Edge(PE) device to remove only the MAC addresses that need to be relearned.

Additional extensions to [RFC4762] MAC Withdrawal procedures are specified to provide optimized MAC flushing for the PBB-VPLS specified in [PBB-VPLS Model].

Table of Contents

1.1. Conventions used in this document

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.

This document uses the terminology defined in [PBB-VPLS Model], [RFC5036], [RFC4447] and [RFC4762]. Throughout this document VPLS means the emulated bridged LAN service offered to a customer. H-VPLS means the hierarchical connectivity or layout of MTU-s and PE devices offering the VPLS [RFC4762]. The terms spoke node and MTU-s in H-VPLS are used interchangeably.

2. Introduction

A method of Virtual Private LAN Service (VPLS), also known as Transparent LAN Service (TLS) is described in [RFC4762]. A VPLS is created using a collection of one or more point-to-point pseudowires (PWs) [RFC4664] configured in a flat, full-mesh topology. The mesh topology provides a LAN segment or broadcast domain that is fully capable of learning and forwarding Ethernet MAC addresses at the PE devices.

This VPLS full mesh core configuration can be augmented with additional non-meshed spoke nodes to provide a Hierarchical VPLS (H-VPLS) service [RFC4762]. Throughout this document this configuration is referred to as "regular" H-VPLS.

[PBB-VPLS Model] describes how Provider Backbone Bridging (PBB) can be integrated with VPLS to benefit from PBB capabilities while continuing to avoid the use of MSTP in the backbone. The combined solution, referred to as PBB-VPLS, results in better scalability in terms of number of service instances, PWs and customer MACs (CMACs) that need to be handled in the VPLS PEs.

A MAC Address Withdrawal mechanism for VPLS is described in [RFC4762] to remove or unlearn MAC addresses for faster convergence on a topology change in resilient H-VPLS topologies.

An example of the usage of the MAC Flush mechanism is a dual-homed H-VPLS where an edge device termed as MTU-s is connected to two PE devices via a primary spoke PW and a backup spoke PW, respectively. Such redundancy is designed to protect against the failure of the primary spoke PW or primary PE device.

When the MTU-s switches over to the backup PW, it is useful to flush the MAC addresses learned in the corresponding VSI in the peer PE devices participating in the full mesh to avoid black holing of frames to those addresses. Note that a forced switchover to the backup PW can be also performed at MTU-s administratively due to maintenance activities on the primary spoke PW. When the backup PW is made active by the MTU-s it triggers an LDP Address Withdraw Message with a list of MAC addresses to be flushed. The message is forwarded over the LDP session(s) associated with the newly activated PW. In order to minimize the impact on LDP convergence time and scalability when a MAC List TLV contains a large number of MAC addresses, many implementations use an LDP Address Withdraw Message with an empty MAC List. Throughout this document the term MAC Flush Message is used to specify an LDP Address Withdraw Message with an empty MAC List described in [RFC4762] unless specified otherwise.

As per the MAC Address Withdrawal processing rules in [RFC4762] a PE device on receiving a MAC flush message removes all MAC addresses associated with the specified VPLS instance (as indicated in the FEC TLV) except for the MAC addresses learned over the newly activated PW. The PE device further triggers a MAC flush message to each remote PE device connected to it in the VPLS full mesh.

This method of MAC flushing is modeled after Topology Change Notification (TCN) in Rapid Spanning Tree Protocol (RSTP)[802.1w]. When a bridge switches from a failed link to the backup link, the bridge sends out a TCN message over the newly activated link. The upstream bridge, upon receiving this message, flushes its entire MAC addresses except the ones received over this link and sends the TCN message out of its other ports in that spanning tree instance. The message is further relayed along the spanning tree by the other bridges. When a PE device in the full-mesh of H-VPLS receives a MAC flush message it also flushes MAC addresses which are not affected due to topology change, thus leading to unnecessary flooding and relearning. This document describes the problem and a solution to optimize the MAC flush procedure in [RFC4762] so it flushes the minimal set of MAC addresses that require relearning when the

topology changes in H-VPLS. The solution proposed in this document is generic and is applicable when MS-PWs are used in interconnecting PE devices in H-VPLS.

[PBB-VPLS Model] describes how PBB can be integrated with VPLS to benefit from PBB capabilities while continuing to avoid the use of MSTP in the backbone. The combined solution referred as PBB-VPLS results in better scalability in terms of number of service instances, PWs and CMACs that need to be handled in the VPLS PEs.

This document also describes extensions to LDP MAC Flush procedures described in [RFC4762] required to build desirable capabilities in a PBB-VPLS solution.

Section 3 covers the problem space. Section 4 describes the solution and the required TLV extensions.

3. Problem Description

3.1. MAC Flush optimization in VPLS resiliency

3.1.1. MAC Flush optimization for regular H-VPLS

Figure 1 describes a dual-homed H-VPLS scenario for a VPLS instance where the problem with the existing MAC flush method in [RFC4762] is explained.

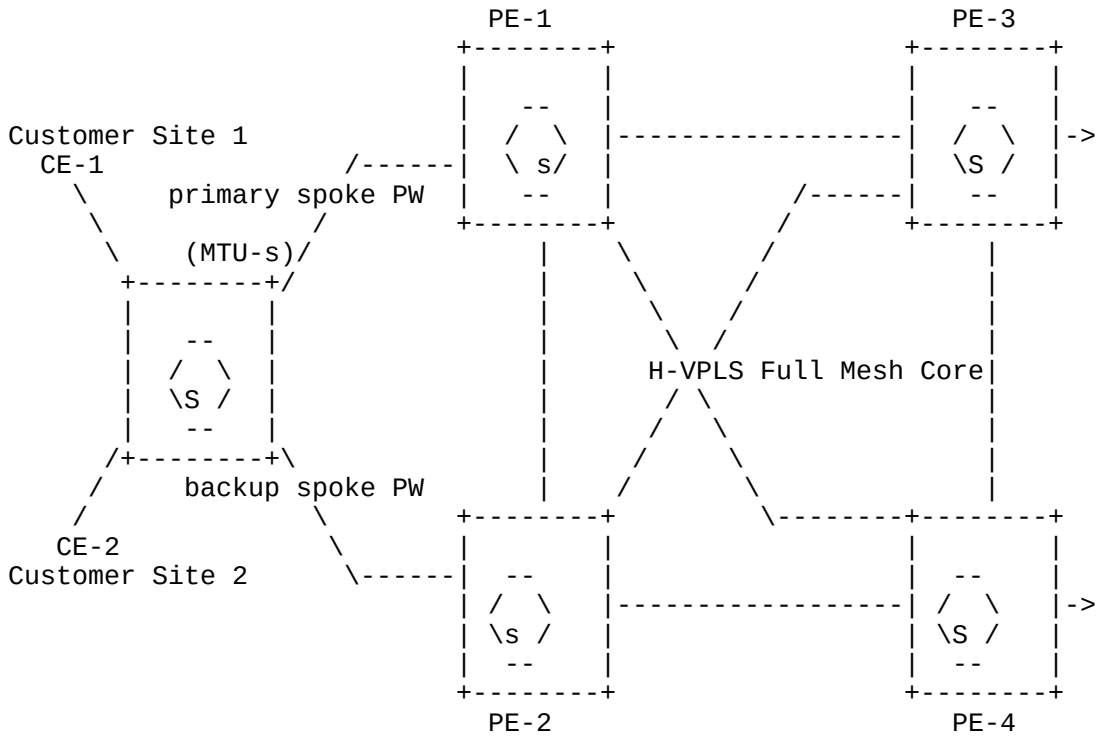


Figure 1: Dual homed MTU-s in two tier hierarchy H-VPLS

In Figure 1, the MTU-s is dual-homed to PE-1 and PE-2. Only the primary spoke PW is active at MTU-s, thus PE-1 is acting as the active device to reach the full mesh in the VPLS instance. The MAC addresses of nodes located at access sites (behind CE1 and CE2) are learned at PE-1 over the primary spoke PW. PE-2, PE-3 and PE-4 learn those MAC addresses on their respective mesh PWs terminating to PE-1.

When MTU-s switches to the backup spoke PW and activates it, PE-2 becomes the active device to reach the full mesh core. Traffic entering the H-VPLS from CE-1 and CE-2 is diverted by the MTU-s to the spoke PW to PE-2. To avoid traffic black holing the MAC addresses that have been learned in the upstream VPLS full-mesh through PE-1 must be relearned or removed from the MAC FIBs of PE-2, PE-3 and PE-4.

As per the processing rules defined in [RFC4762], on activation of the backup PW from MTU-s, a MAC flush message will be sent by MTU-s to PE-2 that will flush all the MAC addresses learned in the VPLS from all the other PWs except the PWs connected to MTU-s.

PE-2 further relays MAC flush messages to all other PE devices in the full mesh. Same processing rule applies at all those PE devices: all the MAC addresses are flushed except the ones learned on the PW to PE2. For example, at PE-3 all of the MAC addresses learned from the PWs connected to PE-1 and PE-4 are flushed and relearned subsequently. Before the relearning happens flooding of unknown destination MAC addresses takes place throughout the network. As the number of PE devices in the full-mesh increases, the number of unaffected MAC addresses flushed in a VPLS instance also increases, thus leading to unnecessary flooding and relearning. With a large number of VPLS instances provisioned in the H-VPLS network topology the amount of unnecessary flooding and relearning increases. An optimization is required that will flush only the MAC addresses learned from the PW connected to PE-1 to minimize the relearning and flooding in the network.

Further the forwarding of the MAC Flush by PE-2 delays the overall MAC flush propagation time into the core PEs in the full mesh. So it is desirable to avoid MAC flush forwarding across multiple PEs as far as possible and yet achieve the same desired MAC flushing action.

3.1.2. MAC Flush optimization for native Ethernet access

The analysis in section 3.1.1 applies also to the native Ethernet access into a VPLS where one active and one or more backup endpoints into two or more VPLS or H-VPLS PEs are being used. Examples of these are [G.8032v2] access rings or any proprietary multi-chassis LAG emulations.

As in the active/standby PWs case from the previous section, upon failure of the active native Ethernet endpoint on PE-1 a MAC Flush optimization is required to ensure that on PE-2, PE-3 and PE-4 only the MAC addresses learned from the PW connected to PE-1 are being flushed.

3.2. Black holing issue in PBB-VPLS

In a PBB-VPLS solution a B-component VPLS (B-VPLS) may be used as the infrastructure for one or more I-component instances. B-VPLS control plane (LDP Signaling) replaces the I-component control plane throughout the MPLS core. This raises an additional challenge related to black hole avoidance in the I-component domain as described in this section. Figure 2 describes the case of a CE device (node A) dual-homed to two I-component instances located on two PBB-VPLS PEs (PE1 and PE2).

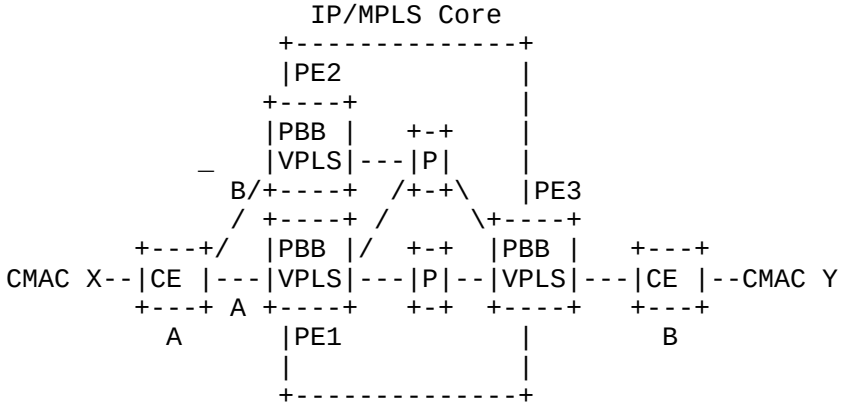


Figure 2: PBB Black holing Issue - CE Dual-Homing use case

The link between PE1 and CE A is active (marked with A) while the link between CE A and PE2 is in Backup/Blocked status. In the network diagram CMAC X is one of the MAC addresses located behind CE A in the customer domain, CMAC Y is behind CE B and the B-VPLS instances on PE1 and PE2 are associated with backbone MAC (BMAC) B1 and BMAC B2, respectively.

As the packets flow from CMAC X to CMAC Y through PE1 with BMAC B1, the remote PEs participating in the I-VPLS (for example, PE3) will

learn the CMAC X associated with BMAC B1 on PE1. Under failure of the link between CE A and PE1, and with the activation of link to PE2, the remote PEs (for example, PE3) will black hole the traffic destined to customer MAC X to BMAC B1 until the aging timer expires or a packet flows from X to Y through the PE2 with B2. This may take a long time (default aging timer is 5 minutes) and may affect a large number of flows across multiple I-components.

A possible solution to this issue is to use the existing LDP MAC Flush as specified in [RFC4762] to flush in the B-VPLS domain the BMAC associated with the PE where the failure occurred. This will automatically flush the CMAC to BMAC association in the remote PEs. This solution though has the disadvantage of producing a lot of unnecessary MAC flushes in the B-VPLS domain as there was no failure or topology change affecting the Backbone domain.

A better solution is required to propagate the I-component events through the backbone infrastructure (B-VPLS) in order to flush only the customer MAC to BMAC entries in the remote PBB-VPLS PEs. As there are no I-VPLS control plane exchanges across the PBB backbone, extensions to the B-VPLS control plane are required to propagate the I-component MAC Flush events across the B-VPLS.

4. Solution description

4.1. MAC Flush Optimization for VPLS resiliency

The basic principle of the optimized MAC flush mechanism is explained with reference to Figure 1.

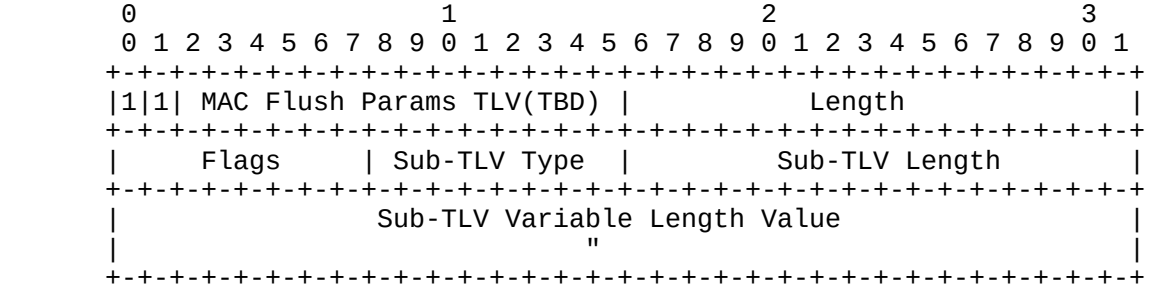
PE-1 would initiate a MAC Flush towards the core on detection of the failure of the primary spoke PW between MTU-s and PE-1 (or status change from active to standby). This method is referred to as a PE initiated MAC Flush throughout this document. The MAC Flush message would indicate to receiving PEs to flush all MACs learned over the PW in the context of the VPLS over which the MAC flush message is received. Each PE device in the full mesh that receives the message identifies the VPLS instance and its respective PW that terminates in PE-1 from the FEC TLV received in the message. Thus the PE device flushes only the MAC addresses learned from that PW connected to PE-1 minimizing the required relearning and the flooding throughout the VPLS domain.

This section defines a generic MAC Flush Parameters TLV for LDP [RFC5036]. Throughout this document the MAC Flush Parameters TLV is referred to as a MAC Flush TLV. A MAC Flush TLV carries information on the desired action at the PE device receiving the message and is

used for optimized MAC flushing in H-VPLS. The MAC Flush TLV is backward compatible and can be used for [RFC4762] style of MAC Flush as explained in section 3.1.

4.1.1. MAC Flush Parameters TLV format

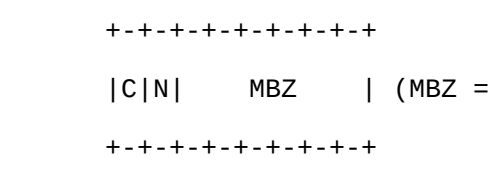
The MAC Flush Parameters TLV is described as below:



The U and F bits are set to forward if unknown so that potential intermediate VPLS PEs unaware of the new TLV can just propagate it transparently. The MAC Flush Parameters TLV type is to be assigned by IANA. The encoding of the TLV follows the standard LDP TLV encoding in [RFC5036].

The TLV value field contains a one byte Flag field used as described below. Further, the TLV value may carry one or more sub-TLVs. Any sub-TLV definition to the above TLV MUST address the actions in combination with other existing sub-TLVs.

The detailed format for the Flags bit vector is described below:



1 Byte Flag field is mandatory. The following flags are defined :

C flag, used to indicate the context of the PBB-VPLS component in which MAC flush is required. For PBB-VPLS there are two contexts of MAC flushing - The Backbone VPLS (B-component VPLS) and Customer

VPLS (I-component VPLS). C flag MUST be ZERO (C=0) when a MAC Flush for the B-VPLS is required. C flag MUST be set (C=1) when the MAC Flush for I-VPLS is required. In the regular H-VPLS case the C flag MUST be ZERO (C=0) to indicate the flush applies to the current VPLS context.

N flag, used to indicate whether a positive (N=0, Flush-all-but-mine) or negative (N=1 Flush-all-from-me) MAC Flush is required. The source (mine/me) is defined either as the PW associated with the LDP session on which the LDP MAC Withdraw was received or with the BMAC(s) listed in the BMAC List Sub-TLV. For the optimized MAC Flush procedure described in this section the flag must be set (N=1).

Detailed usage in the context of PBB-VPLS is explained in section 4.2.

MBZ flags, the rest of the flags MUST be set to zero on transmission and ignored on reception.

4.1.2. Application of MAC Flush TLV in Optimized MAC Flush

For the optimized MAC flush, the MAC Flush TLV MAY be sent as in the existing LDP Address Withdraw Message with an empty MAC List but from the core PE on detection of failure of its local spoke PW. The N bit in the TLV MUST be set to 1. If the optimized MAC Flush procedure is used in a Backbone VPLS or regular VPLS/H-VPLS context the C bit MUST be ZERO (C=0). If it is used in an I-VPLS context the C bit MUST be set (C= 1). See section 4.2 for PBB-VPLS details.

Note that if a MAC Flush TLV is not understood by a receiver then it may result in undesired action. For example if a MAC Flush Parameters TLV is received with N=1 and receiver does not understand that TLV then it would result in flushing of all MACs learned in the VSI except the ones learned over the PW. The MAC Flush TLV SHOULD be placed after the existing TLVs in MAC Flush message in [RFC4762].

For backward compatibility of MAC flush initiation procedures as defined in [RFC4762], the PE-1 MAY send a MAC Flush TLV as an OPTIONAL TLV in the MAC Flush Message with N = 0. This would result in same flushing action at the receiving PE devices as desired in [RFC4762].

4.1.3. MAC Flush TLV Processing Rules for regular H-VPLS

This section describes the processing rules of a MAC Flush TLV that SHOULD be followed in the context of MAC flush procedures in an H-VPLS.

For optimized MAC Flush a multi-homing PE initiates a MAC flush message towards the other related VPLS PEs when it detects a transition (failure or to standby) in an active spoke PW. In such a case the MAC Flush TLV MUST be sent with N = 1. A PE device receiving the MAC Flush TLV SHOULD follow the same processing rules as described in this section.

Note that if MS-PW is used in the VPLS then a MAC flush message is processed only at the T-PE nodes since the S-PE(s) traversed by the MS-PW propagate MAC flush messages without any action. In this section, a PE device signifies only T-PE in the MS-PW case unless specified otherwise.

When a PE device receives a MAC Flush TLV with N = 1, it SHOULD flush all the MAC addresses learned from the PW in the VPLS in the context on which the MAC Flush message is received.

If a MAC Flush TLV is received with N = 0 in the MAC flush message then the receiving PE SHOULD flush the MAC addresses learned from all PWs in the VPLS instance except the ones learned over the PW on which the message is received.

If a PE device receives a MAC flush with the MAC Flush TLV option and a valid MAC address list, it SHOULD ignore the option and deal with MAC addresses explicitly as per [RFC4762].

If a PE device that doesn’t support MAC Flush TLV receives a MAC flush message with this option, it MUST ignore the option and follow the processing rules as per [RFC4762]. However if the MAC Flush Parameters TLV was sent with N = 1 then it may result in wrong flushing action (Positive MAC Flush).

4.1.4. Optimized MAC Flush Procedures

This section explains the optimized MAC flush procedure in the scenario in Figure 1. When the primary spoke PW transition (failure or standby transition) is detected by PE-1, it may send MAC flush messages to PE-2, PE-3 and PE-4 with a MAC Flush TLV and N = 1. Upon receipt of the MAC flush message, PE-2 identifies the VPLS instance that requires the MAC flush from the FEC element in the FEC TLV. On receiving N=1, PE-2 removes all MAC addresses learned from that PW

over which the message is received. Same action is followed by PE-3 and PE-4.

Figure 3 shows another redundant H-VPLS topology to protect against failure of MTU-s device. Provider RSTP may be used as selection algorithm for active and backup PWs in order to maintain the connectivity between MTU devices and PE devices at the edge. It is assumed that PE devices can detect the failure of PWs in either direction through OAM mechanisms such as VCCV procedures for instance.

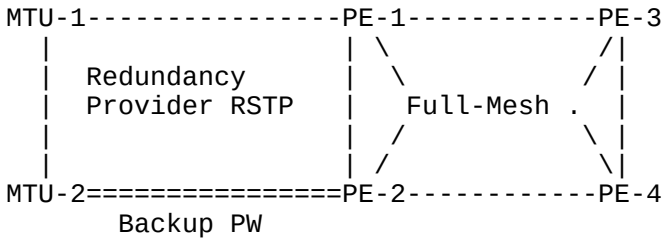


Figure 3: Redundancy with Provider RSTP

MTU-1, MTU-2, PE-1 and PE-2 participate in provider RSTP. By configuration in RSTP it is ensured that the PW between MTU-1 and PE-1 is active and the PW between MTU-2 and PE-2 is blocked (made backup) at the MTU-2 end. When the active PW failure is detected by RSTP, it activates the PW between MTU-2 and PE-2. When PE-1 detects the failing PW to MTU-1, it may trigger a MAC flush into the full mesh with a MAC Flush TLV that carries N=1. Other PE devices in the full mesh that receive the MAC flush message identify their respective PWs terminating on PE-1 and flush all the MAC addresses learned from it.

By default, MTU-2 should still trigger MAC flush as currently defined in [RFC4762] after the backup PW is made active by RSTP. Mechanisms to prevent two copies of MAC withdraws to be sent in such scenarios is out of scope of this document.

[RFC4762] describes multi-domain VPLS services where fully meshed VPLS networks (domains) are connected together by a single spoke PW per VPLS service between the VPLS "border" PE devices. To provide redundancy against failure of the inter-domain spoke, a full mesh of

inter-domain spokes can be setup between border PE devices and provider RSTP may be used for selection of the active inter-domain spoke. In case of an inter-domain spoke PW failure, PE initiated MAC withdrawal may be used for optimized MAC flushing within individual domains.

Further, the procedures are applicable with any native Ethernet access topologies multi-homed to two or more VPLS PEs. The text in section 4.1 applies for the native Ethernet case where active/standby PWs are replaced with the active/standby Ethernet endpoints. An optimized MAC Flush message can be generated by the VPLS-PE that detects the failure in the primary Ethernet access.

4.2. LDP MAC Withdraw Extensions for PBB-VPLS

The use of Address Withdraw messages with MAC List TLV is proposed in [RFC4762] as a way to expedite removal of MAC addresses as the result of a topology change (e.g. failure of a primary link of a VPLS PE and implicitly the activation of an alternate link in a dual-homing use case). These existing procedures apply individually to B-VPLS and I-component domains.

When it comes to reflecting topology changes in access networks connected to I-component across the B-VPLS domain certain additions should be considered as described below.

MAC Switching in PBB is based on the mapping of Customer MACs (CMACs) to Backbone MAC(s) (BMACs). A topology change in the access (I-domain) should just invoke the flushing of CMAC entries in PBB PEs' FIB(s) associated with the I-component(s) impacted by the failure. There is a need to indicate the PBB PE (BMAC source) that originated the MAC Flush message to selectively flush only the MACs that are affected.

These goals can be achieved by adding a new MAC Flush Parameters TLV in the LDP Address Withdraw message to indicate the particular domain(s) requiring MAC flush. On the other end, the receiving PEs may use the information from the new TLV to flush only the related FIB entry/entries in the I-component instance(s).

The following sub-TLVs MUST be included in the MAC Flush Parameters TLV if the C-flag is set to 1:

- PBB BMAC List sub-TLV:

Type: 0x01

Length: value length in octets. At least one BMAC address must be present in the list.

Value: one or a list of 48 bits BMAC addresses. These are the source BMAC addresses associated with the B-VPLS instance that originated the MAC Withdraw message. It will be used to identify the CMAC(s) mapped to the BMAC(s) listed in the sub-TLV.

- PBB ISID List sub-TLV:

Type: 0x02,

Length: value length in octets. Zero indicates an empty ISID list. An empty ISID list means that the flush applies to all the ISIDs mapped to the B-VPLS indicated by the FEC TLV.

Value: one or a list of 24 bits ISIDs that represent the I-component FIB(s) where the MAC Flush needs to take place.

4.2.1. MAC Flush TLV Processing Rules for PBB-VPLS

The following steps describe the details of the processing for the related LDP Address Withdraw message:

. The LDP MAC Withdraw Message, including the MAC Flush Parameters TLV is initiated by the PBB PE(s) experiencing a Topology Change event in one or multiple customer I-component(s).

    o The flags are set accordingly to indicate the type of MAC Flush required for this event: N=0 (Flush-all-but-mine) or N = 1 (Flush-all-from-me), C=1 (Flush only CMAC FIBs).

    o The PBB Sub-TLVs (BMAC and ISID Lists) are included according to the context of topology change.

. On reception of the LDP Address Withdrawal message, the B-VPLS instances corresponding to the FEC TLV in the message must interpret the content of MAC Flush Parameters TLV. If the C-bit is set to 1 then Backbone Core Bridges (BCB) in the PBB-VPLS SHOULD NOT flush their BMAC FIBs. The B-VPLS control plane SHOULD propagate the MAC Flush following the split-horizon grouping and the established B-VPLS topology.

. The usage and processing rules of MAC Flush Parameters TLV in the context of Backbone Edge Bridges (BEB) is as follows:

- o The PBB ISID List is used to determine the particular ISID FIBs (I-VPLS) that need to be flushed. If the ISID List is empty then all the ISID FIBs associated with the receiving B-VPLS SHOULD be flushed.

- o The PBB BMAC List is used to identify from the ISID FIBs in the previous step whether to selectively flush BMAC to CMAC associations depending on the N flag specified below.

. Next, depending on the N flag value the following actions apply:

- o N=0, all the CMACs in the selected ISID FIBs SHOULD be flushed with the exception of the identified CMAC list from the BMAC List mentioned in the message. ("Flush all but the CMACs associated with the BMAC(s) in the BMAC List Sub-TLV from the FIBs associated with the ISID list").

- o N=1, the identified CMAC list SHOULD be flushed ("Flush all the CMACs associated with the BMAC(s) in the BMAC List Sub-TLV from the FIBs associated with the ISID list").

4.2.3 Applicability of MAC Flush Parameters TLV

If a MAC Flush Parameters TLV is received by a BEB in a PBB-VPLS that does not understand the TLV then it may result in undesirable MAC flushing action. It is RECOMMENDED that all PE devices participating in PBB-VPLS support MAC Flush Parameters TLV.

The MAC Flush Parameters TLV is also applicable to regular VPLS contexts. To achieve negative MAC Flush (flush-all-from-me) in a regular VPLS context, the MAC Flush Parameters TLV SHOULD be encoded with C=0 and N = 1 without the inclusion of any Sub-TLVs. Negative MAC flush is highly desirable in scenarios when VPLS access redundancy is provided by Ethernet Ring Protection as specified in [G.8032v2] specification etc.

5. Security Considerations

Control plane aspects:

- LDP security (authentication) methods as described in [RFC5036] is applicable here. Further this document implements security considerations as in [RFC4447] and [RFC4762].

Data plane aspects:

- This specification does not have any impact on the VPLS forwarding plane.

6. IANA Considerations

The Type field in MAC Flush Parameters TLV is defined as 0x406 and is subject to IANA approval.

7. Acknowledgments

The authors would like to thank the following people who have provided valuable comments and feedback on the topics discussed in this document: Marc Lasserre, Ian Cowburn, Dimitri Papadimitriou, Jorge Rabadan, Prashanth Ishwar, Vipin Jain, John Rigby, Ali Sajassi, Wim Henderickx, Jorge Rabadan and Maarten Vissers.

8. References

8.1. Normative References

[RFC4762] Lasserre, M. and Kompella, V. (Editors), "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC5036] Andersson, L., et al. "LDP Specification", RFC5036, October 2007.

[RFC4447] Martini. and et al., "Pseudowire Setup and Maintenance Using Label Distribution Protocol (LDP)", RFC 4447, April 2006.

8.2. Informative References

[PBB-VPLS Model] F. Balus, et Al. "Extensions to VPLS PE model for Provider Backbone Bridging", draft-ietf-l2vpn-pbb-vpls-pe-model-00.txt, May 2009 (work in progress)

[RFC4664] Andersson, L., et al. "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

[802.1w] "IEEE Standard for Local and metropolitan area networks. Common specifications Part 3: Media Access Control (MAC) Bridges. Amendment 2: Rapid Reconfiguration", IEEE Std 802.1w-2001.

[G.8032v2] ITU-T G.8032v2 specification

Author's Addresses

Pranjal Kumar Dutta
Alcatel-Lucent
701 E Middlefield Road,
Mountain View, CA 94043
USA
Email: pranjal.dutta@alcatel-lucent.com

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

Geraldine Calvignac
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex
France
Email: geraldine.calvignac@orange-ftgroup.com

Olen Stokes
Extreme Networks
PO Box 14129
RTP, NC 27709
USA
Email: ostokes@extremenetworks.com

Internet Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2012

Y. Jiang
L. Yong
Huawei
M. Paul
Deutsche Telekom
F. Jounay
France Telecom Orange
July 11, 2011

VPLS PE Model for E-Tree Support
draft-jiang-l2vpn-vpls-pe-etree-04.txt

Abstract

A generic VPLS solution for E-Tree services is proposed which uses VLANs to indicate root/leaf traffic. A VPLS Provider Edge (PE) model is illustrated as an example for the solution. In the solution, E-Tree VPLS PEs are interconnected by tagged PWs, the MAC address based Ethernet forwarding engine and the PW works in the same way as before. A signaling mechanism for E-Tree capability and VLAN mapping negotiation is further described.

Table of Contents

1. Introduction

The E-Tree service is defined in Metro Ethernet Forum (MEF) as a Rooted-Multipoint EVC service. It is a multipoint Ethernet service with special restrictions: the frames from a root may be received by any other root or leaf, and the frames from a leaf may be received by any root, but MUST not be received by a leaf. Further, an E-Tree service may include multiple roots and multiple leaves. Although VPMS

or P2MP multicast is a somewhat simplified version of this service, in fact, there is no exact corresponding terminology in IETF.

[Etree-req] gives the requirements for providing E-Tree solutions in the VPLS and the need to filter leaf to leaf traffic.

[vpls-etree] describes a PW control word based E-Tree solution, where a bit in the PW control word is used to indicate the root/leaf attribute for a packet. The Ethernet forwarder in the VPLS is also extended to filter the leaf-leaf traffic based on the <ingress port, egress port, CW L-bit> tuple.

[Etree-2PW] proposes another E-Tree solution where root and leaf traffic are classified and forwarded in the same VSI but with two separate PWs.

Both solutions are only applicable to "VPLS only" networks.

In fact, VPLS PE usually consists of a bridge module itself (see [RFC4664] and [RFC6246]), moreover, E-Tree services may cross both Ethernet and VPLS domains. Therefore, it is necessary to develop an E-Tree solution both for "VPLS only" scenarios and for interworking between Ethernet and VPLS.

IEEE 802.1 has incorporated the generic E-Tree solution in the latest version of 802.1Q [802.1aq], which is just an improvement on the traditional asymmetric VLAN mechanism. In the solution, VLANs are used to indicate root/leaf attribute of a packet: one VLAN ID is used to indicate the frames originated from the roots and another VLAN ID is used to indicate the frames originated from the leaves. At a leaf port, the bridge can then filter out all the frames from other leaf ports based on the VLAN ID. It is better to reuse the same mechanism in VPLS than to develop a new mechanism. The latter will introduce more complexity to interwork with IEEE 802.1Q solution.

This document introduces how the Ethernet VLAN solution can be used to support generic E-Tree services in the VPLS. The solution proposed here is fully compatible with the IEEE bridge architecture and the IETF PWE3 technology, and VPLS scalability and simplicity is also well kept. With this mechanism, it is also convenient to deploy a converged E-Tree service across both Ethernet and MPLS networks.

Firstly, a typical VPLS PE model is introduced as an example, the model is extended in which a Tree VSI is connected to a VLAN bridge with a dual-VLAN interface.

This document then discusses the PW encapsulation and PW processing such as VLAN mapping options for transporting E-Tree services in a VPLS.

Finally, it describes the signaling extensions for E-Tree support and PE processing procedures.

2. Conventions used in this document

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 [RFC2119].

3. Terminology

E-Tree: a Rooted-Multipoint EVC service according to the definition in MEF

EVC: Ethernet Virtual Connection, as defined in MEF 4.0

T-VSI: Tree VSI, a VSI with E-Tree support

Root AC, an AC attached with a root

Leaf AC, an AC attached with a leaf

Root VLAN, a VLAN ID used to indicate all the frames that are originated at a root AC

Leaf VLAN, a VLAN ID used to indicate all the frames that are originated at a leaf AC

4. PE Model with E-Tree Support

"VPLS only" PE architecture as outlined in Fig. 1 of [Etree-req] is a simplification of the VPLS and PWE3 architecture, several common VPLS PE architectures are discussed in more details in [RFC4664] and [RFC6246].

Therefore, VLAN based E-Tree solution are demonstrated with the help of a typical VPLS PE model. Other PE models are further discussed in Appendix A.

4.1. Existing PE Models

According to [RFC4664], there are at least three models possible for a VPLS PE, including:

o A single bridge module, a single VSI;

o A single bridge module, multiple VSIs;

o Multiple bridge modules, each attaches to a VSI.

The second PE model is commonly used. A typical example is further depicted in Fig. 1 and Fig. 2 [RFC6246], where an S-VLAN bridge module is connected to multiple VSIs each with a single VLAN virtual interface.

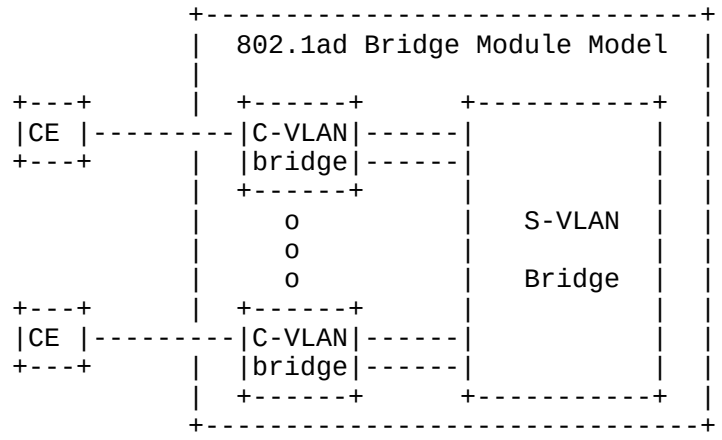


Figure 1 The Model of 802.1ad Bridge Module

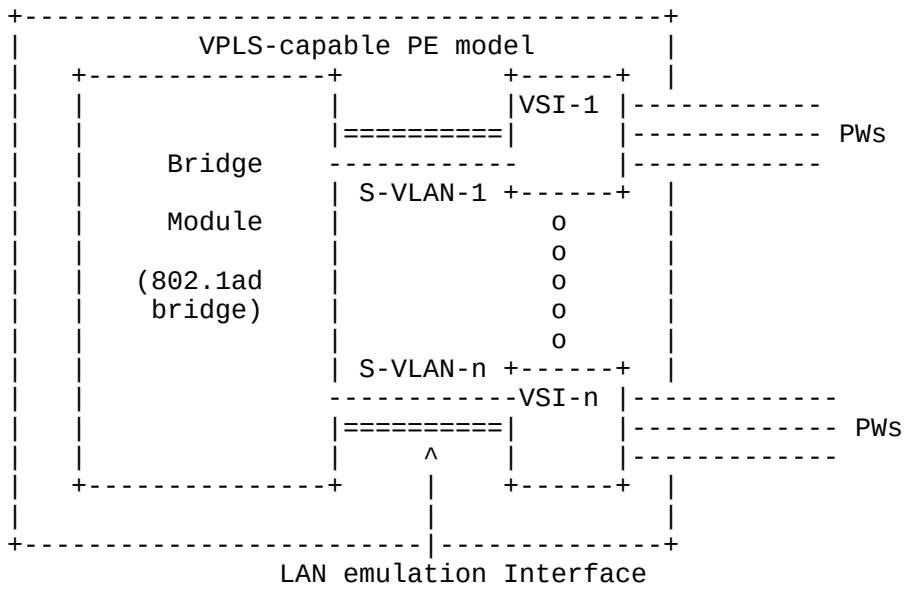


Figure 2 VPLS-capable PE Model

In this PE model, Ethernet frames from Customer Edges (CEs) will cross multiple stages of bridge modules (i.e., C-VLAN and S-VLAN bridge) and a VSI in a PE before being sent on the PW to a remote PE. Therefore, the association between an AC port and a PW on a VSI as required in [vpls-etree] or [Etree-2PW] is difficult, sometimes even impossible.

This model could be further enhanced: When Ethernet frames arrive at a PE, a root VLAN or a leaf VLAN tag is added. Then the frames with the root VLAN tag are transmitted both on the roots and the leaves, while the frames with the leaf VLAN tag are transmitted on the roots but dropped on the leaves (these VLAN tags are removed before the frames are transmitted over the wire). It was demonstrated in [802.1aq] that the E-Tree service in Ethernet networks can be well supported with this mechanism.

Assuming this mechanism is implemented in the bridge module, it is quite straightforward to infer a VPLS PE model with two VSIs to support the E-Tree (as shown in Fig. 3). But this model will require two VSIs per PE and two sets of PWs per E-Tree service, which is poorly scalable in a large MPLS/VPLS network; in addition, both these VSIs have to share their learned MAC addresses.

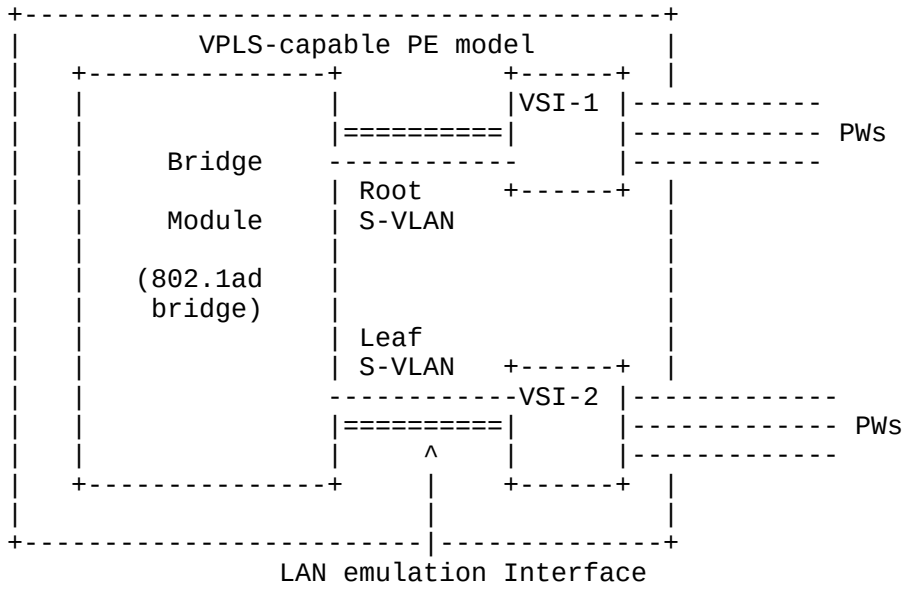


Figure 3 VPLS PE Model for E-Tree with 2 VSIs

4.2. A New PE Model with E-Tree Support

In order to support the E-Tree in a more scalable way, a new VPLS PE model with a single Tree VSI (T-VSI, a VSI with E-Tree support) is proposed. As depicted in Fig. 4, the bridge module is connected to the T-VSI with a dual-VLAN virtual interface, i.e., both the root VLAN and the leaf VLAN are connected to the same T-VSI, and they share the same FIB and work in shared VLAN learning. In this way, only one VPLS instance and one set of PWs is needed per E-Tree service, and the scalability of VPLS is improved.

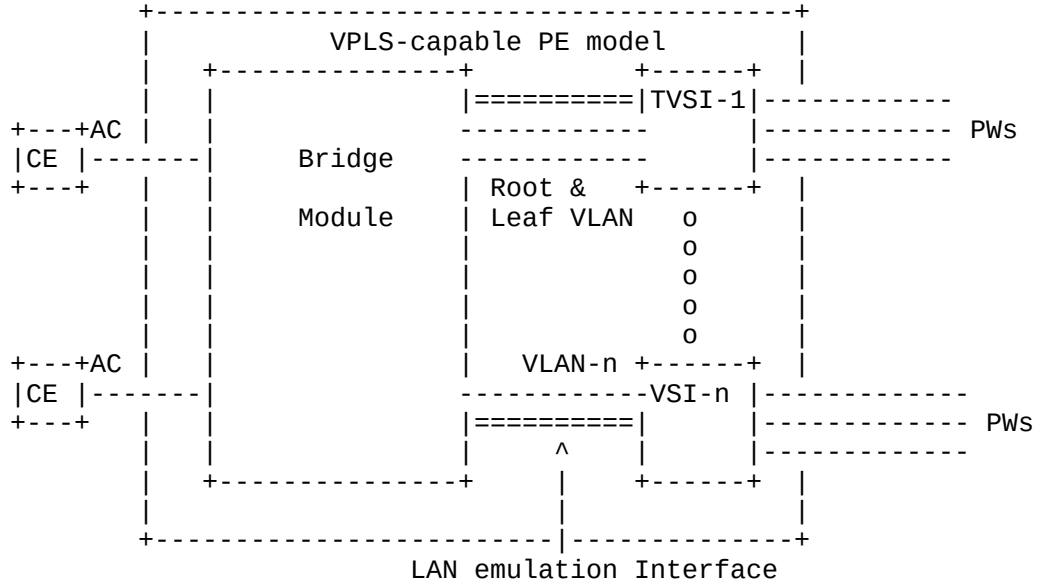


Figure 4 VPLS PE Model for E-Tree with a Single T-VSI

For an untagged port (customer sites attached to the PEs with untagged ports), the Ethernet frames received from the root ACs can be firstly tagged with a root C-VLAN, and then added with another root S-VLAN. Optionally, the frames from the root ACs can be tagged with the root S-VLAN tag directly in the VPLS network domain.

For a C-VLAN tagged port, the Ethernet frames received from the root ACs can be added with a root S-VLAN, or the C-VLAN can be translated to the root S-VLAN in the VPLS network domain.

For an S-VLAN tagged port, the S-VLAN tag in the Ethernet frames received from the root ACs can be translated to the root S-VLAN in the VPLS network domain.

In a similar way, the traffic from the leaf ACs is tagged and transported on the leaf S-VLAN.

This document will use the VLAN in its more general meaning in the latter sections.

5. PW for E-Tree Support

5.1. PW Encapsulation

For a VPLS instance to support an E-Tree service, its Ethernet PW should work in the tagged mode (PW type 0x0004) as described in [RFC4448], and a VLAN tag must be carried in each frame in the PW to indicate the E-Tree root/leaf attribute.

Raw PW may also be used to carry E-Tree service, as the VLAN indicating the E-Tree root/leaf attribute can be translated by the bridge module or by another Ethernet edge device.

A pair of T-VSIs in a VPLS is interconnected with a bidirectional PW. The VLAN indicating root/leaf attribute of the frame is carried in the PW, and the peer PE must drop all the frames with a leaf VLAN on each egress port associated with a leaf.

5.2. VLAN Mapping

There are two ways of manipulating VLANs for an E-Tree in VPLS:

- Global VLAN based that is, provisioning two global VLANs (Root VLAN, Leaf VLAN) across the VPLS network, thus no VLAN mapping is needed at all, or the VLAN mapping is done completely in the Ethernet domains.

- Local VLAN based, that is, provisioning two local VLANs for each PE (which participates in the E-Tree) in the VPLS network independently.

The first method requires no VLAN mapping in the PW, but two unique VLANs must be allocated in the VPLS (they may be provisioned by management or signaled by some control protocols), and the PW processing procedure as described in [RFC4448] applies.

The second method is more scalable in the use of VLANs, but needs a VLAN mapping mechanism in the PW similar to what is already described in Section 4.3 of [RFC4448]. It is assumed that for each PE with E-Tree capability there is a VLAN mapping module that can be enabled when VLAN mapping is needed for a PW. Actual VLAN mapping mode can be provisioned or determined by a signaling protocol as described in Section 6 when PW is being established.

5.3. PW Processing

5.3.1.PW Processing in the VLAN Mapping Mode

In the VLAN Mapping mode, two VPLS PEs with E-Tree capability are inter-connected with a PW (For example, the scenario of Fig. 5 depicts the interconnection of two PEs miscellaneously attached with roots and leaves).

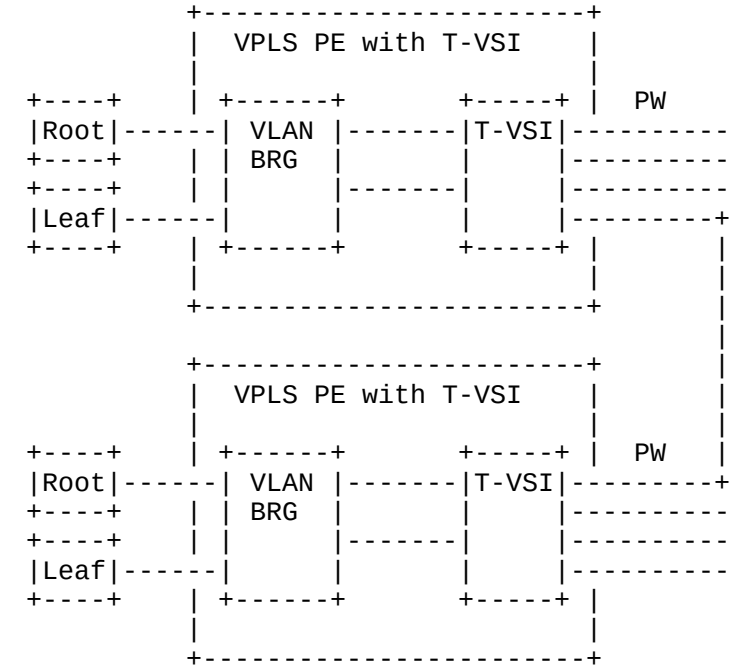


Figure 5 T-VSI Interconnected in the Normal Mode

If a PE is in the VLAN mapping mode for a PW, then in the data plane the PE MUST map the VLAN in each frame as follows:

o Upon transmitting frames on the PW, map from local VLAN to remote VLAN (i.e., the local leaf VLAN in a frame is translated to the remote leaf VLAN; the local root VLAN in a frame is translated to the remote root VLAN).

o Upon receiving frames on the PW, map from remote VLAN to local VLAN, and the frames are further forwarded or dropped in the egress bridge module using the filtering mechanism as described in [802.1aq].

5.3.2.PW Processing in the Compatible Mode

The new VPLS PE model can work in a traditional VPLS network seamlessly in the compatibility mode. As shown in Fig. 6, the VPLS PE with T-VSI can be attached with root and/or leaf nodes, while the VPLS PE with a traditional VSI can only be attached with root nodes.

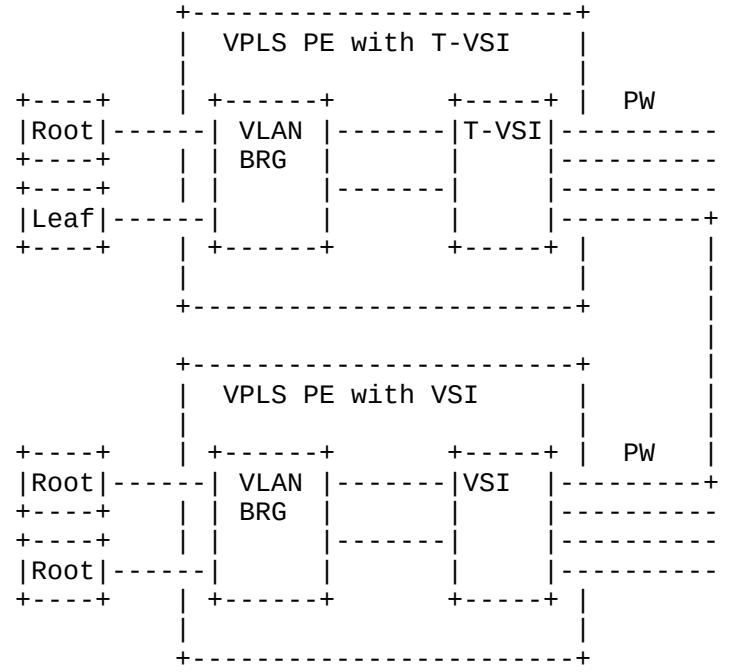


Figure 6 T-VSI interconnected with Traditional VSI

If a PE is in the Compatible mode for a PW, then in the data plane the PE MUST map the VLAN in each frame as follows:

o Upon transmitting frames on the PW, map both local root and local leaf VLAN to the remote VLAN.

o Upon receiving frames on the PW, map the remote VLAN to the local root VLAN.

5.3.3.PW Processing in the Optimized Mode

When two PEs are connected with their T-VSIs and one PE (e.g., PE2) is attached with only leaves, as shown in the scenario of Fig. 6, the peer PE (e.g., PE1) should then work in the optimization mode. In this case, PE1 should not send the frames originated from the local leaf VLAN to PE2, i.e., these frames are dropped rather than transported over the PW. The bandwidth efficiency of the VPLS can thus be improved. The signaling for the PE attached with only leaves is specified in Section 6.

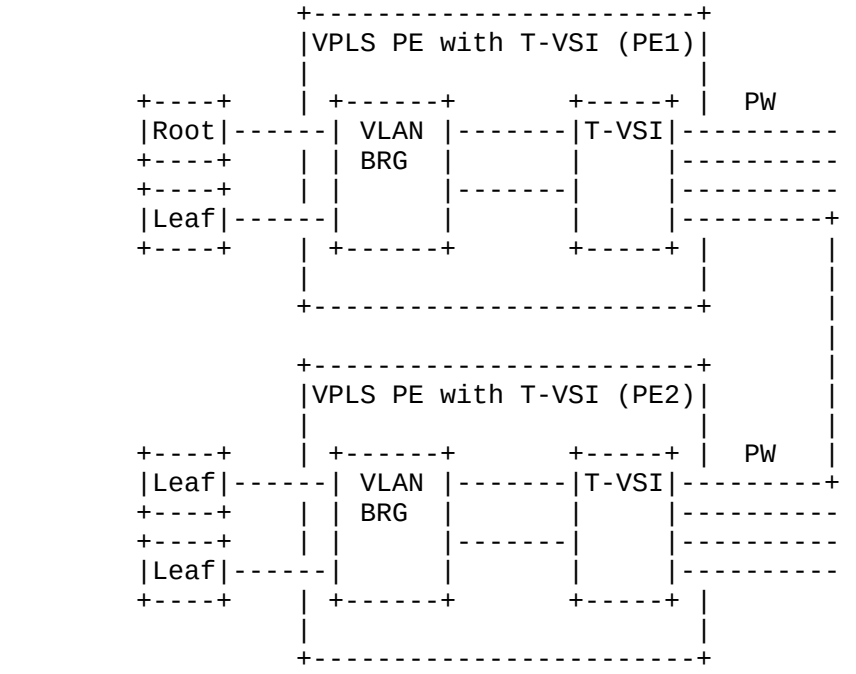


Figure 7 T-VSI interconnected with one side attached with only leaves

If a PE is in the Optimized Mode for a PW, then in the data plane, before proceeding as listed in Section 5.3.1 upon transmit, the PE SHOULD first operate as follows:

o Drop a frame if its VLAN ID matches the local leaf VLAN ID.

6. LDP Extensions for E-Tree Support

In addition to the signaling procedures as specified in [RFC4447], this document proposes a new interface parameter sub-TLV to provision an E-Tree service and negotiate the VLAN mapping function, as follows:

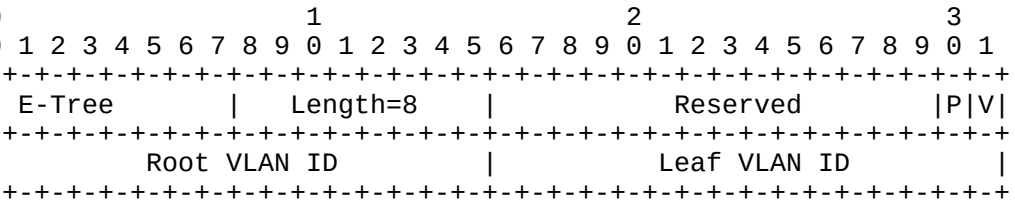


Figure 8 E-Tree Sub-TLV

Where:

- o E-Tree is the sub-TLV identifier to be assigned by IANA.
- o Length is the length of the sub TLV in octets.
- o Reserved bits MUST be set to zero on transmit and be ignored on receive.
- o P is a Leaf-only bit, it is set to 1 to indicate that the PE is attached with only leaves, and set to 0 otherwise.
- o V is a bit indicating the sender's VLAN mapping capability. A PE capable of VLAN mapping MUST set this bit, and clear it otherwise.
- o Root VLAN ID is the value of the local root VLAN.
- o Leaf VLAN ID is the value of the local leaf VLAN.

When setting up a PW for the E-Tree based VPLS, two PEs negotiate the E-Tree support using the above E-Tree sub-TLV. Note PW type of 0x0004 should be used during the PW negotiation.

A PE that wishes to support E-Tree service MUST include an E-Tree Sub-TLV in its PW label mapping message and include its local root VLAN ID and leaf VLAN ID in the TLV. A PE that has the VLAN mapping capability MUST set the V bit to 1, and a PE is attached with only leaves SHOULD set the P bit to 1.

In default, for each PW, VLAN-Mapping-Mode, Compatible-Mode, and Optimized-Mode are all set to FALSE.

A PE that receives a PW label mapping message with an E-Tree Sub-TLV from its peer PE must process it as follows:

1) if the root and leaf VLAN ID in the message match the local root and leaf VLAN ID, then continue to 3);

2) else {

if the bit V is cleared, then {

if the PE is capable of VLAN mapping, then it MUST set VLAN-Mapping-Mode to TRUE;

else {

A label release message with the error code "E-Tree VLAN mapping not supported" is sent to the peer PE and exit the process;

}

}

if the bit V is set, and the PE is capable of VLAN mapping, then the PE with the minimum IP address MUST set VLAN-Mapping-Mode to TRUE;

}

3) If the P bit is set, then:

{

If the PE is a leaf-only node itself, then a label release message with the error code "Leaf to Leaf PW error" is sent to the peer PE and exit the process;

Else the PE SHOULD set the Optimized-Mode to TRUE.

}

If a PE has sent an E-Tree Sub-TLV but does not receive any E-Tree Sub-TLV in its peer’s PW label mapping message, then set Compatible-

Mode to TRUE if the PE is VLAN mapping capable, otherwise a label release message is sent and an error is logged.

Data plane processing for this PW is as following:

If Optimized-Mode is TRUE, then data plane processing is as described in Section 5.3.3.

Else if Compatible-Mode is TRUE, then data plane processing is as described in Section 5.3.2.

Else if VLAN-Mapping-Mode is TRUE, then data plane processing is as described in Section 5.3.1.

PW processing as described in [RFC4448] proceeds as usual.

7. BGP Extensions for E-Tree Support

BGP may also be used to distribute the E-Tree and VLAN mapping information. It is to be specified in the next version.

8. Applicability

The solution is applicable to LDP VPLS [RFC4762] and may also be applicable to BGP VPLS [RFC4761].

The solution is applicable to both "VPLS Only" network and VPLS with Ethernet aggregation network.

9. Security Considerations

To be added in the future version.

10. IANA Considerations

IANA is requested to allocate a value for E-Tree in the Pseudowire Interface Parameters Sub-TLV type registry.

Parameter ID	Length	Description
TBD	8	E-Tree

IANA is requested to allocate a new LDP status code from the registry of name "STATUS CODE NAME SPACE". The following value is suggested:

Range/Value	E	Description
TBD	0	E-Tree VLAN mapping not supported

11. References

11.1. Normative References

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

[RFC4447] Martini, L., and et al, "Pseudowire Setup and Maintenance Using Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4448] Martini, L., and et al, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[RFC4762] Lasserre, M. and Kompella, V., "Virtual Private LAN Services using LDP", RFC 4762, January 2007.

11.2. Informative References

[RFC3985] Bryant, S., and Pate, P., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4664] Andersson, L., and Rosen, E., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

[RFC6246] Sajassi, A., and et al, "Virtual Private LAN Service (VPLS) Interoperability with Customer Edge (CE) Bridges", RFC 6246, June 2011

[ETree-req] Key, R., et al, "Requirements for MEF E-Tree Support in VPLS", draft-key-l2vpn-vpls-etree-reqt-02, October 2010

[vpls-etree] Key, R., and et al, "Extension to VPLS for E-Tree", draft-key-l2vpn-vpls-etree-04, October 2010

[802.1aq] IEEE 802.1aq D3.6, Virtual Bridged Local Area Networks - Amendment 9: Shortest Path Bridging, February 2011

[Etree-2PW] Ram, R., and et al., Extension to LDP-VPLS for E-Tree Using Two PW, draft-ram-l2vpn-ldp-vpls-etree-2pw-00.txt, October 2010

12. Acknowledgments

The authors would like to thank Adrian Farrel and Susan Hares for their valuable comments and advices.

Appendix A. Other PE Models for E-Tree

A.1. PE Model With a VSI and No bridge

If there is no bridge module in a PE, the PE may consist of Native Service Processors (NSPs) as shown in Figure A.1 (adapted from Fig. 5 of [RFC3985]) where any transformation operation for VLANs (e.g., VLAN insertion/removal or VLAN mapping) may be applied. Thus a root VLAN or leaf VLAN can be added by the NSP depending on the UNI type (root/leaf) associated with the AC over which the packet arrives.

Further, when a packet with a leaf VLAN exits a forwarder and arrives at the NSP, the NSP must drop the packet if the egress AC is associated with a leaf UNI.

Tagged PW and VLAN mapping work in the same way as in the typical PE model.

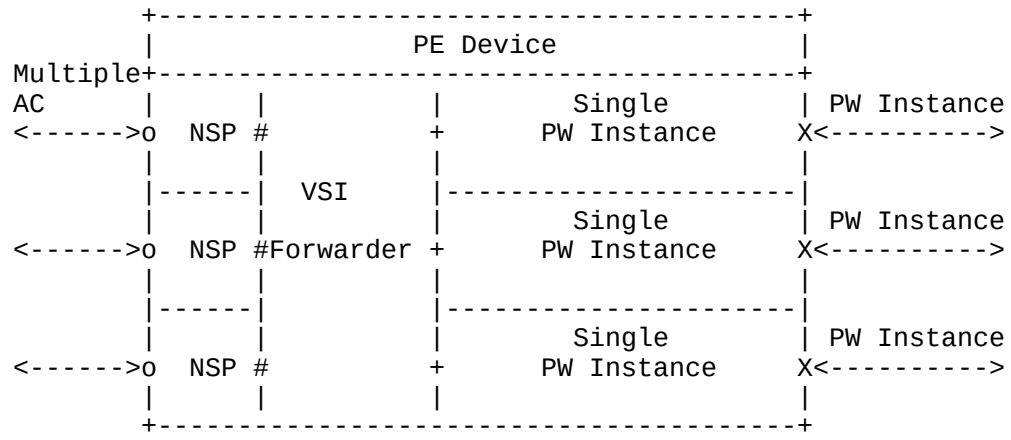


Figure A.1 PE model with a VSI and no bridge module

Authors' Addresses

Yuanlong Jiang
Huawei Technologies Co., Ltd.
Bantian, Longgang district
Shenzhen 518129, China
Email: jiangyuanlong@huawei.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075, USA
Email: lucyyong@huawei.com

Manuel Paul
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin, Germany
Email: manuel.paul@telekom.de

Frederic Jounay
France Telecom Orange
2, avenue Pierre-Marzin
22307 Lannion Cedex, France
Email: frederic.jounay@orange-ftgroup.com

Networking Working Group
Internet-Draft
Intended status: Informational
Expires: January 1, 2012

Z. Liu
China Telecom
L. Jin
R. Chen
ZTE

June 30, 2011

Node redundancy provisioning for VPLS Inter-domain draft-liu-l2vpn-vpls-inter-domain-redundancy-00

Abstract

In many VPLS deployment based on [RFC4762], inter-domain has been deployed without node redundancy, or only with node redundancy in one domain. This document describes how to deploy inter-domain VPLS based on [RFC4762] with node redundancy in both domain. The draft reuses the existing protocols without introducing any new protocols.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 1, 2012.

Copyright Notice

Table of Contents

1. Introduction

In many VPLS deployment based on [RFC 4762], inter-domain has been deployed without node redundancy, or only with node redundancy in one domain. This document describes how to deploy inter-domain VPLS based on [RFC 4762] with node redundancy in both domain. The draft reuses the existing protocols without introducing any new protocols. The domain in this document refers to AS, or other administrative domain.

2. Conventions used in this document

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 RFC2119.

3. Motivation

Inter-AS VPLS has now been wildly deployed between two providers. Usually, the physical link and ASBR between the two providers would carry many kinds of service, then it is important to provider link and node redundancy for such kind of inter-AS service to ensure high availability.

Some current high availability deployments of inter-AS VPLS are provided by MC-LAG (Multi-Chassis Link Aggregation) and [I-D.ietf-pwe3-iccp], but there is a pre-condition that the interconnected link between the two providers are Ethernet link. There are also many interconnection cases between two providers to use POS (Packet over Sonet/SDH) link on which MC-LAG cannot be enabled. Moreover, it is also required for the VPLS between two providers to ensure bandwidth control, QoS, MAC address control and Broadcast/Multicast traffic control. Then from the technical point of view, it is necessary to use PW to interconnect the two VPLS in its corresponding providers, and also to provide link/node redundancy to ensure high availability.

4. Redundancy scenario with ICCP

The following figure presents a typical inter-AS VPLS deployment topology. PE3 and PE4 are the VPLS edge nodes in network of operator A, and PE5 and PE6 are the VPLS edge nodes in network of operator B. The PE3/PE4/PE5/PE6 may be ASBR of the AS, or VPLS PE within its own AS.

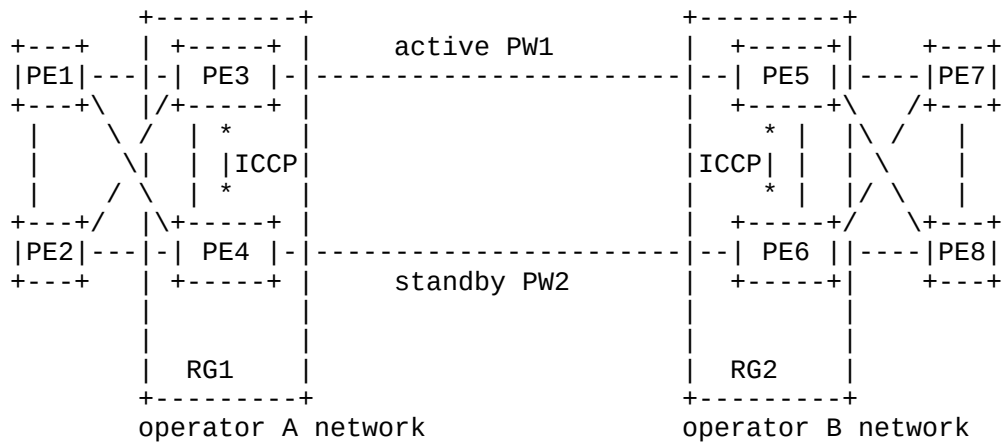


Figure 1

When inter-AS VPLS is deployed with node redundancy on both AS side, node redundancy protocol ICCP[I-D.ietf-pwe3-iccp]SHOULD be implemented on the VPLS edge nodes of the AS, e.g, ICCP should be running between PE3 and PE4, PE5 and PE6.

There are several deployment scenarios for inter-domain VPLS:

- ICCP deployment option: ICCP is deployed on VPLS edge nodes in one domain, or in both domain;
- PW redundancy mode: independent or master/slave;

From the operator’s point of view, it is important to keep the technical balance and technical independence between the two operators. One operator will not highly rely on the other operator’s technical choice for inter-domain VPLS node redundancy. Then it is highly recommended to be the deployment scenario as follows:

- ICCP deployment option: ICCP is deployed on VPLS edge nodes in both domain;
- PW redundancy mode: independent only;

And this draft will only focus on the above deployment option, other options are out of the scope.

5. Node redundancy for VPLS Inter-domain

The PEs in the RG are required to run an inter-chassis communication protocol ([I-D.ietf-pwe3-iccp]) in order to select which pseudowire(s) should be in active/standby state for a given VPLS service instance.

The procedures to select active/standby pseudowire(s):

- The PEs in the RG enable ICCP[I-D.ietf-pwe3-iccp].
- The PEs should establish a PW-RED application connection using the mechanism described in [I-D.ietf-pwe3-iccp], section 9.1.1.
- When the PW-RED application connection first comes up, Each PE MUST advertise it local PW configuration to other PEs that are members of the same RG. As part of the configuration information, the PE should advertise a PW priority value that is used to determine the precedence of a given pseudowire.
- Pseudowire Status Synchronization. A PE MAY re-advertise its PW-RED state in an unsolicited/solicited manner, the detailed mechanism is described in [I-D.ietf-pwe3-iccp], section 9.1.3.

The PEs SHOULD then use PW redundancy bit [I-D.ietf-pwe3-redundancy-bit] or basic PW status bit [RFC4447] to advertise the outcome of the arbitration to the peer PE(s).

Before deploying inter-domain VPLS, the operator MUST negotiate to configure same PW priority at two end-points. If different PW priority value is configured at the two PW end-points, e.g, PE3 and PE5 for PW1, and PE4 and PE6 for PW2 in figure 1, it is possible to select PE3 and PE6 as active for the two domain, then both PW1 and PW2 will be standby according to the independent mode in [I-D.ietf-pwe3-redundancy-bit].

6. MAC Withdraw procedure

It MAY be desirable to remove or unlearn MAC addresses that have been dynamically learned for faster convergence. This is accomplished by sending an LDP Address Withdraw Message. It not only obey the rule of MAC withdraw mechanism as described in [RFC4762], but also obey the rule that the range of MAC Address Withdraw is confined in the same domain for security reason. That means PE SHOULD not advertise MAC Address Withdraw message from one domain to the other. If not, the MAC address withdraw message with empty list originated from one domain would lead to a VPLS in another domain to flush all MAC addresses which is not necessary, and bring potential network instability. Correspondingly, VPLS PE that connects another domain SHOULD also reject any MAC Address Withdraw message received from that domain.

In figure 1, we assume PE3 node failure. The following case will describe MAC Address Withdraw of VPLS Inter-AS in detail. Upon detecting that a remote PE3 that is member of the same RG, has gone downGBP[not]PE4 will send Address Withdraw Message to other PEs that in the same AS(e.g.,PE2,PE3,and PE1), but will not send to PE6. The receiver PEs nodes flush the MAC addresses associated with

corresponding VPLS instance. Upon detecting remote-end PE3 failure, PE5 will send Address Withdraw Message to other PEs that in the same AS(e.g., PE6, PE7, and PE8). The receiver PEs flush the MAC addresses associated with corresponding VPLS instance. When PE6 receives this Address Withdraw Message, it will not re-advertise it to PE4.

7. Load Balancing

It is recommended to configure different PW priority values for different VPLS instance, then the active PW of different VPLS will be running on different PEs, to provide load balancing between the two PE in one domain.

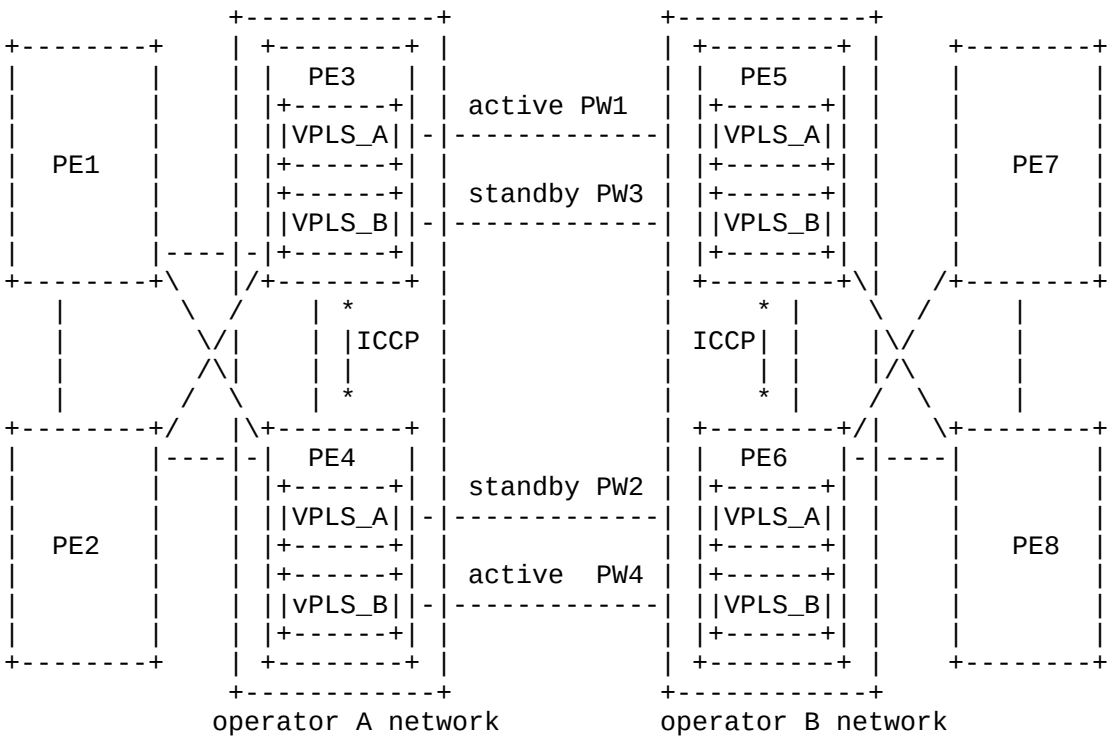


Figure 2

In figure 2, it shows two VPLS inter-AS deployment. VPLS_A will have the PW between PE3 and PE5 with higher priority than the one between PE4 and PE6, while VPLS_B will have the PW between PE4 and PE6 with

higher priority than the one between PE3 and PE5. Then PE3&PE4 or PE5&PE6 can provider load balance among different VPLS instance.

8. Security Considerations

This section will be added in a future version.

9. Normative references

[I-D.ietf-pwe3-iccp]
Martini, L., Salam, S., Sajassi, A., Bocci, M., Matsushima, S., and T. Nadeau, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-05 (work in progress), April 2011.

[I-D.ietf-pwe3-redundancy]
Muley, P., "Pseudowire (PW) Redundancy", draft-ietf-pwe3-redundancy-03 (work in progress), May 2010.

[I-D.ietf-pwe3-redundancy-bit]
Muley, P. and M. Aissaoui, "Pseudowire Preferential Forwarding Status Bit", draft-ietf-pwe3-redundancy-bit-04 (work in progress), March 2011.

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

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

Authors' Addresses

Zhihua Liu
China Telecom
109 Zhongshan Ave.
Guangzhou 510630
P.R.China

Email: zhliu@gsta.com

Lizhong Jin
ZTE Corporation
889 Bibo Road
Shanghai 201203
P.R.China

Email: lizhong.jin@zte.com.cn

Ran Chen
ZTE Corporation
68 Zijinghua Road
Nanjing 210012
P.R.China

Email: chen.ran@zte.com.cn

Network Working Group
Internet Draft
Category: Standard Track

R. Ram, Orckit-Corrigent
D. Cohn, Orckit-Corrigent
R. Key, Telstra
P. Agarwal, Broadcom

Expires: November 18, 2011

May 18, 2011

Extension to LDP-VPLS for E-Tree Using Two PW
draft-ram-l2vpn-ldp-vpls-etree-2pw-02.txt

Status of this Memo

Copyright Notice

Abstract

This document proposes a solution for Metro Ethernet Forum (MEF) Ethernet Tree (E-Tree) support in Virtual Private LAN Service using LDP Signaling (LDP-VPLS) [RFC4762]. The proposed solution is characterized by the use of two PWs between a pair of PEs. This solution is applicable for both VPLS and H-VPLS.

Table of Contents

1. Introduction

This document proposes a solution for Metro Ethernet Forum (MEF) Tree (E-Tree) support in Virtual Private LAN Service using LDP Signaling (LDP-VPLS) [RFC4762].

[Draft ETree VPLS Req] is used as requirement specification.

The proposed solution is characterized by the use of two PWs between a pair of PEs, which requires extension to the current VPLS standard [RFC4762].

This solution is applicable for both VPLS and H-VPLS.

The proposed solution is composed of three main components:

- Current standard LDP-VPLS [RFC4762]
- Extension to LDP-VPLS specified in this document
- PE local split horizon mechanism

2. Conventions used in this document

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 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. The Problem

[Draft ETree VPLS Req] identifies the problem when there are two or more PEs with both Root AC and Leaf AC.

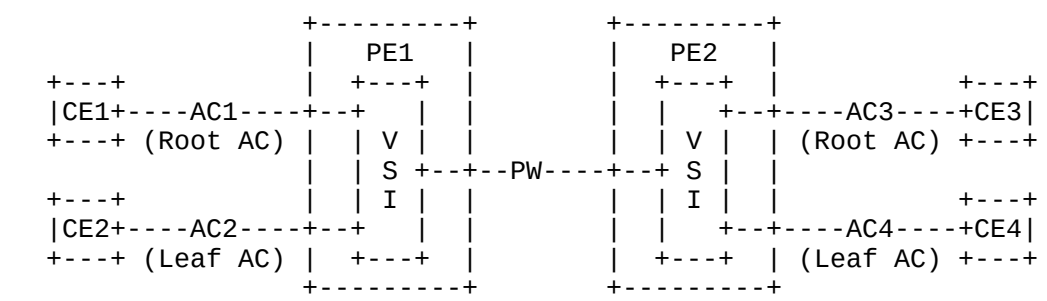


Figure 1: Problem Scenario for Leaf-to-Leaf Communication Restriction

When PE2 receives a frame from PE1 via the Ethernet PW,
- PE2 does not know whether the ingress AC is a Leaf AC or not
- PE2 does not have sufficient information to enforce the Leaf-to-Leaf communication restriction

4. The 2-PW Solution

A simple fix is to carry additional information with each frame on the PW, indicating whether the frame is originated from a Leaf AC or a Root AC on the ingress PE.

The proposed solution uses a pair of PWs to interconnect two VPLS PEs:

- First PW is used for frames originated from Root ACs
- Second PW is used for frames originated from Leaf ACs

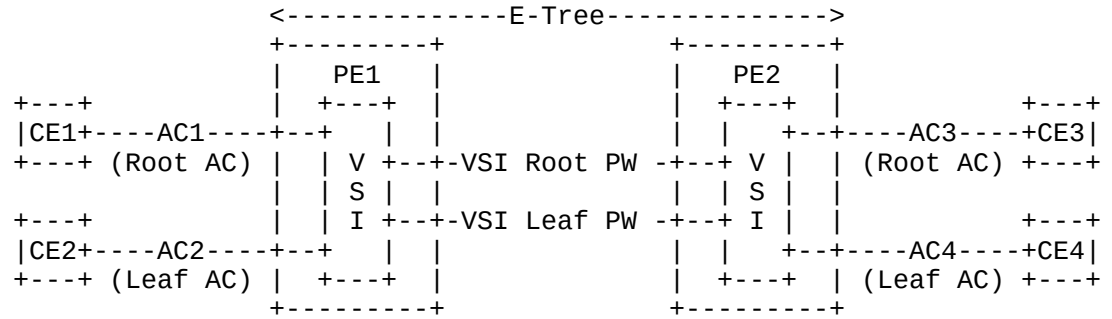


Figure 2: Two-PW Solution for Leaf-to-Leaf Communication Restriction

Extension to current VPLS standard [RFC4762] is required.

5. Extension to VPLS for E-Tree

5.1. AC E-Tree Type

Each AC connected to a specific VPLS instance on a PE MUST have an AC E-Tree Type attribute, either Leaf AC or Root AC. For backward compatibility, the default AC E-Tree Type MUST be Root.

This AC E-Tree Type is locally configured on a PE and no signaling is required between PEs.

5.2. VSI E-Tree Type and Identifier

Two new PW interface parameters (as defined in section 5.5 of [RFC4447]) are defined for use in E-Tree VPLS: VSI E-Tree type and VSI E-Tree identifier.

VSI E-Tree type can be either root or leaf and identifies VSI root PW and VSI leaf PW respectively, as defined in section 4.

VSI E-tree identifier is a number that is used to identify a pair of root and leaf PW as part of the same logical VSI interface.

On reception, the two PWs SHALL be handled as the same logical VSI interface with respect to MAC address learning/forwarding, e.g. traffic SHALL NOT be forwarded between such PWs and MAC addresses arriving at one of the PWs SHALL be learned with a common logical VSI interface.

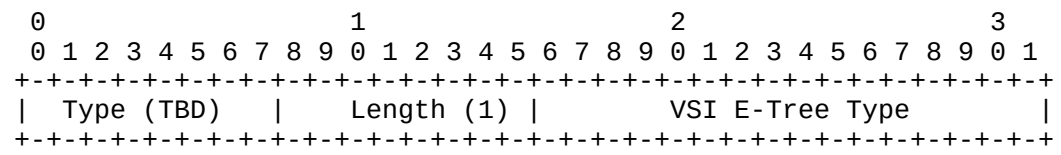
On transmission, the VPLS processing entity SHALL send root-originated traffic via the root PW, and SHALL send leaf-originated traffic via the leaf PW.

The <VSI E-Tree type, VSI E-Tree identifier> pair SHALL be unique in PWs connecting a pair of VPLS PEs.

5.2.1. VSI E-Tree Type Encoding

The VSI E-Tree type field is encoded as an interface parameters sub-TLV (as defined in section 5.5 of [RFC4447]).

The field structure is defined as follows:



VSI E-tree Type can take the following values:

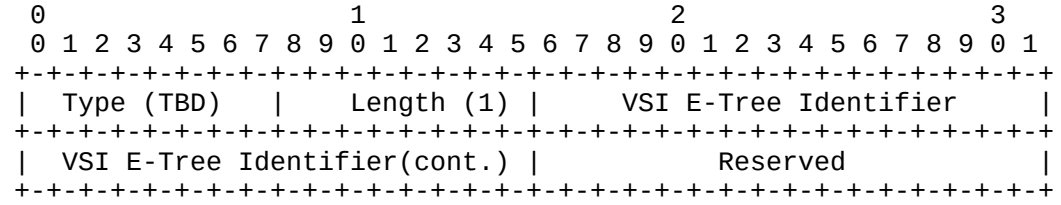
0 E-Tree Root VSI

1 E-Tree Leaf VSI

5.2.2. VSI E-Tree Identifier Encoding

The VSI E-Tree identifier field is encoded as an interface parameters sub-TLV (as defined in section 5.5 of [RFC4447]).

The field structure is defined as follows:



VSI E-tree Identifier is a 32-bit number that is used to identify a pair of root and leaf PW as part of the same logical VSI interface, in the context of a pair of VPLS PEs.

The reserved field SHALL be set to zero.

5.3. Additional Filtering in Data Forwarding

An egress PE SHALL NOT deliver a frame originated at a leaf AC to another leaf AC.

The following specifies how AC E-Tree type per frame is determined:

- A frame received from a root PW indicates that the frame was originated from a root AC
- A frame received from a leaf PW indicates that the frame was originated from a leaf AC.

o For the case where both ingress AC and egress AC are on the same PE, local split horizon implementation on the PE will be sufficient, and is not further discussed in this document.

5.4. Root/Leaf PWs Signaling

Signaling of root and leaf PWs is required only when two PWs are used for interconnecting between pair of VSIs. As explained in section . 5.2. :

o Root VSI E-Tree type SHALL be used to signal a root PW.

o Leaf VSI E-Tree type be used to signal a leaf PW.

PW type signaling rules remain as defined in [RFC4447].

5.5. Supporting Remote AC

When PW is used to interconnect between VSI and a remote AC (e.g. the PW1, PW2 in Figure 3), an Ethernet Raw or Ethernet tagged PW types SHALL be used as defined in [RFC4762].

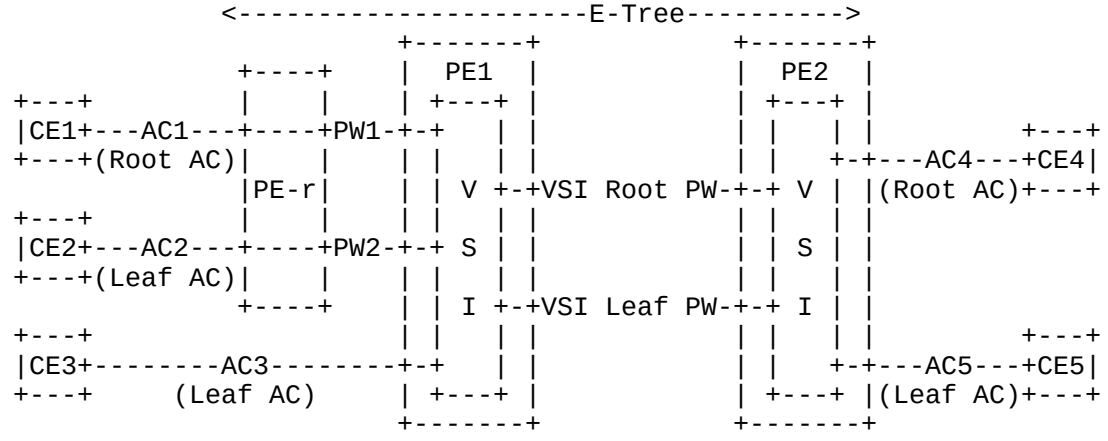


Figure 3: VPLS with Remote AC Connectivity

In addition, the AC type i.e. Root or leaf, SHALL be locally provisioned on the VSI side to specify the remote AC E-Tree Type per PW. Moreover, such PWs that are used for interconnecting between a remote AC and a VSI SHALL considered as separate logical VSI interfaces with respect to MAC address learning/forwarding e.g. traffic forwarding between such PWs is allowed as long as they are not both defined as Leaf.

In Figure 3, AC1 is remotely interconnected to the VPLS service via PW1, and AC2 is remotely interconnected to the VPLS service via PW2.

AC1 is a Root AC and therefore the local type for PW1 in PE1 SHALL be Root.

AC2 is a Leaf AC and therefore the local type for PW2 in PE1 SHALL be Leaf.

6. Backward Compatibility

Root or leaf VSI E-Tree type and identifier parameters SHALL be used only in cases where both PEs are VPLS capable and both support E-Tree root/leaf.

In a case where one of the peers do not support E-Tree, VSI E-Tree type and identifier parameters SHALL NOT be used.

7. Compliance with Requirements

This refers to [Draft ETree VPLS Req] Section 5. Requirements.

The solution prohibits communication between any two Leaf ACs in a VPLS instance.

The solution allows multiple Root ACs in a VPLS instance.

The solution allows Root AC and Leaf AC of a VPLS instance co-exist on any PE.

The solution is applicable to LDP-VPLS [RFC4762].

The solution is applicable to Case 1: Single technology "VPLS Only".

8. Security Considerations

This will be added in later version.

9. IANA Considerations

Additional assignments will be required for the new interface parameter sub-TLV types introduced in Section 4.2. Details will be added in a later version.

10. Acknowledgements

The authors wish to acknowledge the contributions of Luca Martini and Amir Halperin.

11. References

11.1. Normative References

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

[RFC4447] Martini, L., and al, Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP), April 2006

[RFC4762] Lasserre & Kompella, Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling, January 2007

11.2. Informative References

[Draft VPLS ETree Req] Key, et al., Requirements for MEF E-Tree Support in VPLS, draft-key-l2vpn-vpls-etree-reqt-01.txt, September 2010

Authors' Addresses

Rafi Ram
Orckit-Corrigent
126 Yigal Alon st.
Tel Aviv, Israel
Email: rafir@orckit.com

Daniel Cohn
Orckit-Corrigent
126 Yigal Alon st.
Tel Aviv, Israel
Email: danielc@orckit.com

Raymond Key
Telstra
242 Exhibition Street, Melbourne
VIC 3000, Australia
Email: raymond.key@team.telstra.com

Puneet Agarwal
Broadcom
3151 Zanker Road
San Jose, CA 95134
Email: pagarwal@broadcom.com

INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Keyur Patel
Cisco

Expires: January 4, 2011

July 4, 2011

E-VPN Ethernet Segment Route
draft-sajassi-l2vpn-evpn-segment-route-00.txt

Status of this Memo

Copyright and License Notice

Abstract

[E-VPN] defines a solution and architecture for BGP MPLS-based Ethernet VPNs. This document describes an additional BGP route and associated route attributes that enhance the multi-homing capabilities of the solution. These are: the Ethernet Segment Route, the ESI Import Extended Community, the DF Election Attribute and the Inter-chassis Communication Attribute. This draft describes their usage, advantages and encoding.

Table of Contents

1 Introduction

[E-VPN] defines a solution and architecture for BGP MPLS-based Ethernet L2VPN services with advanced multi-homing capabilities. To that end, [E-VPN] defines a new BGP NLRI with 5 route types:

1. Ethernet Auto-Discovery (A-D) route
2. MAC advertisement route
3. Inclusive Multicast Route
5. Selective Multicast Auto-Discovery (A-D) Route
6. Leaf Auto-Discovery (A-D) Route

In this draft, we define one additional route type:

4. Ethernet Segment Route

This route primarily enhances the multi-homing capabilities of the E-VPN solution in the following areas:

- Preventing transient loops and packet duplication
- Support of multi-chassis Ethernet bundles
- Designated Forwarder election with VLAN carving
- Avoiding relearning of subscriber/session state

In addition to the above route, 3 new BGP route attributes are defined: the ESI Import Extended Community attribute, the DF Election attribute and the Inter-chassis Communication attribute.

Section 2 discusses the motivation and usage of the new route and attributes. Section 3 describes the BGP encoding.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2 Motivation and Usage

This section focuses on the reasons for defining the Ethernet Segment route and its associated 3 BGP attributes, and describes its usage in E-VPN.

2.1 Preventing Transient Loops and Packet Duplication

The Designated Forwarder (DF) election procedures defined in [E-VPN] require that each MES constructs a candidate list of DFs from the received Ethernet A-D routes. By default, each MES then independently

chooses the MES with the highest IP address as the elected DF. There is no handshake mechanism between the MESes that are connected to the same Ethernet Segment. As a result of that, during routing transients, different MESes may end up electing different DFs for the same Ethernet Segment due to inconsistent views of the network. If the Ethernet Segment is a multi-homed device, this may lead to transient packet duplication. If the Ethernet Segment is a multi-homed network, the presence of multiple DFs may lead to transient forwarding loops in addition to potential packet duplication.

To eliminate these issues, a handshake mechanism is required between the MES nodes connected to the same Ethernet Segment, to ensure a common view of the network among them. This handshake is performed using the DF Election attribute carried in the Ethernet Segment route, as discussed in the 'DF Election with Paxos Algorithm' section.

2.2 Support of Multi-Chassis Ethernet Bundles

When a CE is multi-homed to a set of MES nodes using the [802.1AX] Link Aggregation Control Protocol (LACP), the MESes must act as if they were a single LACP speaker for the Ethernet links to form a bundle, and operate correctly as a Link Aggregation Group (LAG). To achieve this, the MESes connected to the same multi-homed CE must synchronize LACP configuration and operational data among them. The synchronization is required for the following reasons:

- to determine if the links in the Ethernet bundle are to operate in all-active or hot-standby resiliency mode

- to detect and handle CE mis-configuration when LACP Port Key is configured on the MES

- to detect and handle mis-wiring between CE and MES when LACP Port Key is configured on the MES

- to deterministically agree on which link(s) should join a bundle based on port and system priorities, especially when the number of links exceeds the aggregation capacity of the MESes, and the MES LACP System Priority is higher than the CE's

- to detect and react to actor/partner churn where the LACP speakers are not able to converge

Synchronization of LACP state between MESes is performed using the Inter-chassis Communication attribute carried in the Ethernet Segment route, as described in the 'LACP State Synchronization' section below.

2.3 Designated Forwarder (DF) Election with VLAN Carving

In the case where multiple MES nodes offer redundant connectivity for an Ethernet Segment, it is preferred to elect multiple DFs (one DF per VLAN) in order to distribute the traffic among the redundancy group members. This process of electing different DFs for different VLANs on an Ethernet Segment, for purpose of load-balancing, is referred to as 'VLAN Carving'.

The VLAN carving algorithm must ensure even distribution of VLANs among the MES nodes servicing the same Ethernet Segment. As new MES devices get commissioned or decommissioned, the VLANs must be redistributed over the available devices for even load-balancing. However, in the case of link, port or node failure, the VLAN carving algorithm should ensure that only the affected VLANs are reassigned to different MES(es), and none of the other active VLANs are shuffled. Otherwise, the fault decoupling capability of the redundancy group would be compromised.

VLAN carving requires exchange of information among the MES nodes connected to an Ethernet Segment in order to agree upon how the VLANs will be distributed. Since this information is only relevant to the MES nodes that are directly connected to a specific Ethernet Segment, the exchanges and associated processing should be localized to the redundancy group members.

DF Election with VLAN carving is performed using the DF Election attribute carried in the Ethernet Segment route, as described in the "VLAN Carving" section below.

2.4 Route Scalability with Granular DF Election

[E-VPN] allows for DF election to be performed at the granularity of either an Ethernet Segment or combination of Ethernet Segment and VLAN on that segment. In the latter case, an Ethernet A-D route per (ESI, VLAN) must be advertised by the MES regardless of whether the service interface is port-based, VLAN-based, VLAN bundling-based or VLAN aware bundling-based. In case of port-based and VLAN bundling-based services, these routes are only required for DF election and not for advertising forwarding labels. By using the Ethernet Segment route instead of the Ethernet A-D route for DF election, it is still possible to have per-VLAN DF granularity while significantly reducing the number of BGP routes advertised. For e.g., consider an Ethernet Segment ESI1 used for a port-based service. By using the Ethernet A-D route for per (ESI, VLAN) DF election, 4095 routes are needed. Whereas, using the Ethernet Segment route, only a single route is required.

2.5 Avoiding Relearning of Subscriber/Session State

For certain applications, the MES builds and maintains per subscriber or per session 'soft' state that is used for either optimizing the traffic forwarding or enforcing security. Examples of such per subscriber/session state includes:

- multicast state derived from IGMP or PIM snooping

- IP address to MAC address bindings gleaned from snooping ARP and/or DHCP packets, and used to prevent address spoofing or masquerading

When a set of MES nodes provides multi-homed connectivity for an Ethernet Segment, this 'soft' state is built on the active MES node that forwards and snoops the relevant protocol packets. In case of a link or node failure, the state must be reconstructed on the backup MES (e.g. by waiting for the next IGMP query or ARP message or by issuing unsolicited queries). This may cause traffic disruption and affect the availability of the service. Alternatively, the state can be synchronized among the MES nodes via BGP, and that would enhance the convergence of the service after failure.

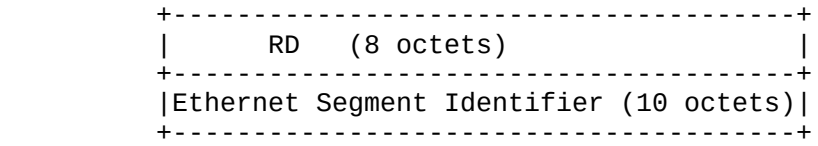
Synchronization of subscriber/session state between MES nodes is performed using the Inter-chassis Communication attribute carried in the Ethernet Segment route, as described in the 'Subscriber/Session State Synchronization' section below.

3 BGP Encoding

This section defines the encoding of the BGP route and attributes.

3.1 Ethernet Segment Route

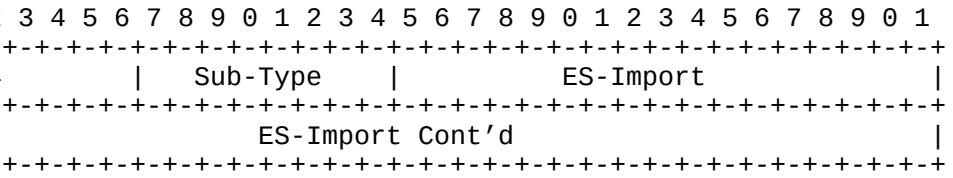
The Ethernet Segment Route is encoded in the E-VPN NLRI defined in [E-VPN] using the Route Type value of 4. The Route Type Specific field of the NLRI is formatted as follows:



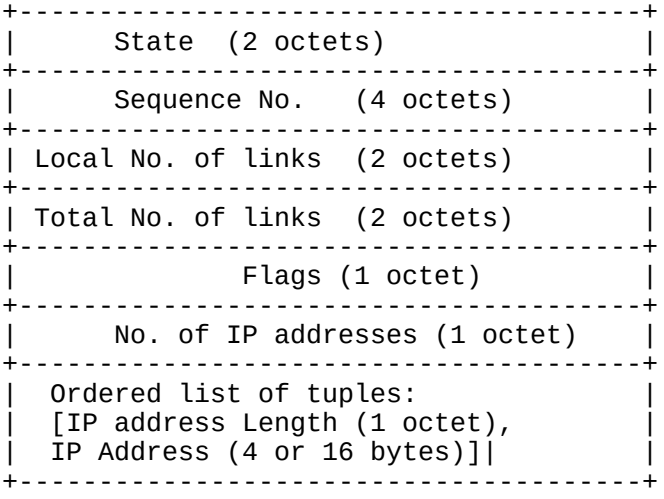
3.2 ES-Import Extended Community

This is a new transitive extended community carried with the Ethernet Segment route. When used, it enables all the MESes connected to the

same multi-homed site to import the Ethernet Segment routes. The value is derived automatically from the ESI by encoding the 6-byte MAC address portion of the ESI in the ES-Import Extended Community. The format of this extended community is as follows:



3.3 DF Election Attribute



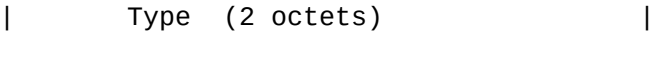
State field can take one of the following values:

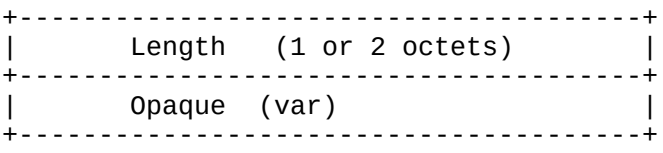
0x0000 Initializing
0x0001 Proposal Pending
0x0002 Promise Pending
0x0003 Active

Flags field is encoded as follows:

7 bits: reserved
Least significant bit: Protecting flag

3.4 Inter-chassis Communication Attribute





4 DF Election with Paxos Algorithm

The procedures in this section guarantee that all MES nodes in a given redundancy group agree on a unique DF for a given Ethernet Segment. This eliminates the problem of transient forwarding loops and transient packet duplicates described above. The procedures can be broken down to the following steps:

1. When a MES discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute and with the 'Initializing' code in the State field of the DF Election attribute.

2. The MES then starts a timer to allow the reception of Ethernet Segment routes from other MES nodes in the same redundancy group.

3. When the timer expires, each MES builds an ordered list of the IP addresses of all the MES nodes connected to the Ethernet Segment (including itself), in increasing numeric value.

4. The first MES in the ordered list then elects itself as the Arbiter Node (AN). It initiates the handshake by sending an Ethernet Segment route with 'Proposal Pending' code in the State field of the DF Election attribute.

5. When a MES node receives an Ethernet Segment route with the 'Proposal Pending' code, it takes one of the following options:

   a. If the receiving MES ranks the transmitting MES's IP address as the top entry in its local ordered list, it acknowledges the handshake by responding with an Ethernet Segment route with the 'Promise Pending' code in the State field of the DF Election attribute. This includes the scenario where the receiving MES forfeits the AN role to another advertising MES with a numerically lower IP address.

   b. If the receiving MES does not rank the transmitting MES's IP address as the top entry in its local ordered list, and the receiving MES had advertised an Ethernet Segment route with the 'Initializing' code or with the 'Proposal Pending' code, then the MES takes no further action.

6. When the AN receives 'Promise Pending' from all of the MES nodes in the ordered list, it sends an updated Ethernet Segment route with the 'Active' code in the DF Election attribute.

7. When the other MES nodes in the redundancy group receive the 'Active' code from the AN, they respond with an updated Ethernet Segment route with the 'Active' code in the DF Election attribute. This concludes the handshake.

In the case where the DF election is performed at the granularity of an Ethernet Segment, i.e. there is a single DF for all VLANs on the segment, the Arbiter Node is effectively the Designated Forwarder for the segment. All the MES nodes start off with their ports, that are connected to the segment, blocked in Step 1 (for multi-destination traffic from core). And in Step 6, the MES confirmed as the AN (i.e. DF) unblocks its port towards the Ethernet Segment. DF election at the granularity of (Ethernet Segment, VLAN) is discussed in the "VLAN Carving" section below.

5 LACP State Synchronization

To support CE multi-homing with multi-chassis Ethernet bundles, the MES nodes connected to a given CE should synchronize [802.1AX] LACP state amongst each other. This includes the following LACP specific configuration parameters:

- System Identifier (MAC Address): uniquely identifies a LACP speaker.
- System Priority: determines which LACP speaker's port priorities are used in the Selection logic.
- Aggregator Identifier: uniquely identifies a bundle within a LACP speaker.
- Aggregator MAC Address: identifies the MAC address of the bundle.
- Aggregator Key: used to determine which ports can join an Aggregator.
- Port Number: uniquely identifies an interface within a LACP speaker.
- Port Key: determines the set of ports that can be bundled.
- Port Priority: determines a port's precedence level to join a bundle in case the number of eligible ports exceeds the maximum number of links allowed in a bundle.

The above information must be synchronized between the MES nodes wishing to form a multi-chassis bundle with a given CE, in order for the former to convey a single LACP peer to that CE. This is required for initial system bring-up and upon any configuration change. Furthermore, the MESes must also synchronize operational (run-time) data, in order for the LACP Selection logic state-machines to

execute. This operational data includes the following LACP operational parameters, on a per port basis:

- Partner System Identifier: this is the CE System MAC address.
- Partner System Priority: the CE LACP System Priority
- Partner Port Number: CE's AC port number.
- Partner Port Priority: CE's AC Port Priority.
- Partner Key: CE's key for this AC.
- Partner State: CE's LACP State for the AC.
- Actor State: PE's LACP State for the AC.
- Port State: PE's AC port status.

The above state needs to be communicated between MESes forming a multi-chassis bundle during LACP initial bring-up, upon any configuration change and upon the occurrence of a failure.

It should be noted that the above configuration and operational state is localized in scope and is only relevant to PEs within a given Redundancy Group, i.e. which connect to the same Ethernet Segment over a given Ethernet bundle. Furthermore, the communication of state changes, upon failures, must occur with minimal latency, in order to minimize the switchover time and consequent service disruption.

Without synchronization of the above parameters, the system is subject to the issues outlined in section 2.2 above.

6 VLAN Carving

It is possible to elect multiple DFs per Ethernet Segment (one per VLAN) by using a slightly modified version of the procedures described in the "DF Election with Paxos Algorithm" section above.

In step 3, each of the MES nodes assigns an ordinal for itself based on the order of its IP address in the list. The first MES in the list (the one with the numerically lowest IP address) is given an ordinal of 0. The ordinals are used to determine which MES node will be the DF for a given VLAN on the Ethernet Segment using the following rule:

Assuming a redundancy group of N MES nodes, the MES with ordinal i is the DF for VLAN V when (V MOD N) = i.

In step 6, the AN unblocks only the VLANs for which it is a DF for the Ethernet Segment.

In step 7, each MES node unblocks only the VLANs for which it is a DF for the Ethernet Segment.

In the case of a port, link or node failure, the AN takes over the forwarding for the affected VLANs on the segment and advertises an updated Ethernet Segment route with the 'Active' code and 'Protecting' flag set in the DF Election attribute. Therefore, when VLAN carving is used, the AN acts as the Backup DF (BDF) for the Ethernet Segment. This ensures that only the affected VLANs are failed over, and none of the other VLANs are shuffled.

When the fault clears, the following procedure is followed to revert the VLANs to the recovering MES:

1. The recovering MES advertises an Ethernet Segment route with the 'Initializing' code in the State field of the DF Election attribute.

2. The recovering MES receives from the other MES nodes Ethernet Segment routes with the 'Active' code in the DF Election attribute. The MES can, then, build its ordered list.

3. The recovering MES advertises an Ethernet Segment route with the 'Proposal-Pending' code in the DF Election attribute. This is meant to indicate to the AN that the recovering MES is ready to take over its VLANs.

4. Upon receiving the route with the 'Proposal Pending' code, the AN blocks all the VLANs that belong to the recovering MES. The AN then advertises an updated Ethernet Segment route with the 'Protecting' flag cleared.

5. Upon receiving the above route from the AN, the recovering MES unblocks the VLANs for which it is the DF. The recovering MES then transmits an Ethernet Segment route with the 'Active' code. This completes the reversion.

If the failed MES is the AN, then the MES node with the second best claim to be AN (i.e. whose IP address is the second in the ordered list) takes over the failed VLANs and advertises an updated Ethernet Segment route with the 'Active' code and 'Protecting' flag set in the DF Election attribute. The procedures for reversion, in this case, are as follows:

1. The recovering AN advertises an Ethernet Segment route with the 'Initializing' code in the State field of the DF Election attribute.

2. The recovering AN receives from the other MES nodes Ethernet Segment routes with the 'Active' code in the DF Election attribute.

3. The recovering AN advertises an Ethernet Segment route with the 'Proposal-Pending' code in the DF Election attribute.

4. The other MES nodes respond to that advertisement with Ethernet Segment routes with the 'Promise-Pending' code in the DF Election attribute. At this point, the BDF blocks all the VLANs that belong to the recovering AN before advertising its Ethernet Segment route, with the 'Promise-Pending' code and 'Protecting' flag cleared.

5. The recovering AN unblocks the VLANs for which it is the DF upon receiving the 'Promise-Pending' advertisements from the BDF. The AN then advertises an Ethernet Segment route with the 'Active' code once it receives the Ethernet Segment route with 'Promise-Pending' code from all of the MES nodes in the redundancy group.

6. The other MES nodes respond with Ethernet Segment routes with the 'Active' code. This marks the end of the reversion.

7 Subscriber/Session State Synchronization

Synchronization of subscriber/session state between MES nodes is performed using the Inter-chassis Communication attribute carried in the Ethernet Segment route. The various applications are responsible for the encoding and decoding of the relevant data, and this is outside the scope of this draft. BGP provides a reliable transport service in this case.

8 Security Considerations

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

9 IANA Considerations

To be added in a later revision.

10 References

10.1 Normative References

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

10.2 Informative References

[E-VPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-raggarwa-sajassi-l2vpn-evpn-02.txt, work in progress,

March, 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.

Author's 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

Keyur Patel
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: keyupate@cisco.com

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 Isaac
Bloomberg

Expires: January 11, 2012 July 11, 2011

PBB E-VPN
draft-sajassi-l2vpn-pbb-evpn-02.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on January 11, 2012.

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
Clarence Filsfils
Dennis Cai
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.

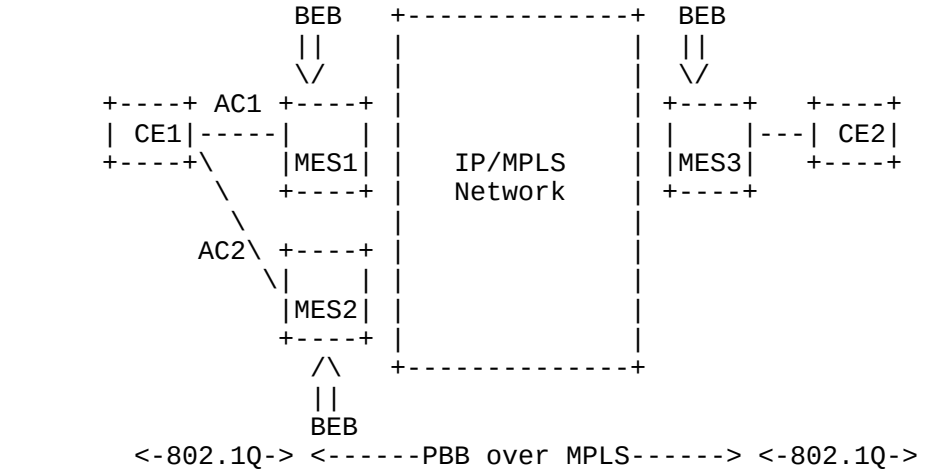


Figure 1: PBB-EVPN Network

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 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.

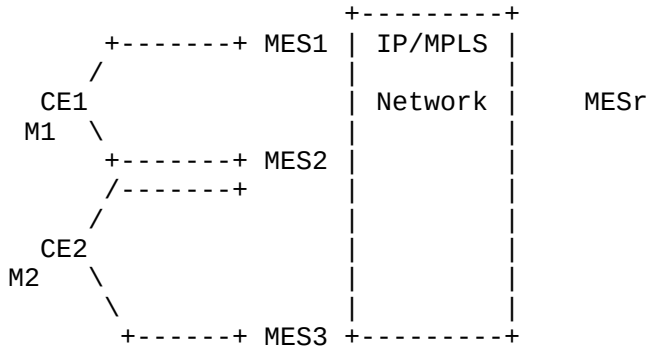


Figure 2: B-MAC Address Assignment

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).

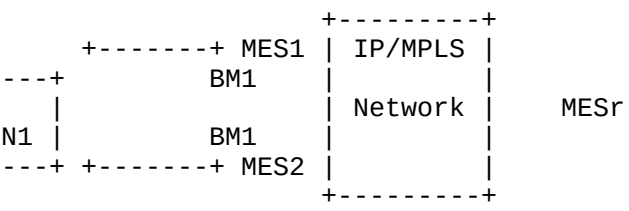


Figure 3: Multi-homed Network

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. 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.

8.1. 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.

8.2. 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.

8.3. 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.

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

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.

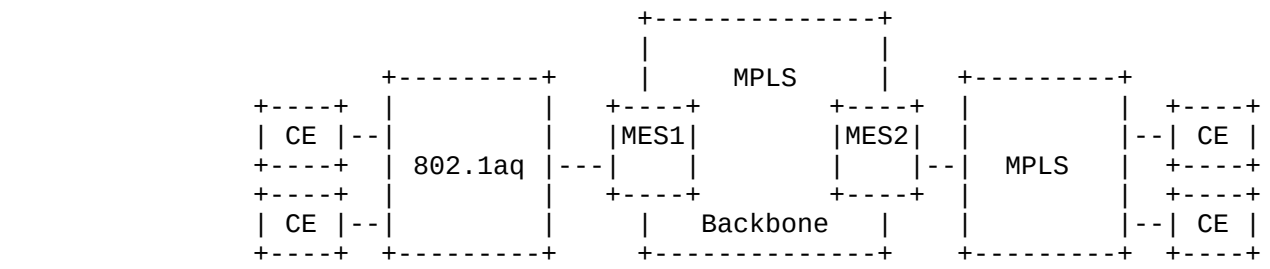


Figure 3: Interoperability with 802.1aq

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.

8.5. 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.

8.6. 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 MES2. MES2 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.

9. Acknowledgements

TBD.

10. Security Considerations

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

11. IANA Considerations

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

12. Intellectual Property Considerations

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

13. Normative References

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

14. 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.

15. 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