

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 30, 2014

S. Sivabalan
S. Boutros
Cisco Systems, Inc.
H. Shah
Ciena Corp.
September 30, 2013

MAC Address Withdrawal over Static Pseudowire
draft-boutros-l2vpn-mppls-tp-mac-wd-02.txt

Abstract

This document specifies a mechanism to signal MAC address withdrawal notification using PW Associated Channel (ACH). Such notification is useful when statically provisioned PWs are deployed in VPLS/H-VPLS environment.

Requirements Language

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].

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 March 26, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	2
3. MAC Withdraw OAM Message	3
4. Operation	4
4.1. Operation of Sender	4
4.2. Operation of Receiver	5
5. IANA Considerations	5
6. References	5
6.1. Normative References	5
6.2. Informative References	6
Authors' Addresses	6

1. Introduction

An LDP-based MAC Address Withdrawal Mechanism is specified in [RFC4762] to remove dynamically learned MAC addresses when the source of those addresses can no longer forward traffic. This is accomplished by sending an LDP Address Withdraw Message with a MAC List TLV containing the MAC addressed to be removed to all other PEs over LDP sessions. When the number of MAC addresses to be removed is large, empty MAC List TLV may be used. [MAC-OPT] describes an optimized MAC withdrawal mechanism which can be used to remove only the set of MAC addresses that need to be re-learned in H-VPLS networks. The solution also provides optimized MAC Withdrawal operations in PBB-VPLS networks.

A PW can be signaled via LDP or can be statically provisioned. In the case of static PW, LDP based MAC withdrawal mechanism cannot be used. This is analogous to the problem and solution described in [RFC4762] where PW OAM message has been introduced to carry PW status TLV using in-band PW Associated Channel. In this document, we propose to use PW OAM message to withdraw MAC address(es) learned via static PW.

2. Terminology

The following terminologies are used in this document:

ACK: Acknowledgement for MAC withdraw message.

LDP: Label Distribution Protocol.

MAC: Media Access Control.

PE: Provide Edge Node.

MPLS: Multi Protocol Label Switching.

PW: PseudoWire.

PW OAM: PW Operations, Administration and Maintenance.

TLV: Type, Length, and Value.

VPLS: Virtual Private LAN Services.

3. MAC Withdraw OAM Message

LDP provides a reliable packet transport for control plackets for dynamic PWs. This can be contrasted with static PWs which rely on re-transmission and acknowledgments (ACK) for reliable OAM packet delivery as described in [RFC6478]. The proposed solution for MAC withdrawal over static PW also relies on re-transmissions and ACKs. However, ACK is mandatory. A given MAC withdrawal notification is sent as a PW OAM message, and the sender keeps re-transmitting the message until it receives an ACK for that message. Once a receiver successfully remove MAC address(es) in response to a MAC address withdraw OAM message, it should not unnecessarily remove MAC address(es) upon getting refresh message(s). To facilitate this, the proposed mechanism uses sequence number, and defines a new TLV to carry the sequence number.

The format of the MAC address withdraw OAM message is shown in Figure 1. The PW OAM message header is exactly the same as what is defined in [RFC6478]. Since the MAC withdrawal PW OAM message is not refreshed forever. A MAC address withdraw OAM message MUST contain a "Sequence Number TLV" otherwise the entire message is dropped. It MAY contain MAC Flush Parameter TLVs defined in [MAC-OPT] when static PWs are deployed in H-VPLS and PBB-VPLS scenarios. The first 2 bits of the sequence number TLV are reserved and MUST be set to 0 on transmit and ignored on receipt.

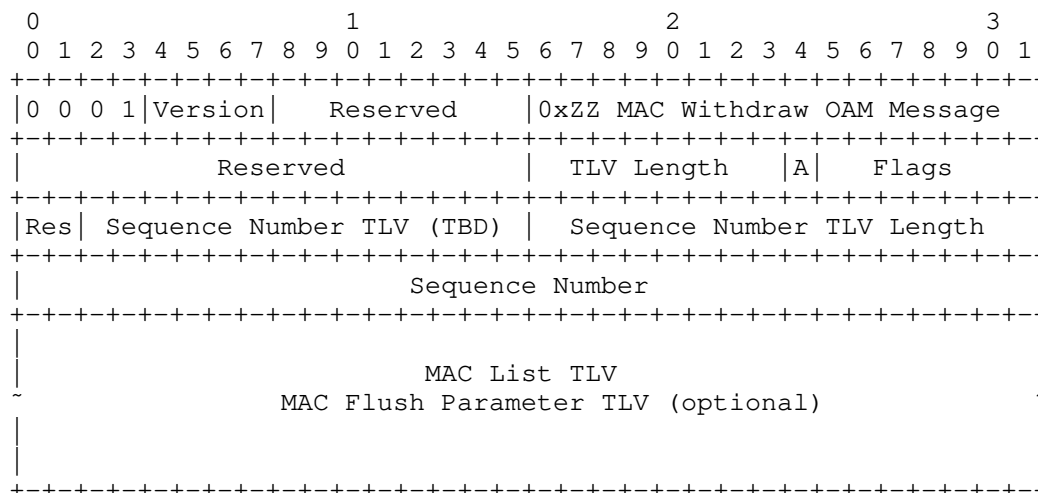


Figure 1: MAC Address Withdraw PW OAM Packet Format

In this section, MAC List TLV and MAC Flush Parameter TLV are collectively referred to as "MAC TLV(s)". The processing rules of MAC List TLV are governed by [RFC4762], and the corresponding rules of MAC Flush Parameter TLV are governed by [MAC-OPT].

"TLV Length" is the total length of all TLVs in the message, and "Sequence Number TLV Length" is the length of the sequence number field.

A single bit (called A-bit) is set to indicate if a MAC withdraw message is for ACK. Also, ACK does not include MAC TLV(s).

Only half of the sequence number space is used. Modular arithmetic is used to detect wrapping of sequence number. When sequence number wraps, all MAC addresses are flushed and the sequence number is reset.

4. Operation

This section describes how the initial MAC withdraw OAM messages are sent and retransmitted, as well as how the messages are processed and retransmitted messages are identified.

4.1. Operation of Sender

Each PW is associated with a counter to keep track of the sequence number of the transmitted MAC withdrawal messages. Whenever a node

sends a new set of MAC TLVs, it increments the transmitted sequence number counter, and include the new sequence number in the message.

The sender expects an ACK from the receiver within a time interval which we call "Retransmit Time" which can be either a default or configured value. If the ACK arrives within the Retransmit Time, the sender assumes that the message transmission is successful. Otherwise, it retransmits the message with the same sequence number as the original message.

4.2. Operation of Receiver

Each PW is associated with a register to keep track of the sequence number of the MAC withdrawal message received last. Whenever a MAC withdrawal message is received, and if the sequence number on the message is greater than the value in the register, the MAC address(es) contained in the MAC TLV(s) is/are removed, and the register is updated with the received sequence number. The receiver sends an ACK whose sequence number is the same as that in the received message.

If the sequence number in the received message is smaller than or equal to the value in the register, the MAC TLV(s) is/are not processed. However, an ACK with the received sequence number MUST be sent as a response. The receiver processes the ACK message as an acknowledgement for all the MAC withdraw messages sent up to the sequence number present in the ACK message and terminates retransmission.

As mentioned above, since only half of the sequence number space is used, the receiver MUST use modular arithmetic to detect wrapping of the sequence number.

5. IANA Considerations

The proposed mechanism requests IANA to assign new channel type (recommended value 0x0028) from the registry named "Pseudowire Associated Channel Types". The description of the new channel type is "Pseudowire MAC Withdraw OAM Channel".

IANA needs to create a new registry for Pseudowire Associated Channel TLVs, and create an entry for "Sequence Number TLV". The recommended value is 0x0001.

6. References

6.1. Normative References

- [MAC-OPT] Dutta, P., Balus, F., Stokes, O., and G. Calvinac, "LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS", draft-ietf-l2vpn-vpls-ldp-mac-opt-08.txt (work in progress), February 2013.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC6478] Martini, L., Swallow, G., Heron, G., and M. Bocci, "Pseudowire Status for Static Pseudowires", RFC 6478, May 2012.

6.2. Informative References

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

Authors' Addresses

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

Email: msiva@cisco.com

Sami Boutros
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
US

Email: sboutros@cisco.com

Himanshu Shah
Ciena Corp.
3939 North First Street
San Jose, CA 95134
US

Email: hshah@ciena.com

Internet Working Group

Internet Draft

Intended status: Standards Track

Y. Jiang

L. Yong

Huawei

M. Paul
Deutsche Telekom

F. Jounay

Orange CH

F. Balus
W. Henderickx
Alcatel-Lucent

A. Sajassi
Cisco

Expires: April 2014

October 10, 2013

VPLS PE Model for E-Tree Support
draft-ietf-l2vpn-vpls-pe-etree-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/lid-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 March 10, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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 PWs which carry the VLAN indicating the E-Tree attribute, the MAC address based Ethernet forwarding engine and the PW work 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	3
2.	Conventions used in this document	4
3.	Terminology	4
4.	PE Model with E-Tree Support	5
4.1.	Existing PE Models	5
4.2.	A New PE Model with E-Tree Support	8
5.	PW for E-Tree Support	9
5.1.	PW Encapsulation	9
5.2.	VLAN Mapping	9
5.3.	PW Processing	11
5.3.1.	PW Processing in the VLAN Mapping Mode	11
5.3.2.	PW Processing in the Compatible Mode	12
5.3.3.	PW Processing in the Optimized Mode	13
6.	Signaling for E-Tree Support	14
6.1.	LDP Extensions for E-Tree Support	14
6.2.	BGP Extensions for E-Tree Support	16
7.	OAM Considerations	18
8.	Applicability	18
9.	Security Considerations	18

10. IANA Considerations	19
11. References	19
11.1. Normative References	19
11.2. Informative References	20
12. Acknowledgments	20
Appendix A. Other PE Models for E-Tree	21
A.1. A PE Model With a VSI and No bridge	21
A.2. A PE Model With external E-Tree interface	22

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-to-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 [802.1Q-2011], which is just an improvement on the traditional asymmetric VLAN mechanism (the use of different VLANs to indicate E-Tree root/leaf attributes and prohibiting leaf-to-leaf

traffic with the help of VLANs was first standardized in IEEE 802.1Q-2003). 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, thus it will not change the FIB (such as installing E-Tree attributes in the FIB), or need any specially tailored implementation. Furthermore, 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 then 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 as defined in MEF 6.1

EVC: Ethernet Virtual Connection, as defined in MEF 4.0

FIB: Forwarding Information Base, or forwarding table

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

C-VLAN, Customer VLAN

S-VLAN, Service VLAN

B-VLAN, Backbone VLAN

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

I-SID, Backbone Service Instance Identifier, as defined in IEEE 802.1ah

4. PE Model with E-Tree Support

"VPLS only" PE architecture as shown 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. It can also be used by other PE models which are 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.

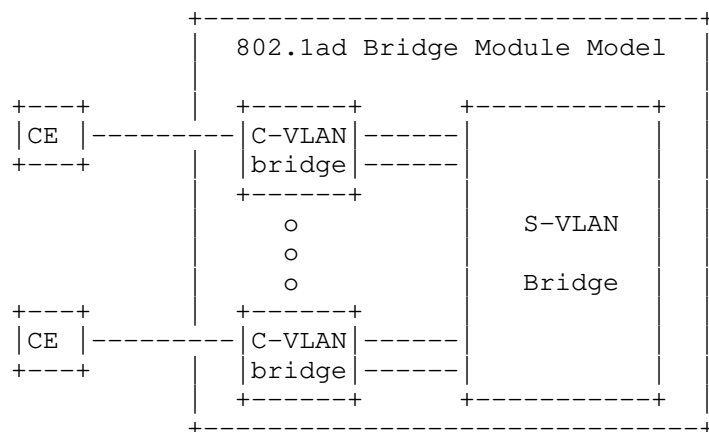


Figure 1 A model of 802.1ad Bridge Module

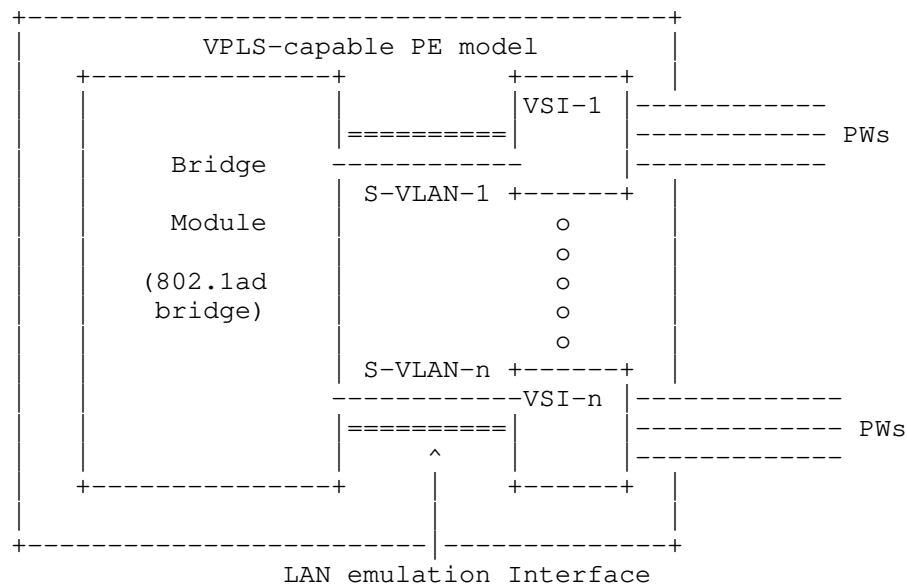


Figure 2 A VPLS-capable PE Model

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 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 to the roots and the leaves, while the frames with the leaf VLAN tag are transmitted to the roots but dropped for the leaves (these VLAN tags are removed before the frames are transmitted over the wire). It was demonstrated in [802.1Q-2011] 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.

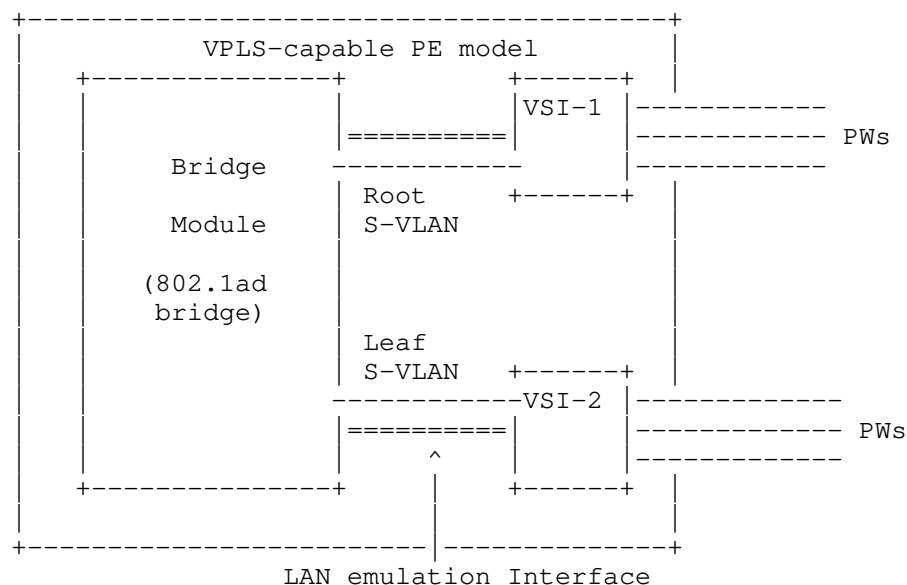


Figure 3 A VPLS PE Model for E-Tree with 2 VSIs

in this case (the E-Tree attribute may also be indicated with two I-SID tags in the bridge module, and the frames are further encapsulated and transported transparently over a single B-VLAN, thus the PBB VPLS works just in the same way as described in [PBB-VPLS] and will be discussed no more in this document). When many S-VLANs are multiplexed in a single AC, the 2nd option has an advantage of both VLAN scalability and MAC address scalability.

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

In all cases, the outermost VLAN in the resulted Ethernet header is used to indicate the E-Tree attribute of an Ethernet frame; this document will use VLAN to refer to this outermost VLAN for simplicity in the latter sections.

5. PW for E-Tree Support

5.1. PW Encapsulation

To support an E-Tree service, T-VSIs in a VPLS must be interconnected with a bidirectional Ethernet PW. The Ethernet PW may 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 frame originated from either root or leaf (the VLAN tag indicating the frame originated from either root or leaf can be translated by a bridge module in the PE or added by an outside Ethernet edge device, even by a customer device). In the tagged PW mode, two service delimiting VLANs must be allocated in the VPLS domain for an E-Tree. PW processing for the tagged PW will be described in Section 5.3 of this document.

Raw PW (PW type 0x0005 in [RFC4448]) may be used to carry E-Tree service for a PW in Compatible mode as shown in Section 5.3.2.

5.2. VLAN Mapping

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

- o 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.

- o 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 service delimiting VLANs must be allocated across the VPLS domain.

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].

Global or local VLANs can be manually configured or provisioned by an OSS system. Alternatively, some automatic VLAN allocation algorithm may be provided in the management plane, but it is out scope of this document.

For both methods, VLAN mapping parameters from a remote PE can be provisioned or determined by a signaling protocol as described in Section 6 when a 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 both root and leaf nodes).

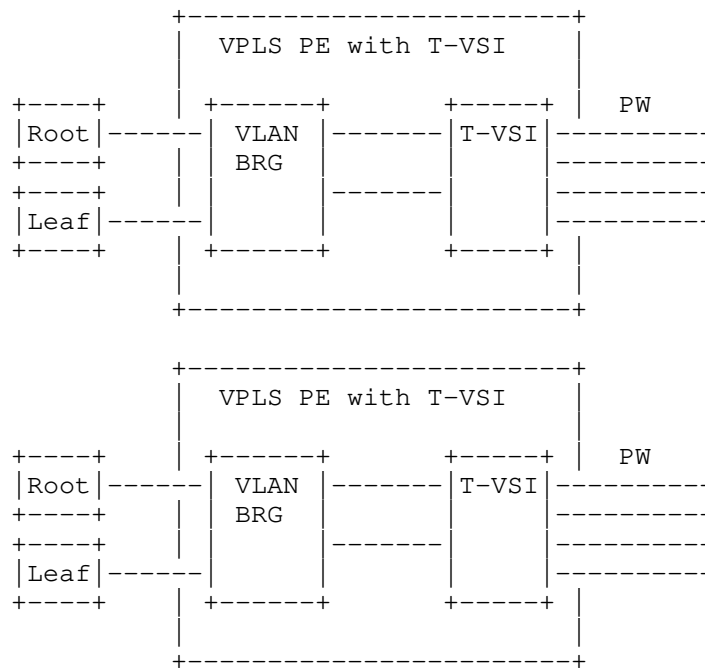


Figure 5 T-VSI Interconnected in the Normal Mode

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 [802.1Q-2011].

The signaling for VLANs is specified in Section 6.

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. A raw PW should be used to connect them.

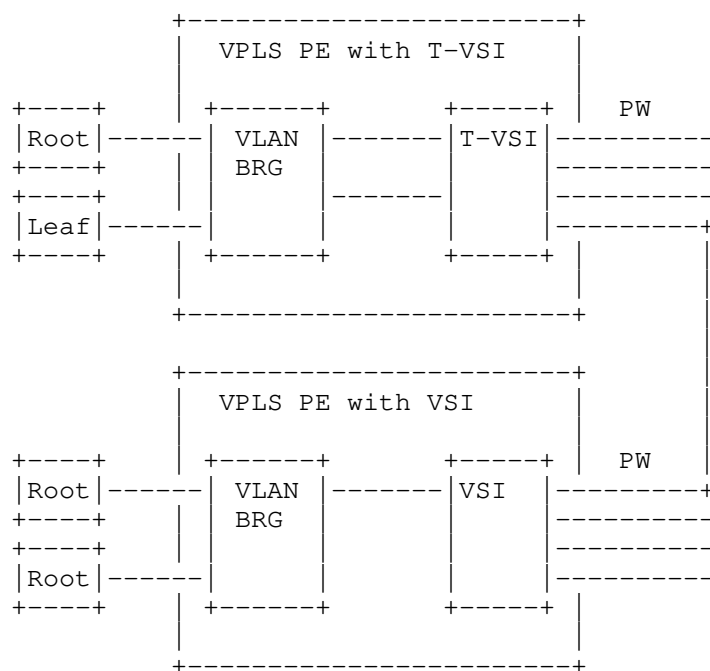


Figure 6 T-VSI interconnected with Traditional VSI

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

- o Upon transmitting frames on the PW, remove the root or leaf VLAN in the frames.
- o Upon receiving frames on the PW, add a VLAN tag with a value of the local root VLAN to the frames.

5.3.3.PW Processing in the Optimized Mode

When two PEs (both have E-Tree capability) are inter-connected and one of them (e.g., PE2) is attached with only leaf nodes, as shown in the scenario of Fig. 7, its peer PE (e.g., PE1) should then work in the optimized 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 leaf nodes is specified in Section 6.

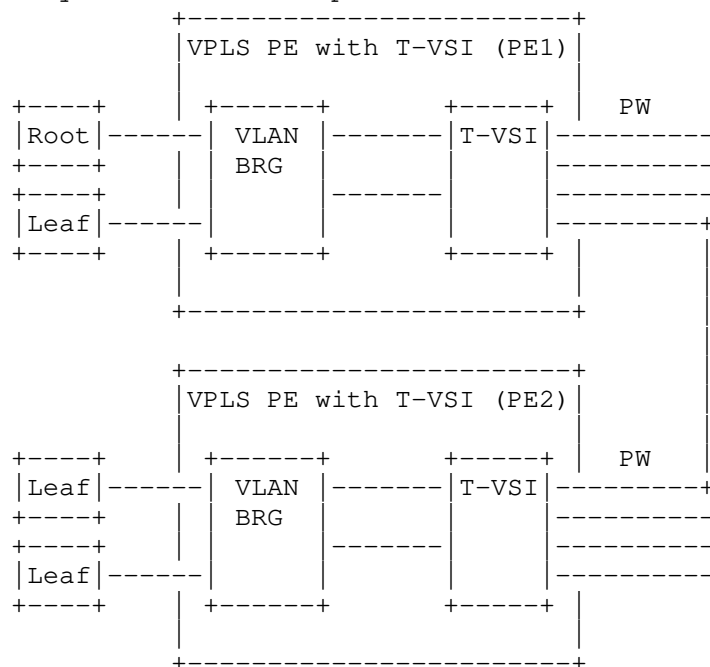


Figure 7 T-VSI interconnected with PE attached with only leaf nodes

If a PE is in the Optimized Mode for a PW, upon transmit, the PE SHOULD first operate as follows:

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

6. Signaling for E-Tree Support

6.1. 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:

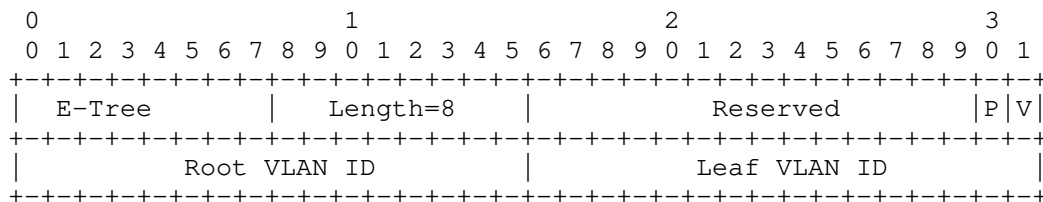


Figure 8 E-Tree Sub-TLV

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 leaf nodes, 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 leaf nodes 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, after saving the VLAN information for the PW, 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 a status code "Leaf to Leaf PW released" 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, The PE SHOULD then

establish a raw PW with this peer as in traditional VPLS and set Compatible-Mode to TRUE for this PW.

Data plane processing for this PW is as following:

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

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

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

PW processing as described in [RFC4448] proceeds as usual for all cases.

6.2. BGP Extensions for E-Tree Support

A new E-Tree extended community is proposed for E-Tree signaling in BGP VPLS:

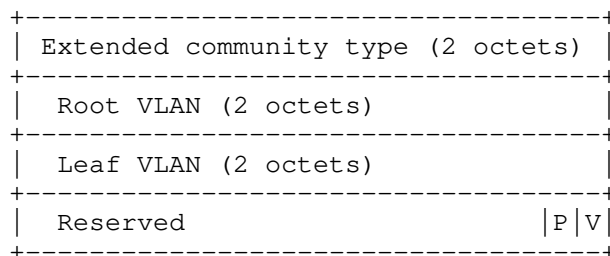


Figure 9 E-Tree Extended Community

Where:

- o Root VLAN ID is the value of the local root VLAN.
- o Leaf VLAN ID is the value of the local leaf VLAN.
- o Reserved, 14 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 leaf nodes, 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.

The PEs attached with both leaf and root nodes MUST support BGP E-Tree signaling as described in this document, and SHOULD support VLAN mapping in their data planes. The traditional PE attached with only root nodes may also participate in an E-Tree service. If some PEs don't support VLAN mapping, global VLANs as per Section 5.2 MUST be provisioned for an E-Tree service.

In BGP VPLS signaling, besides attaching a Layer2 Info Extended Community as detailed in [RFC4761], an E-Tree Extended Community MUST be further attached if a PE wishes to participate in an E-Tree service. The PE MUST include its local root VLAN ID and leaf VLAN ID in the E-Tree Extended Community. A PE attached with only leaf nodes of an E-Tree SHOULD set the P bit in the E-Tree Extended Community to 1.

A PE that receives a BGP UPDATE message with an E-Tree Extended Community from its peer PE, after saving the VLAN information for the PW, must process it as follows (after processing procedures as specified in Section 3.2 of [RFC4761]):

- 1) if the root and leaf VLAN ID in the E-Tree Extended Community 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 {
 - Log with a message "E-Tree VLAN mapping not supported" and exit the process;
 - if the bit V is set, and the PE is capable of VLAN mapping, the PE with the minimum IP address MUST set VLAN-Mapping-Mode to TRUE;
- }
- 3) If the P bit is set {

If the PE is a leaf-only PE itself, then forbids any traffic on the PW;

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

}

A PE which does not recognize this attribute shall ignore it silently. If a PE has sent an E-Tree Extended Community but does not receive any E-Tree Extended Community from its peer, the PE SHOULD then establish a raw PW with this peer as in traditional VPLS, and set Compatible-Mode to TRUE for this PW.

Data plane in the VPLS is the same as described in Section 4.2 of [RFC4761], and data plane processing for a PW is the same as described at the end of Section 6.1.

7. OAM Considerations

VPLS OAM requirements and framework as specified in [RFC6136] are applicable to E-Tree, as both Ethernet OAM frames and data traffic are transported over the same PW.

Ethernet OAM for E-Tree including both service OAM and segment OAM frames shall undergo the same VLAN mapping as the data traffic; and root VLAN SHOULD be applied to segment OAM frames so that they are not filtered.

8. Applicability

The solution is applicable to both LDP VPLS [RFC4762] and BGP VPLS [RFC4761].

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

The solution is also applicable to PBB VPLS networks.

9. Security Considerations

Besides security considerations as described in [RFC4448], [RFC4761] and [RFC4762], this solution prevents leaf to leaf communication in the data plane of VPLS when its PEs are interconnected with PWs. In this regard, security can be enhanced for customers with this solution.

10. IANA Considerations

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

Parameter ID	Length	Description
=====		
TBD	8	E-Tree

IANA is requested to allocate two new LDP status codes from the registry of name "STATUS CODE NAME SPACE". The following values are suggested:

Range/Value	E	Description

TBD	1	E-Tree VLAN mapping not supported
TBD	0	Leaf to Leaf PW released

IANA is requested to allocate a value for E-Tree in the registry of BGP Extended Community.

Type Value	Name
=====	
TBD	E-Tree Info

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.
- [RFC4761] Kompella, K. and Rekhter, Y., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007

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

[RFC6136] Sajassi, A. and Mohan, D., "L2VPN OAM Requirements and Framework", RFC 6136, March 2011

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-ietf-l2vpn-etree-req-05, Work in Progress

[ETree-frwk] Key, R., and et al, "A Framework for E-Tree Service over MPLS Network", draft-ietf-l2vpn-etree-frwk-03, Work in Progress

[802.1Q-2011] IEEE 802.1Q, Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks, August 2011

[PBB-VPLS] Balus, F., and et al., Extensions to VPLS PE model for Provider Backbone Bridging, draft-ietf-l2vpn-pbb-vpls-pe-model-07, Work in Progress

12. Acknowledgments

The authors would like to thank Adrian Farrel, Susan Hares and Shane Amante for their valuable advices, thank Ben Mack-crane, Edwin Mallette, Donald Fedyk, Dave Allan, Giles Heron, Raymond Key, Josh Rogers, Sam Cao and Daniel Cohn for their valuable comments and discussions.

Appendix A. Other PE Models for E-Tree

A.1. A 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.

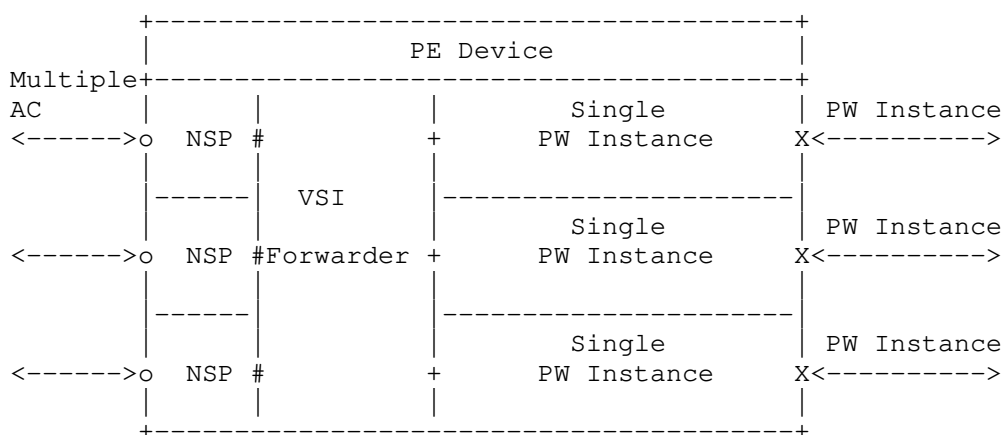


Figure A.1 A PE model with a VSI and no bridge module

This PE model may be used by an MTU-s in an H-VPLS network, or an N-PE in an H-VPLS network with non-bridging edge devices, wherein a spoke PW can be treated as an AC in this model.

A.2. A PE Model With external E-Tree interface

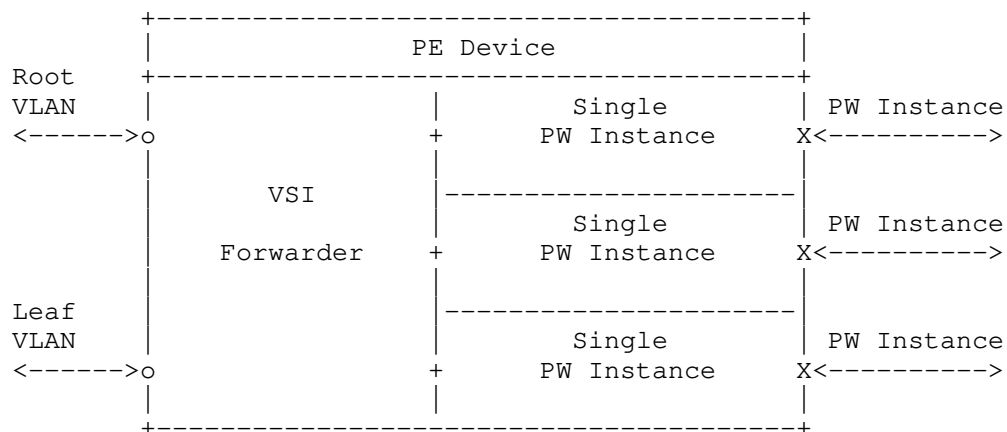


Figure A.2 A PE model with external E-Tree interface

A more simplified PE model is depicted in A.2, where Root/Leaf VLANs are directly or indirectly over a single PW connected to a same VSI forwarder in a PE, any transformation of E-Tree VLANs, e.g., VLAN insertion/removal or VLAN mapping, can be performed by some outer equipments, and the PE may further translate these VLANs into its own local VLANs. This PE model may be used by an N-PE in an H-VPLS network with bridging-capable devices, or scenarios such as providing E-Tree Network-to-Network (NNI) interfaces.

Authors' Addresses

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

Lucy Yong
Huawei USA
207 Estrella Xing
Georgetown TX, USA 78628
Email: lucyyong@huawei.com

Manuel Paul
Deutsche Telekom
Winterfeldtstr. 21
10781 Berlin, Germany
Email: manuel.paul@telekom.de

Frederic Jounay
Orange CH
4 rue caudray 1020 Renens, Switzerland
Email: frederic.jounay@orange.ch

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

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp, Belgium
Email: wim.henderickx@alcatel-lucent.com

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

Layer 2 Virtual Private Networks
Internet-Draft
Intended status: Informational
Expires: April 11, 2014

O. Dornon
J. Kotalwar
Alcatel-Lucent
V. Hemige

R. Qiu
Z. Zhang
Juniper Networks, Inc.
October 8, 2013

PIM Snooping over VPLS
draft-ietf-l2vpn-vpls-pim-snooping-05

Abstract

This document describes the procedures and recommendations for VPLS PEs to facilitate replication of multicast traffic to only certain ports (behind which there are interested PIM routers and/or IGMP hosts) via PIM Snooping and PIM Proxy.

With PIM Snooping, PEs passively listen to certain PIM control messages to build control and forwarding states while transparently flooding those messages. With PIM Proxy, PEs do not flood PIM Join/Prune messages but only generate their own and send out of certain ports, based on the control states built from downstream Join/Prune messages. PIM Proxy is required when PIM Join suppression is enabled on the CE devices and useful to reduce PIM control traffic in a VPLS domain.

The document also describes PIM Relay, which can be viewed as light-weight proxy, where all downstream Join/Prune messages are simply forwarded out of certain ports but not flooded to avoid triggering PIM Join suppression on CE devices.

Requirements Language

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].

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 April 11, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	5
1.1. Multicast Snooping in VPLS	5
1.2. Assumptions	6
1.3. Definitions	7
2. PIM Snooping for VPLS	7
2.1. PIM protocol background	7
2.2. General Rules for PIM Snooping in VPLS	8
2.2.1. Preserving Assert Trigger	8
2.3. Some Considerations for PIM Snooping	9
2.3.1. Scaling	9
2.3.2. IPv6	10
2.3.3. PIM-SM (*,*,RP)	10
2.4. PIM Snooping vs PIM Proxy	10
2.4.1. Differences between PIM Snooping, Relay and Proxy	10
2.4.2. PIM Control Message Latency	11
2.4.3. When to Snoop and When to Proxy	12
2.5. Discovering PIM Routers	13
2.6. PIM-SM and PIM-SSM	14
2.6.1. Building PIM-SM Snooping States	14
2.6.2. Explanation for per (S,G,N) states	17
2.6.3. Receiving (*,G) PIM-SM Join/Prune Messages	17
2.6.4. Receiving (S,G) PIM-SM Join/Prune Messages	19
2.6.5. Receiving (S,G,rpt) Join/Prune Messages	21
2.6.6. Sending Join/Prune Messages Upstream	21
2.7. Bidirectional-PIM (PIM-BIDIR)	22
2.8. Interaction with IGMP Snooping	23
2.9. PIM-DM	23
2.9.1. Building PIM-DM Snooping States	23
2.9.2. PIM-DM Downstream Per-Port PIM(S,G,N) State Machine	24
2.9.3. Triggering ASSERT election in PIM-DM	24
2.10. PIM Proxy	24
2.10.1. Upstream PIM Proxy behavior	24
2.11. Directly Connected Multicast Source	25
2.12. Data Forwarding Rules	25
2.12.1. PIM-SM Data Forwarding Rules	26
2.12.2. PIM-DM Data Forwarding Rules	27
3. IANA Considerations	28
4. Security Considerations	28
5. Contributors	28
6. Acknowledgements	29
7. References	29
7.1. Normative References	29
7.2. Informative References	29
Appendix A. PIM-BIDIR Thoughts	30
A.1. PIM-BIDIR Data Forwarding Rules	30
Appendix B. Example Network Scenario	31

B.1. Pim Snooping Example	32
B.2. PIM Proxy Example with (S,G) / (*,G) interaction	34
Authors' Addresses	40

1. Introduction

In Virtual Private LAN Service (VPLS), the Provider Edge (PE) devices provide a logical interconnect such that Customer Edge (CE) devices belonging to a specific VPLS instance appear to be connected by a single LAN. Forwarding Information Base for a VPLS instance is populated dynamically by source MAC address learning. Once a unicast MAC address is learned and associated with a particular Attachment Circuit (AC) or PseudoWire (PW), a frame destined to that MAC address only needs to be sent on that AC or PW.

For a frame not addressed to a known unicast MAC address, flooding has to be used. This happens with the following so called BUM traffic:

- o B: The destination MAC address is a broadcast address,
- o U: The destination MAC address is unknown (has not been learned),
- o M: The destination MAC address is a multicast address.

Multicast frames are flooded because a PE cannot know where multicast members reside. VPLS solutions (i.e., [VPLS-LDP] and [VPLS-BGP]) perform replication for multicast traffic at the ingress PE devices. As stated in the VPLS Multicast Requirements draft [VPLS-MCAST-REQ], there are two issues with VPLS Multicast today:

- o A. Multicast traffic is replicated to non-member sites.
- o B. Replication of PWs on shared physical path.

Issue A can be solved by Multicast Snooping - PEs learn sites with multicast members by snooping multicast protocol control messages and forward IP multicast traffic only to member sites. This document describes the procedures to achieve that when PIM is running between the CE devices. Issue B is outside the scope of this document and discussed in [VPLS-MCAST-TREES].

While this document is in the context of VPLS, the procedures apply to regular layer-2 switches interconnected by physical connections as well. In that case, the PW related concept/procedures are not applicable and that's all.

1.1. Multicast Snooping in VPLS

IGMP Snooping procedures described in [IGMP-SNOOP] make sure that IP multicast traffic is only sent out of the following:

- o Attachment Circuits (ACs) connecting to hosts that report related group membership
- o ACs connecting to routers
- o PseudoWires (PWs) connecting to remote PEs that have the above described ACs

Notice that traffic is always sent out of ports connecting to routers, even those on which there are no snooped group memberships, because IGMP Snooping alone can not determine if there are interested receivers beyond those routers. To further restrict traffic sent to those routers, PIM Snooping can be used, and this document describes the procedures, including the rules when both IGMP and PIM are active in a VPLS instance.

Note that for both IGMP and PIM, the term Snooping is used loosely, referring to the fact that a layer-2 device peeks into layer-3 routing protocol messages to build relevant control and forwarding states. Depending on how the control messages are handled (transparently flooded, selectively forwarded, or consumed and then regenerated), the procedure/process may be called Snooping or Proxy in different contexts.

Unless explicitly noted, the procedures in this document are used for either PIM Snooping or PIM Proxy, and we will largely refer to PIM "Snooping" in this document. The PIM Proxy specific procedures are described in Section 2.6.6. Differences that need to be observed while implementing one or the other and recommendations on which method to employ in different scenarios are noted in section Section 2.4.

This document also describes PIM Relay, which can be viewed as light-weight Proxy. Unless explicitly noted, in the rest of the document Proxy implicitly includes Relay as well.

1.2. Assumptions

The document assumes that the reader has a good understanding of the PIM protocols. The text in this draft is written in the same style as the PIM RFCs to help correlate the concepts and to make it easier to follow. In order to avoid replicating the text relating to PIM protocol handling here, this draft cross references into definitions of macros and procedures from the PIM RFCs, and assumes that the user will infer such detail from those PIM RFCs. Deviations in protocol handling specific to PIM Snooping are specified in this draft.

1.3. Definitions

There are several definitions referenced in this document that are well described in the PIM RFCs [PIM-SM], PIM-BIDIR, PIM-DM]. The following definitions and abbreviations are used throughout this document:

- o A port is defined as either an attachment circuit (AC) or a Pseudo-Wire (PW).
- o When we say a PIM message is 'received' on a port, it means that a PIM Snooping PE snooped the PIM message.

Abbreviations used in the document:

- o S: IP Address of the Multicast Source.
- o G: IP Address of the Multicast Group.
- o N: Upstream Neighbor field in a Join/Prune/Graft message.
- o Rport(N): Port on which neighbor N is learnt.

Other definitions are explained in the sections where they are introduced.

2. PIM Snooping for VPLS

2.1. PIM protocol background

PIM is a multicast routing protocol running between routers, which are CE devices in a VPLS. PIM shares many of the common characteristics of a routing protocol, such as discovery messages (e.g., neighbor discovery using Hello messages), topology information (e.g., multicast tree), and error detection and notification (e.g., dead timer and designated router election). PIM does not participate in exchange of unicast routing databases, but it uses the unicast routing table to provide reverse path information for building multicast trees. There are a few variants of PIM. In [PIM-DM], multicast data is pushed towards the members similar to broadcast mechanism but routers without attached receivers will prune back towards the source. Unlike PIM-DM, other PIM flavors (PIM-SM [PIM-SM], PIM-SSM [PIM-SSM], and PIM-BIDIR [PIM-BIDIR]) employs a pull methodology via explicit joins instead of push technique.

PIM routers periodically exchange Hello messages to discover and maintain stateful sessions with neighbors. After neighbors are

discovered, PIM routers can signal their intentions to join or prune specific multicast groups. This is accomplished by having downstream routers send an explicit Join/Prune message (for the sake of generalization, consider Graft messages for PIM-DM as Join messages) to the upstream routers. The Join/Prune message can be group specific (*,G) or group and source specific (S,G).

2.2. General Rules for PIM Snooping in VPLS

The following rules for the correct operation of PIM snooping MUST be followed.

- o PIM Snooping MUST NOT affect the operation of customer layer-2 protocols (e.g., BPDUs) or layer-3 protocols.
- o PIM messages and multicast data traffic forwarded by PEs MUST follow the split-horizon rule for mesh PWs.
- o PIM snooping states in a PE MUST be per VPLS instance.
- o PIM assert triggers MUST be preserved to the extent necessary to avoid sending duplicate traffic to the same PE (see Section 2.2.1).

2.2.1. Preserving Assert Trigger

In PIM-SM/DM, there are scenarios where multiple routers could be forwarding the same multicast traffic on a LAN. When this happens, using PIM Assert Election process by sending PIM Assert Messages, routers ensure that only the Assert Winner forwards traffic on the LAN. The Assert Election is a data driven event and happens only if a router sees traffic on the interface to which it should be forwarding the traffic. In the case of VPLS with snooping, two routers may forward the same flow at the same time but each copy may reach different set of PEs, and that is acceptable from the point of view of avoiding duplicate traffic. If the two copies may reach the same PE then the sending routers must be able to see each other's traffic, in order to trigger Assert Election and stop duplicate traffic.

To achieve that, PIM-SM Snooping MUST not only forward multicast traffic for an (S,G) on the ports on which they snooped Joins(S,G)/ Joins(*,G), but also towards the upstream neighbor(s)). In other words, the ports on which the upstream neighbors are learnt must be added to the outgoing port list along with the ports on which Joins are snooped.

Similarly, PIM-DM Snooping SHOULD make sure that asserts can be

triggered (Section 2.9.3).

The above logic needs to be facilitated without breaking VPLS Split Horizon Rules. i.e. traffic should not be forwarded on the port on which it was received, and traffic arriving on a PW MUST NOT be forwarded onto other PW(s).

2.3. Some Considerations for PIM Snooping

The PIM Snooping solution described here requires a PE to examine and operate on only PIM Hello and PIM Join/Prune packets. The PE does not need to examine any other PIM packets.

Most of the procedures in PIM Snooping in the handling of PIM Hellos and PIM Join/Prune packets are very similar to that of a PIM Router.

However, the PE does not need to have any routing tables like is required in PIM Multicast Routing. It knows how to forward Join/Prunes by looking at the Upstream Neighbor field in the Join/Prune packets.

The PE does not need to know about Rendezvous Points (RP) and does not have to maintain any RP Set. All that is transparent to a PIM Snooping PE.

In the following sub-sections, we list some considerations and observations for the implementation of PIM Snooping in VPLS.

2.3.1. Scaling

Snooping needs to be employed on ACs at the downstream PEs to prevent traffic from being sent out of ACs unnecessarily. Snooping techniques can also be employed on PWs at the upstream PEs to prevent traffic from being sent to PEs unnecessarily. This may work well for small to medium scale deployments. However, if there are a large number of VPLS instances with a large number of PEs per instances, then the amount of snooping required at the upstream PEs can overwhelm the upstream PEs.

There are two methods to reduce the burden on the upstream PEs. One is to use PIM Proxy as described in Section 2.6.6, to reduce the control messages forwarded by a PE. The other is not to snoop on the PWs at all, but PEs signal the snooped states to other PEs out of band via BGP, as described in [VPLS-MCAST-TREES]. In this document, it is assumed that Snooping is performed on PWs.

2.3.2. IPv6

In VPLS, PEs forward Ethernet frames received from CEs and as such are agnostic of the layer-3 protocol used by the CEs. However, as an IGMP and PIM snooping PE, the PE would have to look deeper into the IP and IGMP/PIM packets and build snooping state based on that. The PIM Protocol specifications handle both IPv4 and IPv6. The specification for PIM Snooping in this draft can be applied to both IPv4 and IPv6 payloads.

2.3.3. PIM-SM (*,*,RP)

This draft does not address (*,*,RP) states in the VPLS network. Although [PIM-SM] specifies that routers MUST support (*,*,RP) states, there are very few implementations that actually support it in actual deployments, and it is being removed from the PIM protocol in its ongoing advancement process in IETF. Given that, this draft omits the specification relating to (*,*,RP) support.

2.4. PIM Snooping vs PIM Proxy

The document has previously alluded to PIM Snooping/Relay/Proxy. Details on the PIM Proxy/Relay solution are discussed in Section 2.6.6. In this section, a brief description and comparison are given.

2.4.1. Differences between PIM Snooping, Relay and Proxy

Differences between PIM Snooping and Proxy/Relay can be summarized as the following:

PIM Snooping	PIM Relay	PIM Proxy
Join/Prune messages snooped and flooded everywhere	Join/Prune messages snooped; forwarded as is out of certain upstream ports	Join/Prune messages consumed. Regenerated ones sent out of certain upstream ports
No PIM packets generated.	No PIM packets generated	New Join/Prune messages generated
CE Join Suppression not allowed	CE Join Suppression allowed	CE Join Suppression allowed

Note that the differences apply only to PIM Join/Prune messages. PIM

Hello messages are snooped and flooded in all cases.

Other than the above differences, most of the procedures are common to PIM Snooping and PIM Proxy/Relay, unless specifically stated otherwise.

Pure PIM Snooping PEs simply snoop on PIM packets as they are being forwarded in the VPLS. As such they truly provide transparent LAN services since no customer packets are modified or consumed or new packets introduced in the VPLS. It is also simpler to implement than PIM Proxy. However for PIM Snooping to work correctly, it is a requirement that CE routers MUST disable Join suppression in the VPLS.

Given that a large number of existing CE deployments do not support disabling of Join suppression and given the operational complexity for a provider to manage disabling of Join suppression in the VPLS, it becomes a difficult solution to deploy. Another disadvantage of PIM Snooping is that it does not scale as well as PIM Proxy. If there are a large number of CEs in a VPLS, then every CE will see every other CE's Join/Prune messages.

PIM Proxy/Relay has the advantage that it does not require Join suppression to be disabled in the VPLS. Multicast as a VPLS service can be very easily provided without requiring any changes on the CE routers. PIM Proxy/Relay helps scale VPLS Multicast since Join/Prune messages are only sent to certain upstream ports instead of flooded, and in case of full Proxy (vs. Relay) the PEs intelligently generate only one Join/Prune message for a given flow.

PIM Proxy however loses the transparency argument since Join/Prunes could get modified or even consumed at a PE. Also, new packets could get introduced in the VPLS. However, this loss of transparency is limited to PIM Join/Prune packets. It is in the interest of optimizing multicast in the VPLS and helping a VPLS network scale much better. Data traffic will still be completely transparent.

2.4.2. PIM Control Message Latency

A PIM Snooping/Proxy/Relay PE snoops on PIM Hello packets while transparently flooding them in the VPLS. As such there is no latency introduced by the VPLS in the delivery of PIM Hello packets to remote CEs in the VPLS.

A PIM Snooping PE snoops on PIM Join/Prune packets while transparently flooding them in the VPLS. There is no latency introduced by the VPLS in the delivery of PIM Join/Prune packets when PIM Snooping is employed.

A PIM Proxy/Relay PE does not simply flood PIM Join/Prune packets. This can result in additional latency for a downstream CE to receive multicast traffic after it has sent a Join. When a downstream CE prunes a multicast stream, the traffic should stop flowing to the CE with no additional latency introduced by the VPLS.

Performing only proxy of Join/Prune and not Hello messages keeps the PE behavior very similar to that of a PIM router without introducing too much additional complexity. It keeps the PIM Proxy solution fairly simple. Since Join/Prunes are forwarded by a PE along the slow-path and all other PIM packet types are forwarded along the fast-path, it is very likely that packets forwarded along the fast-path will arrive "ahead" of Join/Prune packets at a CE router (note the stress on the fact that fast-path messages will never arrive after Join/Prunes). Of particular importance are Hello packets sent along the fast-path. We can construct a variety of scenarios resulting in out of order delivery of Hellos and Join/Prune messages. However, there should be no deviation from normal expected behavior observed at the CE router receiving these messages out of order.

2.4.3. When to Snoop and When to Proxy

From the above descriptions, factors that affect the choice of Snooping/Relay/Proxy include:

- o Whether CEs do Join Suppression or not
- o Whether Join/Prune latency is critical or not
- o Whether the scale of PIM protocol message/states in a VPLS requires the scaling benefit of Proxy

Of the above factors, Join Suppression is the hard one - pure Snooping can only be used when Join Suppression is disabled on all CEs. The latency associated with Relay/Proxy is implementation dependent and may not be a concern at all with a particular implementation. The scaling benefit may not be important either, in that on a real LAN with Explicit Tracking (ET) a PIM router will need to receive and process all PIM Join/Prune messages as well.

A PIM router indicates that Join Suppression is disabled if the T-bit is set in the LAN Prune Delay option of its Hello message. If all PIM routers on a LAN set the T-bit, Explicit Tracking is possible, allowing an upstream router to track all the downstream neighbors that have Join states for any (S,G) or (*,G). That has two benefits:

- o No need for PrunePending process - the upstream router may immediately stop forwarding data when it receives a Prune from the last downstream neighbor, and immediately prune to its upstream if that's for the last downstream interface.
- o For management purpose, the upstream router knows exactly which downstream routers exist for a particular Join State.

While full Proxy can be used with or without Join Suppression on CEs and does not interfere with an upstream CE's bypass of PrunePending process, it does proxy all its downstream CEs as a single one to the upstream, removing the second benefit mentioned above.

Therefore, the general rule is that if Join Suppression is enabled on CEs then Proxy or Relay MUST be used and if Suppression is known to be disabled on all CEs then either Snooping, Relay, or Proxy MAY be used while Snooping or Relay SHOULD be used.

An implementation MAY choose dynamic determination of which mode to use, through the tracking of the above mentioned T-bit in all snooped PIM Hello messages, or MAY simply require static provisioning.

2.5. Discovering PIM Routers

A PIM Snooping PE MUST snoop on PIM Hellos received on ACs and PWs. i.e. the PE transparently floods the PIM Hello while snooping on it. PIM Hellos are used by the snooping PE to discover PIM routers and their characteristics.

For each neighbor discovered by a PE, it includes an entry in the PIM Neighbor Database with the following fields:

- o Layer 2 encapsulation for the Router sending the PIM Hello.
- o IP Address and address family of the Router sending the PIM Hello.
- o Port (AC / PW) on which the PIM Hello was received.
- o Hello TLVs

The PE should be able to interpret and act on Hello TLVs currently defined in the PIM RFCs. The TLVs of particular interest in this document are:

- o Hello-Hold-Time
- o Tracking Support

- o DR Priority

Please refer to [PIM-SM] for a list of the Hello TLVs. When a PIM Hello is received, the PE MUST reset the neighbor-expiry-timer to Hello-Hold-Time. If a PE does not receive a Hello message from a router within Hello-Hold-Time, the PE MUST remove that neighbor from its PIM Neighbor Database. If a PE receives a Hello message from a router with Hello-Hold-Time value set to zero, the PE MUST remove that router from the PIM snooping state immediately.

From the PIM Neighbor Database, a PE MUST be able to use the procedures defined in [PIM-SM] to identify the PIM Designated Router in the VPLS instance. It should also be able to determine if Tracking Support is active in the VPLS instance.

2.6. PIM-SM and PIM-SSM

The key characteristic of PIM-SM and PIM-SSM is explicit join behavior. In this model, multicast traffic is only forwarded to locations that specifically request it. The root node of a tree is the Rendezvous Point (RP) in case of a shared tree (PIM-SM only) or the first hop router that is directly connected to the multicast source in the case of a shortest path tree. All the procedures described in this section apply to both PIM-SM and PIM-SSM, except for the fact that there is no (*,G) state in PIM-SSM.

2.6.1. Building PIM-SM Snooping States

PIM-SM and PIM-SSM Snooping states are built by snooping on the PIM-SM Join/Prune messages received on AC/PWs.

The downstream state machine of a PIM-SM snooping PE very closely resembles the downstream state machine of PIM-SM routers. The downstream state consists of:

Per downstream (Port, *, G):

- o DownstreamJPState: One of { "NoInfo" (NI), "Join" (J), "Prune Pending" (PP) }

Per downstream (Port, *, G, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Per downstream (Port, S, G):

- o DownstreamJPState: One of { "NoInfo" (NI), "Join" (J), "Prune Pending" (PP) }

Per downstream (Port, S, G, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Per downstream (Port, S, G, rpt):

- o DownstreamJPRptState: One of { "NoInfo" (NI), "Pruned" (P), "Prune Pending" (PP) }

Per downstream (Port, S, G, rpt, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Where S is the address of the multicast source, G is the Group address and N is the upstream neighbor field in the Join/Prune message. Notice that unlike on PIM-SM routers where PPT and ET are per (Interface, S, G), PIM Snooping PEs have to maintain PPT and ET per (Port, S, G, N). The reasons for this are explained in Section 2.6.2.

Apart from the above states, we define the following state summarization macros.

UpstreamNeighbors(*,G): If there is one or more Join(*,G) received on any port with upstream neighbor N and ET(N) is active, then N is added to UpstreamNeighbors(*,G). This set is used to determine if a Join(*,G) or a Prune(*,G) with upstream neighbor N needs to be sent upstream.

UpstreamNeighbors(S,G): If there is one or more Join(S,G) received on any port with upstream neighbor N and ET(N) is active, then N is added to UpstreamNeighbors(S,G). This set is used to determine if a Join(S,G) or a Prune(S,G) with upstream neighbor N needs to be sent upstream.

UpstreamPorts(*,G): This is the set of all Rport(N) ports where N is in the set UpstreamNeighbors(*,G). Multicast Streams forwarded using a (*,G) match MUST be forwarded to these ports in addition to downstream ports. So UpstreamPorts(*,G) MUST be added to OutgoingPortList(*,G).

UpstreamPorts(S,G): This is the set of all Rport(N) ports where N is in the set UpstreamNeighbors(S,G). UpstreamPorts(S,G) MUST be added to OutgoingPortList(S,G).

InheritedUpstreamPorts(S,G): This is the union of UpstreamPorts(S,G) and UpstreamPorts(*,G).

UpstreamPorts(S,G,rpt): If PruneDesired(S,G,rpt) becomes true, then this set is set to UpstreamPorts(*,G). Otherwise, this set is empty. UpstreamPorts(*,G) (-) UpstreamPorts(S,G,rpt) MUST be added to OutgoingPortList(S,G).

UpstreamPorts(G): This set is the union of all the UpstreamPorts(S,G) and UpstreamPorts(*,G) for a given G. Proxy (S,G) Join/Prune and (*,G) Join/Prune messages MUST be sent to a subset of UpstreamPorts(G) as specified in Section 2.6.6.1.

PWPorts: This is the set of all PWs.

OutgoingPortList(*,G): This is the set of all ports to which traffic needs to be forwarded on a (*,G) match.

OutgoingPortList(S,G): This is the set of all ports to which traffic needs to be forwarded on an (S,G) match.

See Section 2.12 on Data Forwarding Rules for the specification on how OutgoingPortList is calculated.

NumETsActive(Port,*,G): Number of (Port,*,G,N) entries that have Expiry Timer running. This macro keeps track of the number of Join(*,G)s that are received on this Port with different upstream neighbors.

NumETsActive(Port,S,G): Number of (Port,S,G,N) entries that have Expiry Timer running. This macro keeps track of the number of Join(S,G)s that are received on this Port with different upstream neighbors.

RpfVectorTlvs(*,G): RPF Vectors [RPF-VECTOR] are TLVs that may be present in received Join(*,G) messages. If present, they must be copied to RpfVectorTlvs(*,G).

RpfVectorTlvs(S,G): RPF Vectors [RPF-VECTOR] are TLVs that may be present in received Join(S,G) messages. If present, they must be copied to RpfVectorTlvs(S,G).

Since there are a few differences between the downstream state machines of PIM-SM Routers and PIM-SM snooping PEs, we specify the

details of the downstream state machine of PIM-SM snooping PEs at the risk of repeating most of the text documented in [PIM-SM].

2.6.2. Explanation for per (S,G,N) states

In PIM Routing protocols, states are built per (S,G). On a router, an (S,G) has only one RPF-Neighbor. However, a PIM Snooping PE does not have the Layer 3 routing information available to the routers in order to determine the RPF-Neighbor for a multicast flow. It merely discovers it by snooping the Join/Prune message. A PE could have snooped on two or more different Join/Prune messages for the same (S,G) that could have carried different Upstream-Neighbor fields. This could happen during transient network conditions or due to dual-homed sources. A PE cannot make assumptions on which one to pick, but instead must facilitate the CE routers decide which Upstream Neighbor gets elected the RPF-Neighbor. And for this purpose, the PE will have to track downstream and upstream Join/Prune per (S,G,N).

2.6.3. Receiving (*,G) PIM-SM Join/Prune Messages

A Join(*,G) or Prune(*,G) is considered "received" if the following conditions are met:

- o The port on which it arrived is not Rport(N) where N is the upstream-neighbor N of the Join/Prune(*,G), or,
- o if both RPort(N) and the arrival port are PWs, then there exists at least one other (*,G,Nx) or (Sx,G,Nx) state with an AC UpstreamPort.

For simplicity, the case where both RPort(N) and the arrival port are PWs is referred to as PW-only Join/Prune in this document. The PW-only Join/Prune handling is so that the RPort(N) PW can be added to the related forwarding entries' OutgoingPortList to trigger Assert, but that is only needed for those states with AC UpstreamPort. Note that in PW-only case, it is OK for the arrival port and RPort(N) to be the same. See Appendix Appendix B for examples.

When a router receives a Join(*,G) or a Prune(*,G) with upstream neighbor N, it must process the message as defined in the state machine below. Note that the macro computations of the various macros resulting from this state machine transition is exactly as specified in the PIM-SM RFC [PIM-SM].

We define the following per-port (*,G,N) macro to help with the state machine below.

Figure 1 : Downstream per-port (*,G) state machine in tabular form

Event	Previous State		
	NoInfo (NI)	Join (J)	Prune-Pend
Receive Join(*,G)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)
Receive Prune(*,G) and NumETsActive<=1	-	-> PP state Start PPT(N)	-> PP state
Receive Prune(*,G) and NumETsActive>1	-	-> J state Start PPT(N)	-
PPT(N) expires	-	-> J state Action PPTEpiry(N)	-> NI state Action PPTEpiry(N)
ET(N) expires and NumETsActive<=1	-	-> NI state Action ETExpiry(N)	-> NI state Action ETExpiry(N)
ET(N) expires and NumETsActive>1	-	-> J state Action ETExpiry(N)	-

Action RxJoin(N) :

If ET(N) is not already running, then start ET(N). Otherwise restart ET(N). If N is not already in UpstreamNeighbors(*,G), then add N to UpstreamNeighbors(*,G) and trigger a Join(*,G) with upstream neighbor N to be forwarded upstream. If there are RPF Vector TLVs in the received (*,G) message and if they are different from the recorded RpfVectorTlvs(*,G), then copy them into RpfVectorTlvs(*,G).

Action PPTEpiry(N) :

Same as Action ETExpiry(N) below, plus Send a Prune-Echo(*,G) with upstream-neighbor N on the downstream port.

Action ETExpiry(N) :

Disable timers ET(N) and PPT(N). Delete neighbor state (Port,*,G,N). If there are no other (Port,*,G) states with NumETsActive(Port,*,G) > 0, transition DownstreamJPState to NoInfo. If there are no other (Port,*,G,N) state (different ports but for the same N), remove N from UpstreamPorts(*,G) - this also serves as a trigger for US FSM (JoinDesired(*,G,N) becomes FALSE).

2.6.4. Receiving (S,G) PIM-SM Join/Prune Messages

A Join(S,G) or Prune(S,G) is considered "received" if the following conditions are met:

- o The port on which it arrived is not Rport(N) where N is the upstream-neighbor N of the Join/Prune(S,G), or,
- o if both RPort(N) and the arrival port are PWs, then there exists at least one other (*,G,Nx) or (S,G,Nx) state with an AC UpstreamPort.

For simplicity, the case where both RPort(N) and the arrival port are PWs is referred to as PW-only Join/Prune in this document. The PW-only Join/Prune handling is so that the RPort(N) PW can be added to the related forwarding entries' OutgoingPortList to trigger Assert, but that is only needed for those states with AC UpstreamPort. See Appendix Appendix B for examples.

When a router receives a Join(S,G) or a Prune(S,G) with upstream neighbor N, it must process the message as defined in the state machine below. Note that the macro computations of the various macros resulting from this state machine transition is exactly as specified in the PIM-SM RFC [PIM-SM].

Figure 2: Downstream per-port (S,G) state machine in tabular form

Event	Previous State		
	NoInfo (NI)	Join (J)	Prune-Pend
Receive Join (S,G)	-> J state Action RxJoin (N)	-> J state Action RxJoin (N)	-> J state Action RxJoin (N)
Receive Prune (S,G) and NumETsActive<=1	-	-> PP state Start PPT(N)	-> PP state
Receive Prune (S,G) and NumETsActive>1	-	-> J state Start PPT(N)	-
PPT(N) expires	-	-> J state Action PPTEpiry (N)	-> NI state Action PPTEpiry (N)
ET(N) expires and NumETsActive<=1	-	-> NI state Action ETExpiry (N)	-> NI state Action ETExpiry (N)
ET(N) expires and NumETsActive>1	-	-> J state Action ETExpiry (N)	-

Action RxJoin(N):

If ET(N) is not already running, then start ET(N). Otherwise, restart ET(N).

If N is not already in UpstreamNeighbors(S,G), then add N to UpstreamNeighbors(S,G) and trigger a Join(S,G) with upstream neighbor N to be forwarded upstream. If there are RPF Vector TLVs in the received (S,G) message and if they are different from the recorded RpfVectorTlvs(S,G), then copy them into RpfVectorTlvs(S,G).

Action PPTEpiry(N):

Same as Action ETExpiry(N) below, plus Send a Prune-Echo(S,G) with upstream-neighbor N on the downstream port.

Action ETEpiry(N):

Disable timers ET(N) and PPT(N). Delete neighbor state (Port,S,G,N). If there are no other (Port,S,G) states with NumETsActive(Port,S,G) > 0, transition DownstreamJPState to NoInfo. If there are no other (Port,S,G,N) state (different ports but for the same N), remove N from UpstreamPorts(S,G) - this also serves as a trigger for US FSM (JoinDesired(S,G,N) becomes FALSE).

2.6.5. Receiving (S,G,rpt) Join/Prune Messages

A Join(S,G,rpt) or Prune(S,G,rpt) is "received" when the port on which it was received is not also the port on which the upstream-neighbor N of the Join/Prune(S,G,rpt) was learnt.

While it is important to ensure that the (S,G) and (*,G) state machines allow for handling per (S,G,N) states, it is not as important for (S,G,rpt) states. It suffices to say that the downstream (S,G,rpt) state machine is the same as what is defined in section 4.5.4 of the PIM-SM RFC [PIM-SM].

2.6.6. Sending Join/Prune Messages Upstream

This section applies only to a PIM Proxy/Relay PE and not to a PIM Snooping PE.

A full PIM Proxy (not Relay) PE MUST implement the Upstream FSM for which the procedures are similar to what is defined in section 4.5.6 of [PIM-SM].

For the purposes of the Upstream FSM, a Join or Prune message with upstream neighbor N is "seen" on a PIM Snooping PE if the port on which the message was received is also Rport(N), and the port is an AC. The AC requirement is needed because a Join received on the Rport(N) PW must not suppress this PE's Join on that PW.

A PIM Relay PE does not implement the Upstream FSM. It simply forwards received Join/Prune messages out of the same set of upstream ports as in the PIM Proxy case.

In order to correctly facilitate assert among the CE routers, such Join/Prunes need to sent not only towards the upstream neighbor, but also on certain PWs as described below.

If RpfVectorTlvs(*,G) is not empty, then it must be encoded in a Join(*,G) message sent upstream.

If RpfVectorTlvs(S,G) is not empty, then it must be encoded in a

Join(S,G) message sent upstream.

2.6.6.1. Where to send Join/Prune messages

The following rules apply, to both forwarded (in case of PIM Relay), refresh and triggered (in case of PIM Proxy) (S,G)/(*,G) Join/Prune messages.

- o The upstream neighbor field in the Join/Prune to be sent is set to the N in the corresponding Upstream FSM.
- o if Rport(N) is an AC, send the message to Rport(N).
- o Additionally, if OutgoingPortList(X,G,N) contains at least one AC, then the message MUST be sent to at least all the PWs in UpstreamPorts(G) (for (*,G)) or InheritedUpstreamPorts(S,G) (for (S,G)). Alternatively, the message MAY be sent to all PWs.

Sending to a subset of PWs as described above guarantees that if traffic (of the same flow) from two upstream routers were to reach this PE, then the two routers will receive from each other, triggering assert.

Sending to all PWs guarantees that if two upstream routers both send traffic for the same flow (even if it is to different sets of downstream PEs), then they'll receive from each other, triggering assert.

2.7. Bidirectional-PIM (PIM-BIDIR)

PIM-BIDIR is a variation of PIM-SM. The main differences between PIM-SM and Bidirectional-PIM are as follows:

- o There are no source-based trees, and source-specific multicast is not supported (i.e., no (S,G) states) in PIM-BIDIR.
- o Multicast traffic can flow up the shared tree in PIM-BIDIR.
- o To avoid forwarding loops, one router on each link is elected as the Designated Forwarder (DF) for each RP in PIM-BIDIR.

The main advantage of PIM-BIDIR is that it scales well for many-to-many applications. However, the lack of source-based trees means that multicast traffic is forced to remain on the shared tree.

As described in [PIM-BIDIR], parts of a PIM-BIDIR enabled network may forward traffic without exchanging Join/Prune messages, for instance between DF's and the RPL.

As the described procedures for Pim snooping rely on the presence of Join/Prune messages, enabling Pim snooping on PIM-BIDIR networks could break the PIM-BIDIR functionality. Deploying Pim snooping on PIM-BIDIR enabled networks will require some further study. Some thoughts are gathered in Appendix A.

2.8. Interaction with IGMP Snooping

Whenever IGMP Snooping is enabled in conjunction with PIM Snooping in the same VPLS instance the PE SHOULD follow these rules:

- o To maintain the list of multicast routers and ports on which they are attached, the PE SHOULD NOT use the rules as described in RFC4541 [IGMP-SNOOP] but SHOULD rely on the neighbors discovered by PIM Snooping . This list SHOULD then be used to apply the forwarding rule as described in 2.1.1.(1) of RFC4541 [IGMP-SNOOP].
- o If the PE supports proxy-reporting, an IGMP membership learned only on a port to which a PIM neighbor is attached but not elsewhere SHOULD NOT be included in the summarized upstream report sent to that port.

2.9. PIM-DM

The characteristics of PIM-DM is flood and prune behavior. Shortest path trees are built as a multicast source starts transmitting.

2.9.1. Building PIM-DM Snooping States

PIM-DM Snooping states are built by snooping on the PIM-DM Join, Prune, Graft and State Refresh messages received on AC/PWs and State-Refresh Messages sent on AC/PWs. By snooping on these PIM-DM messages, a PE builds the following states per (S,G,N) where S is the address of the multicast source, G is the Group address and N is the upstream neighbor to which Prunes/Grafts are sent by downstream CEs:

Per PIM (S,G,N):

Port PIM (S,G,N) Prune State:

- * DownstreamPState(S,G,N,Port): One of {"NoInfo" (NI), "Pruned" (P), "PrunePending" (PP)}
- * Prune Pending Timer (PPT)

- * Prune Timer (PT)
- * Upstream Port (valid if the PIM(S,G,N) Prune State is "Pruned").

2.9.2. PIM-DM Downstream Per-Port PIM(S,G,N) State Machine

The downstream per-port PIM(S,G,N) state machine is as defined in section 4.4.2 of [PIM-DM] with a few changes relevant to PIM Snooping. When reading section 4.4.2 of [PIM-DM] for the purposes of PIM-Snooping please be aware that the downstream states are built per (S, G, N, Downstream-Port) in PIM-Snooping and not per {Downstream-Interface, S, G} as in a PIM-DM router. As noted in the previous Section 2.9.1, the states (DownstreamPState) and timers (PPT and PT) are per (S,G,N,P).

2.9.3. Triggering ASSERT election in PIM-DM

Since PIM-DM is a flood-and-prune protocol, traffic is flooded to all routers unless explicitly pruned. Since PIM-DM routers do not prune on non-RPF interfaces, PEs should typically not receive Prunes on Rport(RPF-neighbor). So the asserting routers should typically be in pim_oiflist(S,G). In most cases, assert election should occur naturally without any special handling since data traffic will be forwarded to the asserting routers.

However, there are some scenarios where a prune might be received on a port which is also an upstream port (UP). If we prune the port from pim_oiflist(S,G), then it would not be possible for the asserting routers to determine if traffic arrived on their downstream port. This can be fixed by adding pim_iifs(S,G) to pim_oiflist(S,G) so that data traffic flows to the UP ports.

2.10. PIM Proxy

As noted earlier, PIM Snooping will work correctly only if Join Suppression is disabled in the VPLS. If Join Suppression is enabled in the VPLS, then PEs MUST do PIM Proxy/Relay for VPLS Multicast to work correctly. This section applies specifically to the full Proxy case and not Relay.

2.10.1. Upstream PIM Proxy behavior

A PIM Proxy PE consumes Join/Prune messages and regenerates PIM Join/Prune messages to be sent upstream by implementing Upstream FSM as specified in the PIM RFC. This is the only difference from PIM Relay.

The source IP address in PIM packets sent upstream SHOULD be the address of a PIM downstream neighbor in the corresponding join/prune state. The address picked MUST NOT be the upstream neighbor field to be encoded in the packet. The layer 2 encapsulation for the selected source IP address MUST be the encapsulation recorded in the PIM Neighbor database for that IP address.

2.11. Directly Connected Multicast Source

If there is a source in the CE network that connects directly into the VPLS instance, then multicast traffic from that source MUST be sent to all PIM routers on the VPLS instance apart from the IGMP receivers in the VPLS. If there is already (S,G) or (*,G) snooping state that is formed on any PE, this will not happen per the current forwarding rules and guidelines. So, in order to determine if traffic needs to be flooded to all routers, a PE must be able to determine if the traffic came from a host on that LAN. There are three ways to address this problem:

- o The PE would have to do ARP snooping to determine if a source is directly connected.
- o Another option is to have configuration on all PEs to say there are CE sources that are directly connected to the VPLS instance and disallow snooping for the groups for which the source is going to send traffic. This way traffic from that source to those groups will always be flooded within the provider network.
- o A third option is to require that sources of CE multicast traffic must be behind a router.

This document recommends the third option - sources traffic must be behind a router.

2.12. Data Forwarding Rules

First we define the rules that are common to PIM-SM and PIM-DM PEs. Forwarding rules for each protocol type is specified in the sub-sections.

If there is no matching forwarding state, then the PE SHOULD discard the packet, i.e., the UserDefinedPortList below SHOULD be empty.

The following general rules MUST be followed when forwarding multicast traffic in a VPLS:

- o Traffic arriving on a port MUST NOT be forwarded back onto the same port.
- o Due to VPLS Split-Horizon rules, traffic ingressing on a PW MUST NOT be forwarded to any other PW.

2.12.1. PIM-SM Data Forwarding Rules

Per the rules in [PIM-SM] and per the additional rules specified in this document,

```
OutgoingPortList(*,G) = immediate_olist(*,G) (+)
                        UpstreamPorts(*,G) (+)
                        Rport(PimDR)
```

```
OutgoingPortList(S,G) = inherited_olist(S,G) (+)
                        UpstreamPorts(S,G) (+)
                        (UpstreamPorts(*,G) (-)
                        UpstreamPorts(S,G,rpt)) (+)
                        Rport(PimDR)
```

[PIM-SM] specifies how `immediate_olist(*,G)` and `inherited_olist(S,G)` are built. `PimDR` is the IP address of the PIM DR in the VPLS.

The PIM-SM Snooping forwarding rules are defined below in pseudocode:


```
BEGIN
  iif is the incoming port of the multicast packet.
  S is the Source IP Address of the multicast packet.
  G is the Destination IP Address of the multicast packet.

  If there is (S,G) state on the PE
  Then
    OutgoingPortList = OutgoingPortList(S,G)
  Else if there is (*,G) state on the PE
  Then
    OutgoingPortList = OutgoingPortList(*,G)
  Else
    OutgoingPortList = UserDefinedPortList
  Endif

  If iif is an AC
  Then
    OutgoingPortList = OutgoingPortList (-) iif
  Else
    ## iif is a PW
    OutgoingPortList = OutgoingPortList (-) PWPorts
  Endif

  Forward the packet to OutgoingPortList.
END
```

First if there is (S,G) state on the PE, then the set of outgoing ports is OutgoingPortList(S,G).

Otherwise if there is (*,G) state on the PE, the set of outgoing ports is OutgoingPortList(*,G).

The packet is forwarded to the selected set of outgoing ports while observing the general rules above in Section 2.12

2.12.2. PIM-DM Data Forwarding Rules

The PIM-DM Snooping data forwarding rules are defined below in pseudocode:

```
BEGIN
    iif is the incoming port of the multicast packet.
    S is the Source IP Address of the multicast packet.
    G is the Destination IP Address of the multicast packet.

    If there is (S,G) state on the PE
    Then
        OutgoingPortList = olist(S,G)
    Else
        OutgoingPortList = UserDefinedPortList
    Endif

    If iif is an AC
    Then
        OutgoingPortList = OutgoingPortList (-) iif
    Else
        ## iif is a PW
        OutgoingPortList = OutgoingPortList (-) PWPorts
    Endif

    Forward the packet to OutgoingPortList.
END
```

If there is forwarding state for (S,G), then forward the packet to olist(S,G) while observing the general rules above in section Section 2.12

[PIM-DM] specifies how olist(S,G) is constructed.

3. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

4. Security Considerations

Security considerations provided in VPLS solution documents (i.e., [VPLS-LDP] and [VPLS-BGP]) apply to this document as well.

5. Contributors

Yetik Serbest, Suresh Boddapati co-authored earlier versions.

Karl (Xiangrong) Cai and Princy Elizabeth made significant contributions to bring the specification to its current state, especially in the area of Join forwarding rules.

6. Acknowledgements

Many members of the L2VPN and PIM working groups have contributed to and provided valuable comments and feedback to this draft, including Vach Kompella, Shane Amante, Sunil Khandekar, Rob Nath, Marc Lassere, Yuji Kamite, Yiqun Cai, Ali Sajassi, Jozef Raets, Himanshu Shah (Ciena), Himanshu Shah (Alcatel-Lucent).

7. References

7.1. Normative References

- [PIM-BIDIR] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, 2007.
- [PIM-DM] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast Version 2 - Dense Mode Specification", RFC 3973, 2005.
- [PIM-SM] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast- Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, 2006.
- [PIM-SSM] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, 2006.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, 1997.
- [RPF-VECTOR] Wijnands, I., Boers, A., and E. Rosen, "The Reverse Path Forwarding (RPF) Vector TLV", RFC 5496, 2009.

7.2. Informative References

- [IGMP-SNOOP] Christensen, M., Kimball, K., and F. Solensky, "Considerations for IGMP and MLD Snooping PEs", RFC 4541, 2006.

[VPLS-BGP]

Kompella, K. and Y. Rekhter, "Virtual Private LAN Service using BGP for Auto-Discovery and Signaling", RFC 4761, 2007.

[VPLS-LDP]

Lasserre, M. and V. Kompella, "Virtual Private LAN Services using LDP Signaling", RFC 4762, 2007.

[VPLS-MCAST-REQ]

Kamite, Y., Wada, Y., Serbest, Y., Morin, T., and L. Fang, "Requirements for Multicast Support in Virtual Private LAN Services", RFC 5501, 2009.

[VPLS-MCAST-TREES]

Aggarwal, R., Kamite, Y., Fang, L., and Y. Rekhter, "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-11, Work in Progress.

Appendix A. PIM-BIDIR Thoughts

This section describes some guidelines that may be used to preserve PIM-BIDIR functionality in combination with Pim Snooping.

In order to preserve PIM-BIDIR Pim snooping routers need to set up forwarding states so that :

- o on the RPL all traffic is forwarded to all Rport(N)
- o on any other interface traffic is always forwarded to the DF

The information needed to setup these states may be obtained by :

- o determining the mapping between group(range) and RP
- o snooping and storing DF election information
- o determining where the RPL is, this could be achieved by static configuration, or by combining the information mentioned in previous bullets.

A.1. PIM-BIDIR Data Forwarding Rules

The PIM-BIDIR Snooping forwarding rules are defined below in pseudocode:

```
BEGIN
  iif is the incoming port of the multicast packet.
  G is the Destination IP Address of the multicast packet.

  If there is forwarding state for G
  Then
    OutgoingPortList = olist(G)
  Else
    OutgoingPortList = UserDefinedPortList
  Endif

  If iif is an AC
  Then
    OutgoingPortList = OutgoingPortList (-) iif
  Else
    ## iif is a PW
    OutgoingPortList = OutgoingPortList (-) PWPorts
  Endif

  Forward the packet to OutgoingPortList.
END

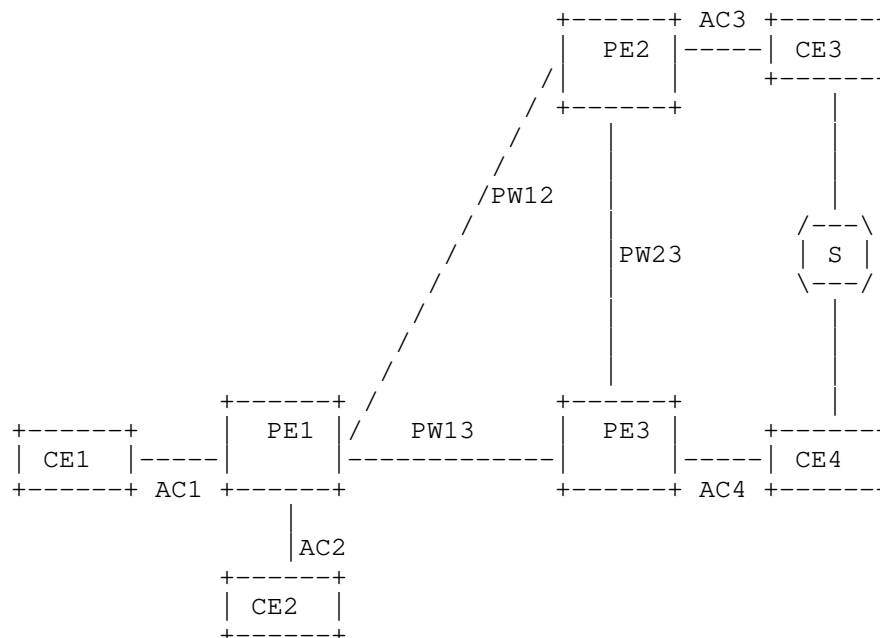
If there is forwarding state for G, then forward the packet to
olist(G) while observing the general rules above in Section 2.12

[PIM-BIDIR] specifies how olist(G) is constructed.
```

Appendix B. Example Network Scenario

Let us consider the scenario in Figure 3.

An Example Network for Triggering Assert



In the examples below, $JT(\text{Port}, S, G, N)$ is the downstream Join Expiry Timer on the specified Port for the (S, G) with upstream neighbor N .

B.1. Pim Snooping Example

In the network depicted in Figure 3, S is the source of a multicast stream (S, G) . $CE1$ and $CE2$ both have two ECMP routes to reach the source.

1. $CE1$ Sends a $Join(S, G)$ with $UpstreamNeighbor(S, G) = CE3$.
2. $PE1$ snoops on the $Join(S, G)$ and builds forwarding states since it is received on an AC. It also floods the $Join(S, G)$ in the VPLS. $PE2$ snoops on the $Join(S, G)$ and builds forwarding state since the $Join(S, G)$ is targeting a neighbor residing on an AC. $PE3$ does not create forwarding state for (S, G) because this is a PW-only join and there is neither existing $(*, G)$ state with an AC in $UpstreamPorts(*, G)$ nor an existing (S, G) state with an AC in $UpstreamPorts(S, G)$. Both $PE2$ and $PE3$ will also flood the $Join(S, G)$ in the VPLS

The resulting states at the PEs is as follows:

At $PE1$:

$JT(AC1, S, G, CE3) = JP_HoldTime$
 $UpstreamNeighbors(S, G) = \{ CE3 \}$

```

UpstreamPorts(S,G)      = { PW12 }
OutgoingPortList(S,G)   = { AC1, PW12 }

```

At PE2:

```

JT(PW12,S,G,CE3)        = JP_HoldTime
UpstreamNeighbors(S,G)   = { CE3 }
UpstreamPorts(S,G)       = { AC3 }
OutgoingPortList(S,G)    = { PW12, AC3 }

```

At PE3:

No (S,G) state

3. The multicast stream (S,G) flows along CE3 -> PE2 -> PE1 -> CE1
4. Now CE2 sends a Join(S,G) with Upstream Neighbor(S,G) = CE4.
5. All PEs snoop on the Join(S,G), build forwarding state and flood the Join(S,G) in the VPLS. Note that for PE2 even though this is a PW-only join, forwarding state is built on this Join(S,G) since PE2 has existing (S,G) state with an AC in UpstreamPorts(S,G)

The resulting states at the PEs:

At PE1:

```

JT(AC1,S,G,CE3)          = active
JT(AC2,S,G,CE4)          = JP_HoldTime
UpstreamNeighbors(S,G)    = { CE3, CE4 }
UpstreamPorts(S,G)        = { PW12, PW13 }
OutgoingPortList(S,G)     = { AC1, PW12, AC2, PW13 }

```

At PE2:

```

JT(PW12,S,G,CE4)         = JP_HoldTime
JT(PW12,S,G,CE3)         = active
UpstreamNeighbors(S,G)    = { CE3, CE4 }
UpstreamPorts(S,G)        = { AC3, PW23 }
OutgoingPortList(S,G)     = { PW12, AC3, PW23 }

```

At PE3:

```

JT(PW13,S,G,CE4)         = JP_HoldTime
UpstreamNeighbors(S,G)    = { CE4 }
UpstreamPorts(S,G)        = { AC4 }
OutgoingPortList(S,G)     = { PW13, AC4 }

```

6. The multicast stream (S,G) flows into the VPLS from the two CEs CE3 and CE4. PE2 forwards the stream received from CE3 to PW23 and PE3 forwards the stream to AC4. This facilitates the CE routers to trigger assert election. Let us say CE3 becomes the assert winner.
7. CE3 sends an Assert message to the VPLS. The PEs flood the Assert message without examining it.

8. CE4 stops sending the multicast stream to the VPLS.
9. CE2 notices an RPF change due to Assert and sends a Prune(S,G) with Upstream Neighbor = CE4. CE2 also sends a Join(S,G) with Upstream Neighbor = CE3.
10. All the PEs start a prune-pend timer on the ports on which they received the Prune(S,G). When the prune-pend timer expires, all PEs will remove the downstream (S,G,CE4) states.

Resulting states at the PEs:

At PE1:

JT(AC1,S,G,CE3)	= active
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { PW12 }
OutgoingPortList(S,G)	= { AC1, AC2, PW12 }

At PE2:

JT(PW12,S,G,CE3)	= active
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { AC3 }
OutgoingPortList(S,G)	= { PW12, AC3 }

At PE3:

JT(PW13,S,G,CE3)	= JP_HoldTime
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { PW23 }
OutgoingPortList(S,G)	= { PW13, PW23 }

Note that at this point at PE3, since there is no AC in OutgoingPortList(S,G) and no (*,G) or (S,G) state with an AC in UpstreamPorts(*,G) or UpstreamPorts(S,G) respectively, the existing (S,G) state at PE3 can also be removed. So finally:

At PE3:

No (S,G) state

Note that at the end of the assert election, there should be no duplicate traffic forwarded downstream and traffic should flow only on the desired path. Also note that there are no unnecessary (S,G) states on PE3 after the assert election.

B.2. PIM Proxy Example with (S,G) / (*,G) interaction

In the same network, let us assume CE4 is the Upstream Neighbor towards the RP for G.

JPST(S,G,N) is the JP sending timer for the (S,G) with upstream neighbor N.

1. CE1 Sends a Join(S,G) with Upstream Neighbor(S,G) = CE3.
2. PE1 consumes the Join(S,G) and builds forwarding state since the Join(S,G) is received on an AC.

PE2 consumes the Join(S,G) and builds forwarding state since the Join(S,G) is targeting a neighbor residing on an AC.

PE3 consumes the Join(S,G) but does not create forwarding state for (S,G) since this is a PW-only join and there is neither existing (*,G) state with an AC in UpstreamPorts(*,G) nor an existing (S,G) state with an AC in UpstreamPorts(S,G)

The resulting states at the PEs is as follows:

PE1 states:

```
JT(AC1,S,G,CE3)      = JP_HoldTime
JPST(S,G,CE3)        = t_periodic
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)    = { PW12 }
OutgoingPortList(S,G) = { AC1, PW12 }
```

PE2 states:

```
JT(PW12,S,G,CE3)     = JP_HoldTime
JPST(S,G,CE3)        = t_periodic
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)    = { AC3 }
OutgoingPortList(S,G) = { PW12, AC3 }
```

PE3 states:

No (S,G) state

Joins are triggered as follows:

PE1 triggers a Join(S,G) targeting CE3. Since the Join(S,G) was received on an AC and is targeting a neighbor that is residing across a PW, the triggered Join(S,G) is sent on all PWs.

PE2 triggers a Join(S,G) targeting CE3. Since the Joins(S,G) is targeting a neighbor residing on an AC, it only sends the join on AC3.

PE3 ignores the Join(S,G) since this is a PW-only join and there is neither existing (*,G) state with an AC in UpstreamPorts(*,G) nor an existing (S,G) state with an AC in UpstreamPorts(S,G)

3. The multicast stream (S,G) flows along CE3 -> PE2 -> PE1 -> CE1.
4. Now let us say CE2 sends a Join(*,G) with UpstreamNeighbor(*,G) = CE4.

5. PE1 consumes the Join(*,G) and builds forwarding state since the Join(*,G) is received on an AC.

PE2 consumes the Join(*,G) and though this is a PW-only join, forwarding state is build on this Join(*,G) since PE2 has existing (S,G) state with an AC in UpstreamPorts(S,G). However, since this is a PW-only join, PE2 only adds the PW towards PE3 (PW23) into UpstreamPorts(*,G) and hence into OutgoingPortList(*,G). It does not add the PW towards PE1 (PW12) into OutgoingPortsList(*,G)

PE3 consumes the Join(*,G) and builds forwarding state since the Join(*,G) is targeting a neighbor residing on an AC.

The resulting states at the PEs is as follows:

PE1 states:

```
JT(AC1, *, G, CE4)      = JP_HoldTime
JPST(*, G, CE4)         = t_periodic
UpstreamNeighbors(*, G) = { CE4 }
UpstreamPorts(*, G)     = { PW13 }
OutgoingPortList(*, G)  = { AC2, PW13 }

JT(AC1, S, G, CE3)      = active
JPST(S, G, CE3)         = active
UpstreamNeighbors(S, G) = { CE3 }
UpstreamPorts(S, G)     = { PW12 }
OutgoingPortList(S, G)  = { AC1, PW12, PW13 }
```

PE2 states:

```
JT(PW12, *, G, CE4)     = JP_HoldTime
UpstreamNeighbors(*, G) = { CE4 }
UpstreamPorts(G)        = { PW23 }
OutgoingPortList(*, G)  = { PW23 }

JT(PW12, S, G, CE3)     = active
JPST(S, G, CE3)         = active
UpstreamNeighbors(S, G) = { CE3 }
UpstreamPorts(S, G)     = { AC3 }
OutgoingPortList(S, G)  = { PW12, AC3, PW23 }
```

PE3 states:

```
JT(PW13, *, G, CE4)     = JP_HoldTime
JPST(*, G, CE4)         = t_periodic
UpstreamNeighbors(*, G) = { CE4 }
UpstreamPorts(*, G)     = { AC4 }
OutgoingPortList(*, G)  = { PW13, AC4 }
```

Joins are triggered as follows:

PE1 triggers a Join(*,G) targeting CE4. Since the Join(*,G) was received on an AC and is targeting a neighbor that is residing across a PW, the triggered Join(S,G) is sent on all PWs.

PE2 does not trigger a Join(*,G) based on this join since this is a PW-only join.

PE3 triggers a Join(*,G) targeting CE4. Since the Join(*,G) is targeting a neighbor residing on an AC, it only sends the join on AC4.

6. In case traffic is not flowing yet (i.e. step 3 is delayed to come after step 6) and in the interim JPST(S,G,CE3) on PE1 expires, causing it to send a refresh Join(S,G) targeting CE3, since the refresh Join(S,G) is targeting a neighbor that is residing across a PW, the refresh Join(S,G) is sent on all PWs.

7. Note that PE1 refreshes its JT timer based on reception of refresh joins from CE1 and CE2

PE2 consumes the Join(S,G) and refreshes the JT(PW12,S,G,CE3) timer.

PE3 consumes the Join(S,G). It also builds forwarding state on this Join(S,G), even though this is a PW-only join, since now PE2 has existing (*,G) state with an AC in UpstreamPorts(*,G). However, since this is a PW-only join, PE3 only adds the PW towards PE2 (PW23) into UpstreamPorts(S,G) and hence into OutgoingPortList(S,G). It does not add the PW towards PE1 (PW13) into OutgoingPortList(S,G).

PE3 States:

JT(PW13,*,G,CE4)	= active
JPST(S,G,CE4)	= active
UpstreamNeighbors(*,G)	= { CE4 }
UpstreamPorts(*,G)	= { AC4 }
OutgoingPortList(*,G)	= { PW13, AC4 }
JT(PW13,S,G,CE3)	= JP_HoldTime
UpstreamNeighbors(*,G)	= { CE3 }
UpstreamPorts(*,G)	= { PW23 }
OutgoingPortList(*,G)	= { PW13, AC4, PW23 }

Joins are triggered as follows:

PE2 already has (S,G) state, so it does not trigger a Join(S,G) based on reception of this refresh join.

PE3 does not trigger a Join(S,G) based on this join since this is a PW-only join.

8. The multicast stream (S,G) flows into the VPLS from the two

CEs, CE3 and CE4. PE2 forwards the stream received from CE3 to PW12 and PW23. At the same time PE3 forwards the stream received from CE4 to PW13 and PW23.

The stream received over PW12 and PW13 is forwarded by PE1 to AC1 and AC2.

The stream received by PE3 over PW23 is forwarded to AC4. The stream received by PE2 over PW23 is forwarded to AC3. Either of these facilitates the CE routers to trigger assert election.

9. CE3 and/or CE4 send(s) Assert message(s) to the VPLS. The PEs flood the Assert message(s) without examining it.
10. CE3 becomes the (S,G) assert winner and CE4 stops sending the multicast stream to the VPLS.
11. CE2 notices an RPF change due to Assert and sends a Prune(S,G,rpt) with Upstream Neighbor = CE4.
12. PE1 consumes the Prune(S,G,rpt) and since PruneDesired(S,G,Rpt,CE4) is TRUE, it triggers a Prune(S,G,rpt) to CE4. Since the prune is targeting a neighbor across a PW, it is sent on all PWs.

PE2 consumes the Prune(S,G,rpt) and does not trigger any prune based on this Prune(S,G,rpt) since this was a PW-only prune.

PE3 consumes the Prune(S,G,rpt) and since PruneDesired(S,G,rpt,CE4) is TRUE it sends the Prune(S,G,rpt) on AC4.

PE1 states:

```

JT(AC2,*,G,CE4)           = active
JPST(*,G,CE4)             = active
UpstreamNeighbors(*,G)    = { CE4 }
UpstreamPorts(*,G)        = { PW13 }
OutgoingPortList(*,G)     = { AC2, PW13 }

JT(AC2,S,G,CE4)           = JP_Holdtime with FLAG sgrpt prune
JPST(S,G,CE4)             = none, since this is sent along
                           with the Join(*,G) to CE4 based
                           on JPST(*,G,CE4) expiry
UpstreamPorts(S,G,rpt)    = { PW13 }
UpstreamNeighbors(S,G,rpt) = { CE4 }

JT(AC1,S,G,CE3)           = active
JPST(S,G,CE3)             = active
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { PW12 }

```

OutgoingPortList(S,G) = { AC1, PW12, AC2 }

At PE2:

```

JT(PW12,*,G,CE4)      = active
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)     = { PW23 }
OutgoingPortList(*,G)  = { PW23 }

JT(PW12,S,G,CE4)      = JP_Holdtime with FLAG sgrpt prune
JPST(S,G,CE4)          = none, since this was created
                        off a PW-only prune
UpstreamPorts(S,G,rpt) = { PW23 }
UpstreamNeighbors(S,G,rpt) = { CE4 }

JT(PW12,S,G,CE3)      = active
JPST(S,G,CE3)          = active
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)     = { AC3 }
OutgoingPortList(*,G)  = { PW12, AC3 }

```

At PE3:

```

JT(PW13,*,G,CE4)      = active
JPST(*,G,CE4)          = active
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)     = { AC4 }
OutgoingPortList(*,G)  = { PW13, AC4 }

JT(PW13,S,G,CE4)      = JP_Holdtime with S,G,rpt prune flag
JPST(S,G,CE4)          = none, since this is sent along
                        with the Join(*,G) to CE4 based
                        on JPST(*,G,CE4) expiry
UpstreamNeighbors(S,G,rpt) = { CE4 }
UpstreamPorts(S,G,rpt)  = { AC4 }

JT(PW13,S,G,CE3)      = active
JPST(S,G,CE3)          = none, since this state is
                        created by PW-only join
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)     = { PW23 }
OutgoingPortList(S,G)  = { PW23 }

```

Even in this example, at the end of the (S,G) / (*,G) assert election, there should be no duplicate traffic forwarded downstream and traffic should flow only to the desired CEs.

However, the reason we don't have duplicate traffic is because one of the CEs stops sending traffic due to assert, not because we don't

have any forwarding state in the PEs to do this forwarding.

Authors' Addresses

Olivier Dornon
Alcatel-Lucent
50 Copernicuslaan
Antwerp, B2018

Email: olivier.dornon@alcatel-lucent.com

Jayant Kotalwar
Alcatel-Lucent
701 East Middlefield Rd.
Mountain View, CA 94043

Email: jayant.kotalwar@alcatel-lucent.com

Venu Hemige

Email: vhemige@gmail.com

Ray Qiu
Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089

Email: rqiujuniper.net

Jeffrey Zhang
Juniper Networks, Inc.
10 Technology Park Drive
Westford, MA 01886

Email: zzhang@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 21, 2014

K. Patel
S. Boutros
J. Liste
Cisco Systems
B. Wen
Comcast
October 18, 2013

Flow-Aware Transport of Pseudowires Extension for BGP
draft-keyupate-l2vpn-fat-pw-bgp-00.txt

Abstract

[RFC6391] describes a mechanism that uses an additional label (Flow Label) in the MPLS label stack that allows Label Switch Routers to balance flows within Pseudowires at a finer granularity than the individual Pseudowires across the Equal Cost Multiple Paths (ECMPs) that exists within the Packet Switched Network (PSN).

Furthermore, [RFC6391] defines the LDP protocol extensions required to synchronize the flow label states between the ingress and egress PEs when using the signaling procedures defined in the [RFC4447].

This draft defines protocol extensions required to synchronize flow label states among PEs when using the BGP-based signaling procedures defined in [RFC4761]. These protocol extensions are equally applicable to point-to-point L2VPNs defined in [RFC6624].

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 April 21, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	2
1.1. Requirements Language	3
2. Modifications to Layer 2 Info Extended Community	3
3. Signaling the Presence of the Flow Label	5
4. Acknowledgements	6
5. Contributors	6
6. IANA Considerations	6
7. Security Considerations	6
8. References	6
8.1. Normative References	6
8.2. Informative References	7
Authors' Addresses	7

1. Introduction

A pseudowire (PW) [RFC3985] is normally transported over one single network path, even if multiple Equal Cost Multiple Paths (ECMPs) exist between the ingress and egress PW provider edge (PE) equipment. This is required to preserve the characteristics of the emulated service. The use of a single path to preserve the packet delivery order remains the default mode of operation of a PW and is described in [RFC4385], [RFC4928].

Using the principles defined in [RFC6391], this draft augments the BGP-signaling procedures of [RFC4761] and [RFC6624] to allow an OPTIONAL mode that may be employed when the use of ECMPs is known to be beneficial to the operation of the PW.

High bandwidth Ethernet-based services are a prime example that benefits from the ability to load-balance flows in a PW over multiple PSN paths. In general, this notion is applicable to cases where the ratio between the PW access speed and the PSNs core link bandwidth is large.

To achieve the load-balancing goal, [RFC6391] introduces the notion of an additional Label Stack Entry (LSE) (Flow label) located at the bottom of the stack (right after PW LSE). Label Switching Routers (LSRs) commonly generate a hash of the label stack in order to discriminate and distribute flows over available ECMPs. The presence of the Flow label (closely associated to a flow determined by the ingress PE) will normally provide the greatest entropy.

Furthermore, following the procedures for Inter-AS scenarios described in [RFC4761] section 3.4, the Flow label should never be handled by the ASBRs, only the terminating PEs on each AS will be responsible for popping or pushing this label. This is equally applicable to Method B [section 3.4.2] of [RFC4761] where ASBRs are responsible for swapping the PW label as traffic traverses from ASBR to PE and ASBR to ASBR directions. Therefore, the Flow label will remain untouched across AS boundaries.

1.1. Requirements Language

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. Modifications to Layer 2 Info Extended Community

The Layer 2 Info Extended Community is used to signal control information about the pseudowires to be setup. The extended community format is described in [RFC4761]. The format of this extended community is described as:

Extended community type (2 octets)
Encaps Type (1 octet)
Control Flags (1 octet)
Layer-2 MTU (2 octet)
Reserved (2 octets)

Layer 2 Info Extended Community

Control Flags:

This field contains bit flags relating to the control information about pseudowires. This field is augmented with a definition of 2 new flags field.

0	1	2	3	4	5	6	7	
+	+	+	+	+	+	+	+	+
	MBZ		T		R		C	
+	+	+	+	+	+	+	+	+

(MBZ = MUST Be Zero)

Control Flags Bit Vector

With Reference to the Control Flags Bit Vector, the following bits in the Control Flags are defined; the remaining bits, designated MBZ, MUST be set to zero when sending and MUST be ignored when receiving this Extended Community.

- S Defined in [RFC4761].
- C Defined in [RFC4761].
- R When the bit value is 1, the PE is able to receive a Pseudowire packet with a flow label present. When the bit value is 0, the PE is unable to receive a Pseudowire packet with the flow label present.
- T When the bit value is 1, the PE is requesting the ability to send a Pseudowire packet that includes a flow label. When the bit value is 0, the PE is indicating that it will not send a Pseudowire packet containing a flow label.

3. Signaling the Presence of the Flow Label

As part of the Pseudowire signaling procedures described in [RFC4761], a Layer 2 Info Extended Community is advertised in the VPLS BGP NLRI. This draft recommends that the Control Flags field of this extended community be used to synchronize the flow label states amongst PEs for a given L2VPN.

A PE that wishes to send a flow label in a Pseudowire packet MUST include in its VPLS BGP NLRI a Layer 2 Info Extended Community using Control Flags field with T = 1.

A PE that is willing to receive a flow label in a Pseudowire packet MUST include in its VPLS BGP NLRI a Layer 2 Info Extended Community using Control Flags field with R = 1.

A PE that receives a VPLS BGP NLRI containing a Layer 2 Info Extended Community with R = 0 MUST NOT include a flow label in the Pseudowire packet.

Therefore, a PE sending a Control Flags field with T = 1 and receiving a Control Flags field with R = 1 MUST include a flow label in the Pseudowire packet. Under all other combinations, a PE MUST NOT include a flow label in the Pseudowire packet.

The signaling procedures in [RFC4761] state that the unspecified bits in the Control Flags field (bits 0-5) MUST be set to zero when sending and MUST be ignored when receiving. The signaling procedure described here is therefore backwards compatible with existing implementations. A PE not supporting the extensions described in this draft will always sent a value of ZERO in the position assigned by this draft to the R bit and therefore, flow labels will never be inserted by its remote peers.

Note that what is signaled is the desire to include the flow LSE in the label stack. The value of the flow label is a local matter for the ingress PE, and the label value itself is not signaled.

4. Acknowledgements

The authors would like to thank Bertrand Duvivier for the review and comments.

5. Contributors

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

Eric Lent

6. IANA Considerations

7. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4271].

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [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.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

8.2. Informative References

- [RFC2842] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.
- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.
- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, May 2012.

Authors' Addresses

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Sami Boutros
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: sboutros@cisco.com

Jose Liste
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: jliste@cisco.com

Internet-DraftFlow-Aware Transport of PWs Extension for BGP October 2013

Bin Wen
Comcast
1701 John F Kennedy Blvd
Philadelphia, PA 19103
USA

Email: bin_wen@cable.comcast.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 23, 2014

Z. Li
S. Zhuang
Huawei Technologies
October 20, 2013

An Architecture of Central Controlled Layer 2 Virtual Private Network
(L2VPN)
draft-li-l2vpn-ccvpn-arch-00

Abstract

With the emergence of Software Defined Networks (SDN), the architecture of forwarding and control element separation will develop faster. In the central controlled framework, control functionality of L2VPN can be done only by the Controllers. Consequently it can reduce control functionality in network nodes. This document defines the architecture of central controlled L2VPN and corresponding protocol extension requirement.

Requirements Language

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].

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 April 23, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Architecture	3
3.1. Application Scenarios	4
4. Solutions and Protocol Extensions	5
4.1. Overview	5
4.2. PW Establishment	5
4.3. PW Redundancy	6
4.4. MAC Withdraw	6
4.5. Capability Negotiation	7
5. IANA Considerations	7
6. Security Considerations	7
7. Normative References	7
Authors' Addresses	8

1. Introduction

With the development of network technologies, carrier's networks become increasingly complex. New technologies are required to integrate traditional switching networks such as ATM and FR networks through IP/MPLS networks. Layer 2 VPN (L2VPN) is therefore introduced. MPLS L2VPN transmits Layer 2 VPN services over an MPLS network. MPLS L2VPN enables operators to provide L2VPN services over different media, such as Asynchronous Transfer Mode (ATM), Frame Relay (FR), virtual local area network (VLAN), Ethernet, and Point-to-Point Protocol (PPP) in a unified MPLS network. Simply, the MPLS L2VPN indicates that Layer 2 data is transmitted transparently over an MPLS network. For the users, the MPLS network functions as a Layer 2 switched network through which Layer 2 connections can be set up between nodes. Layer 2 connections can be set up in virtual leased line (VLL) mode and virtual private LAN service (VPLS) mode.

With the emergence of Software Defined Networks (SDN), various services (e.g. L2VPN, L3VPN, MVPN) have been considered to deploy in a central controlled mode. The architecture of central controlled BGP is defined in [I-D.li-idr-cc-bgp-arch]. In the central controlled framework, control functionality of L2VPN can be done only by the Controllers. Consequently it can reduce control functionality in network nodes.

This document defines an architecture of Central Controlled L2VPN and corresponding protocol extension requirement.

2. Terminology

BGP: Border Gateway Protocol

FEC: Forwarding Equivalence Class

I2RS: Interface to Routing System

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

LDP: Label Distribution Protocol

MPLS: Multi-Protocol Label Switching

PW: Pseudo-Wire

SDN: Software-Defined Network

VPN: Virtual Private Network

VPLS: Virtual Private LAN Service

VSI: Virtual Switch Instance

3. Architecture

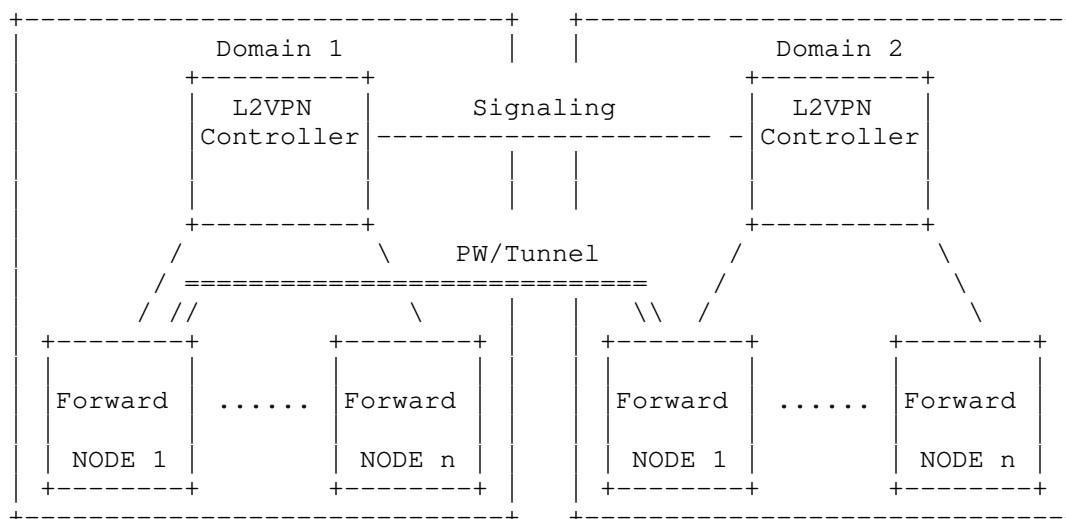


Figure 1: An Architecture of Central Controlled L2VPN

The figure above shows the architecture of central controlled L2VPN, which consists of two essential network elements: L2VPN Controller and Forward Node. In the architecture, there is no L2VPN related control functionality in Forward Nodes. L2VPN Controller controls all the Forward Nodes by download forwarding entries to control the forwarding behavior of the node. L2VPN Controllers need to communicate with each other via extension of existing protocols, e.g. BGP, LDP, etc. In this architecture, the L2VPN service set up between the forward nodes is proxied by the BGP Controllers.

The architecture defined in this document applies to VPLS, VPWS. Extension to EVPN will be described in a future version.

3.1. Application Scenarios

There are three application scenarios for deployment of Central Controlled L2VPN service.

Scenario 1: Partial Deployment

Some network nodes are upgrading to support Central Controlled L2VPN, the other nodes are retained as legacy network nodes. The new network nodes are controlled by L2VPN Controller. In this scenarios, the protocol extensions are applied between the legacy node and the controller.

Scenario 2: Multiple Controller within a Single AS

In this scenario, there are multiple controllers in a single AS to be responsible for setup of Central Controlled L2VPN service. The network will be partitioned into multiple domains in which one Central Controller controls a set of nodes. The protocol extensions are applied between the controllers.

Scenario 3: Multiple Controller within Multiple ASes

In this scenario, there are multiple controllers in different ASes to be responsible for setup of Central Controlled L2VPN service. Each AS has at least one Central Controller. The protocol extensions SHOULD support inter-AS application.

4. Solutions and Protocol Extensions

4.1. Overview

There are two options to implement the architecture of Central Controlled L2VPN.

Option 1:

Using BGP to distribute label and fulfill the other control functionality for central controlled L2VPN service.

This option can be applied to multiple scenarios.

Option 2:

Using LDP to distribute label and fulfill the other control functionality for central controlled L2VPN service.

This option is a transitional manner, mainly being used for communicating between legacy network node and L2VPN Controller.

4.2. PW Establishment

There are following procedures to set up PW in the Central Controlled L2VPN service:

1. Auto Discovery: The controller SHOULD advertise the address list of Forward Nodes participating in a specific VPLS or VLL to other controllers. After the communication between the controllers, the controller can discover the PW that should be set up between the Forwarding Nodes controlled by this controller and the Forwarding Nodes controlled by other controllers.

2. PW Label Allocation: After the process of auto discovery, the controller will advertise the label mapping message to the other controller for the PW which should be set up between a pair of Forward Nodes. The addresses of the local Forward Node and the remote Forward Node SHOULD be carried in the message to differentiate the PWs.

3. PW Forwarding Entry Creation: When receive the label mapping, the controller will find the tunnel to the Forward Node identified by the address information of the Local Forwarding Node in the message. Then the controller will create PW forwarding entry with PW label and the tunnel information and download the forwarding entry to the specified Forward Node identified by the address information of the remote Forward Node in the message .

4.3. PW Redundancy

[RFC4447] defines the PW redundancy mechanism and specifies the PW Status TLV to transmit the PW forwarding status. In the Central Controlled L2VPN, when advertise the status for the PW between the controllers, the addresses of the Local Forward Node and the Remote Forward Node should be carried to specify the specific PW. The similar TLV like PW Status TLV SHOULD also be defined to carry the status of the specific PW.

When the controller receives the message carried the PW status information, it will set the PW on the specified Forwarding Node as the state specified by the message.

4.4. MAC Withdraw

[RFC4762] describes a mechanism to remove 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 does not require relearning due to such topology change.

[I-D.ietf-l2vpn-vpls-ldp-mac-opt] 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 [I-D.ietf-l2vpn-pbb-vpls-pe-model] .

For Central Controlled L2VPN, L2VPN Controller needs to develop an ability to remove the MAC of the specific Forward Node. When a Forward Node within a L2VPN Controller wants to remove MAC Addresses that has been sent to the remote endpoint, the Controller needs to send MAC Withdraw Messages on behalf of the Forward Node. In the

message, the address of the remote forward node should be carried. When the other controller receive the message, it will remove the specified MAC addresses on the Forward Node identified by the address of the remote forward node in the message.

4.5. Capability Negotiation

To ensure backward compatibility with existing implementations, the capability for Central Controlled L2VPN SHOULD be negotiated between the controllers. The capability is advertised to each other by the controllers. After the successful negotiation of the capability, the other control functionalities for the central controlled L2VPN can be done by the controller.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

TBD.

7. Normative References

- [I-D.ietf-l2vpn-pbb-vpls-pe-model]
Balus, F., Sajassi, A., and N. Bitar, "Extensions to VPLS PE model for Provider Backbone Bridging", draft-ietf-l2vpn-pbb-vpls-pe-model-07 (work in progress), June 2013.
- [I-D.ietf-l2vpn-vpls-ldp-mac-opt]
Dutta, P., Balus, F., Stokes, O., and G. Calvignac, "LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS", draft-ietf-l2vpn-vpls-ldp-mac-opt-08 (work in progress), February 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [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.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC6718] Muley, P., Aissaoui, M., and M. Bocci, "Pseudowire Redundancy", RFC 6718, August 2012.
- [RFC6870] Muley, P. and M. Aissaoui, "Pseudowire Preferential Forwarding Status Bit", RFC 6870, February 2013.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 24, 2014

Z. Li
J. Zhang
Huawei Technologies
October 21, 2013

Using BGP between PE and CE in EVPN
draft-li-l2vpn-evpn-pe-ce-00

Abstract

This document specifies protocols and procedures of using BGP as PE-CE control protocol for carrying customer MAC routing information.

Requirements Language

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].

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 April 24, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	2
3. Application Scenarios	2
4. BGP E-VPN NLRI Extensions	3
5. Exchanging C-MAC Routes	4
5.1. Originating MAC Route at the CE router	4
5.2. Receiving a MAC Route by the PE router	5
6. IANA Considerations	6
7. Security Considerations	6
8. Normative References	6
Authors' Addresses	6

1. Introduction

[I-D.ietf-l2vpn-evpn] describes protocols and procedures for BGP MPLS based Ethernet VPNs. BGP is used for MAC learning by exchanging customer MAC routing information between PEs in the control plane instead of MAC learning between PEs in the data plane. It also states that mac learning between PEs and CEs MAY be done in the control plane, but it does not define the detailed protocols and procedures. This document specifies protocols and procedures of using BGP as PE-CE control protocol for carrying customer MAC routing information. This can provide some benefits such as fast convergence in some situation..

2. Terminology

This document uses terminology described in [I-D.ietf-l2vpn-evpn].

3. Application Scenarios

There are some benefits when control plane is introduced between PE and CE in EVPN network. The following illustrates the benifits with an example of fast convergence in the event of PE to CE network failure.

[I-D.ietf-l2vpn-evpn] defines a mechanism to efficiently and quickly signal, to remote PE nodes, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet segment. This mechanism optimizes the withdrawal of MAC advertisement routes, and then optimizes the network convergence time in the event of PE to CE failures. But it still can not fully

provide convergence time that is independent of the number of MAC addresses learned by the PE. There exist a situation where the network convergence time is dependent on the local MAC learning of PE and the advertisement of them to remote PE.

To illustrate this with an example, consider two PEs (PE1 and PE2) connected to a multi-homed Ethernet Segment ES1. All-Active redundancy mode is assumed. A given MAC address M1 is learned by PE1 but not PE2. On PE3, the following states may arise:

T1- When the MAC Advertisement Route from PE1 and the Ethernet A-D routes per ES1 from PE1 and PE2 are received, PE3 can forward traffic destined to M1 to both PE1 and PE2.

T2- After T1, when the ES1 connected to PE1 fails, PE1 MUST withdraw its Ethernet A-D route per ES1, then PE3 forwards traffic destined to M1 to PE2 only.

T3- After T2, PE1 MUST also withdraw the MAC advertisement routes (M1) that are impacted by the failure. Before PE2 learns M1 and advertises a MAC route for M1, PE3 will treat traffic to M1 as unknown unicast. If the behavior is to drop the unknown unicast based on administrative policy, the traffic to M1 on PE3 will be interrupted. Note that had PE2 also advertised a MAC route for M1 before PE1 withdraws its MAC route, then PE3 would have continued forwarding traffic destined to M1.

In the above example, once the local MAC learning of PE was done via control plane, both PE1 and PE2 advertise a MAC route for M1, then PE3 could continue forwarding traffic destined to M1 in the event of ES1 connected to PE1 or PE2 fails. In this case, the network convergence time is not dependent of the local MAC learning and advertisement of MAC addresses learned by the PE any more. The benefit can be achieved in case of single-active redundancy mode.

4. BGP E-VPN NLRI Extensions

A new route type is defined for EVPN NLRI to advertise customer MAC route between PE and CE in EVP:

+ 6 - Customer MAC advertisement route

A customer MAC advertisement route type specific EVPN NLRI consists of the following:

Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octet)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (4 or 16 octets)

It should be noted that the Route Distinguisher (RD) is not used since the customer MAC routes are always exchanged in the context of unawareness of Ethernet VPN.

5. Exchanging C-MAC Routes

This section describes the procedures of exchanging customer MAC routes between PE and CE. This document assumes that a CE and a PE exchange MAC routes over a direct BGP session, and also BGP is just only used to carrying MAC information from CE to PE in current version. Usage of carrying MAC information from PE to CE using BGP will be described in future version.

5.1. Originating MAC Route at the CE router

When a CE receives packets in a given VLAN from interfaces, other than interfaces connected to the PE, it learns MAC addresses in the data plane. If the given VLAN is in the setting of VLANs across the Ethernet links attached to a given PE, the CE MAY advertises the MAC addresses it learns in the data plane to the given PE, using MP-BGP and the specified MAC Route, in the control plane. The MAC route is constructed as follows:

- + The field of the Ethernet Segment Identifier is reserved for future use.
- + The Ethernet Tag ID is set the VLAN ID from which the MAC addresses is learned.

- + The MAC address length field is in bits and it is typically set to 48. However this specification enables specifying the MAC address as a prefix; in which case, the MAC address length field is set to the length of the prefix. This provides the ability to aggregate MAC addresses if the deployment environment supports that.
- + The MAC address is set to the value of MAC address the CE learned. The encoding of a MAC address MUST be the 6-octet MAC address specified by [802.1D-ORIG] [802.1D-REV]. If the MAC address is advertised as a prefix then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as 6 octets.
- + The IP Address field is optional. By default, the IP Address Length field is set to 0 and the IP address field is omitted from the route. When a valid IP address or address prefix needs to be advertised (e.g., for ARP suppression purposes or for inter-subnet switching), it is then encoded in this route. In this case, the IP Address Length field is in bits and it is the length of the IP prefix. This provides the ability to advertise IP address prefixes when the deployment environment supports that.
- + The encoding of an IP address MUST be either 4 octets for IPv4 or 16 octets for IPv6. When the IP address is advertised as a prefix, then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as either 4 or 16 octets. The length field of Ethernet NLRI is sufficient to determine whether an IP address/prefix is encoded in this route and if so, whether the encoded IP address/prefix is IPv4 or IPv6.
- + The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising CE.

It should be noted that the BGP advertisement for the MAC route does not need to carry the Route Target (RT) attributes because of its unawareness of Ethernet VPN.

5.2. Receiving a MAC Route by the PE router

When a PE receives a MAC route from a CE, it learns the MAC addresses advertised in the MAC route in the control plane and associates the MAC addresses with the Ethernet Segment from which it can reach to the advertising CE and the VLAN carried in the MAC route.

The PE SHOULD install forwarding state for the associated MAC addresses based on the Ethernet Segment and VLAN inferred from the MAC route.

In addition, the PE SHOULD advertise the MAC addresses it learns from CE in the control plane, to all the other PEs in the associated EVPN instance, using MP-BGP and the MAC Advertisement route defined in [I-D.ietf-l2vpn-evpn]. For example, the PE learns a MAC address M1 on a multi-homed Ethernet Segment (ES1) and on a VLAN 10, and the VLAN 10 is bundled to EVPN A. The PE SHOULD advertise the MAC address M1 to all the other PEs in EVPN A.

The construction of the MAC Advertisement route and procedures of handling the MAC Advertisement route on receiving it are specified in [I-D.ietf-l2vpn-evpn].

6. IANA Considerations

This document requires IANA to assign a new route type value for E-VPN NLRI.

7. Security Considerations

There are no additional security aspects beyond those of E-VPN ([I-D.ietf-l2vpn-evpn]).

8. Normative References

- [I-D.ietf-l2vpn-evpn]
Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04 (work in progress), July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Junlin Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jackey.zhang@huawei.com

L2VPN Workgroup
Internet Draft

Intended status: Standards Track

J. Rabadan
W. Henderickx
S. Palislaamovic
Alcatel-Lucent

F. Balus
Nuage Networks

A. Isaac
Bloomberg

Expires: April 24, 2014

October 21, 2013

IP Prefix Advertisement in E-VPN
draft-rabadan-l2vpn-evpn-prefix-advertisement-01

Abstract

E-VPN provides a flexible control plane that allows intra-subnet connectivity in an IP/MPLS and/or an NVO-based network. In Data Centers, there is also a need for a dynamic and efficient inter-subnet connectivity across Tenant Systems and End Devices that can be physical or virtual and may not support their own routing protocols. This document defines a new E-VPN route type for the advertisement of IP Prefixes and explains some use-case examples where this new route-type is used.

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), 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/lid-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 16, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Terminology	3
2. Introduction and problem statement	3
2.1 Inter-subnet connectivity requirements in Data Centers	3
2.2 The requirement for advertising IP prefixes in E-VPN	6
2.3 The requirement for a new E-VPN route type	7
3. The BGP E-VPN IP Prefix route	9
3.1 IP Prefix Route encoding	9
4. Benefits of using the E-VPN IP Prefix route	11
5. IP Prefix next-hop use-cases	12
5.1 TS IP address next-hop use-case	12
5.2 Floating IP next-hop use-case	15
5.3 IRB IP next-hop use-case	16
5.4 ESI next-hop ("Bump in the wire") use-case	18
6. Conclusions	20
7. Conventions used in this document	21
8. Security Considerations	21
9. IANA Considerations	21
10. References	21
10.1 Normative References	21
10.2 Informative References	21
11. Acknowledgments	21
12. Authors' Addresses	21

1. Terminology

GW IP: Gateway IP Address

IPL: IP address length

IRB: Integrated Routing and Bridging interface

ML: MAC address length

NVE: Network Virtualization Edge

TS: Tenant System

VA: Virtual Appliance

Overlay next-hop: object used in the IP Prefix route, as described in this document. It can be an IP address in the tenant space or an ESI, and identifies the next-hop to be used in IP lookups for a given IP Prefix at the routing context importing the route.

Underlay next-hop: IP address sent by BGP along with any E-VPN route, i.e. BGP next-hop. It identifies the NVE sending the route and it is used at the receiving NVE as the VXLAN destination VTEP or NVGRE destination end-point.

2. Introduction and problem statement

Inter-subnet connectivity is required within the Data Center, therefore IP Prefixes must be advertised in the control plane. This section explains why IP-VPN [RFC4364] procedures are not recommended for such advertisements and why the existing E-VPN MAC route type does not meet the Data Center requirements for the advertisement of IP Prefixes, hence a new E-VPN route type is proposed.

Section 2.1 describes the inter-subnet connectivity requirements in Data Centers. Section 2.2 and 2.3 explain why neither IP-VPN nor the existing E-VPN route types meet the requirements for IP Prefix advertisements. Once the need for a new E-VPN route type is justified, sections 2 and 3 will describe this route type and how it is used in some specific use cases.

2.1 Inter-subnet connectivity requirements in Data Centers

[E-VPN] is used as the control plane for a Network Virtualization Overlay (NVO3) solution in Data Centers (DC), where Network Virtualization Edge (NVE) devices can be located in Hypervisors or TORs, as described in [E-VPN-OVERLAYS].

If we use the term Tenant System (TS) to designate a physical or virtual system identified by MAC and IP addresses, and connected to an E-VPN instance, the following considerations apply:

- o The Tenant Systems may be Virtual Machines (VMs) that generate traffic from their own MAC and IP.
- o The Tenant Systems may be Virtual Appliance entities (VAs) that forward traffic to/from IP addresses of different End Devices seating behind them.
 - o These VAs can be firewalls, load balancers, NAT devices, other appliances or virtual gateways with virtual routing instances.
 - o These VAs do not have their own routing protocols and hence rely on the E-VPN NVEs to advertise the routes on their behalf.
 - o In all these cases, the VA will forward traffic to the Data Center using its own source MAC but the source IP will be the one associated to the End Device seating behind or a translated IP address (part of a public NAT pool) if the VA is performing NAT.
 - o Note that the same IP address could exist behind two of these TS. One example of this would be certain appliance resiliency mechanisms, where a virtual IP or floating IP can be owned by one of the two VAs running the resiliency protocol (the master VA). VRRP is one particular example of this. Another example is multi-homed subnets, i.e. the same subnet is connected to two VAs.
 - o Although these VAs provide IP connectivity to VMs and subnets behind them, they do not always have their own IP interface connected to the E-VPN NVE, e.g. layer-2 firewalls are examples of VAs not supporting IP interfaces.

The following figure illustrates some of the examples described above.

Figure 1 DC inter-subnet use-cases

- o TS1 is a VM that generates/receives traffic from/to IP1, where IP1 belongs to the E-VPN 10 subnet.
- o TS2 and TS3 are Virtual Appliances (VA) that generate/receive traffic from/to the subnets and hosts seating behind them (SN1, SN2, SN3, IP4 and IP5). Their IP addresses (IP2 and IP3) belong to the E-VPN subnet and they can also generate/receive traffic. When these VAs receive packets destined to their own MAC addresses (M2 and M3) they will route the packets to the

proper subnet or host. These VAs do not support routing protocols to advertise the subnets connected to them and can move to a different server and NVE when the Cloud Management System decides to do so. These VAs may also support redundancy mechanisms for some subnets, similar to VRRP, where a floating IP is owned by the master VA and only the master VA forwards traffic to a given subnet. E.g.: vIP23 in figure 1 is a floating IP that can be owned by TS2 or TS3 depending on who the master is. Only the master will forward traffic to SN1.

- o Integrated Routing and Bridging interfaces IRB1, IRB2 and IRB3 have their own IP addresses that belong to the E-VPN 10 subnet too. These IRB interfaces connect the E-VPN 10 subnet to Virtual Routing and Forwarding (VRF) instances that can route the traffic to other connected subnets for the same tenant (within the DC or at the other end of the WAN).
- o TS4 is a layer-2 VA that provides connectivity to subnets SN5, SN6 and SN7, but does not have an IP address itself in the E-VPN 10. TS4 is connected to a physical port on NVE5 assigned to Ethernet Segment Identifier 4.

All the above DC use cases require inter-subnet forwarding and therefore the individual host routes and subnets MUST be advertised:

- a) From the NVEs (since VAs and VMs do not run routing protocols) and
- b) Associated to an overlay next-hop that can be a VA IP address, a floating IP address, and IRB IP address or an ESI.

2.2 The requirement for advertising IP prefixes in E-VPN

In all the inter-subnet connectivity cases discussed in section 2.1 there is a need to advertise IP prefixes. The advertisement of such prefixes must meet certain requirements, specific to NVO-based Data Centers:

- o The data plane in NVO-based Data Centers is not based on IP over a GRE or MPLS tunnel as required by [RFC4364], but Ethernet over an IP tunnel, such as VXLAN or NVGRE.
- o The IP prefixes in the DC must be advertised with a flexibility that does not exist in IP-VPNs today. For instance:
 - a) The advertised overlay next-hop for a given IP prefix can be an IRB IP address (see section 5.3), a floating IP address (see section 5.2) or even an ESI (see section 5.4).

- b) As stated by [E-VPN-OVERLAYS], VXLAN or NVGRE virtual identifiers can have a global or a local scope. The implementation MUST support the flexibility to advertise IP Prefixes associated to a global identifier (32-bit value encoded in the E-VPN Ethernet Tag ID) or a locally significant identifier (20-bit value encoded in the MPLS label field). At the moment, [RFC4364] can only advertise Prefixes associated to a locally significant identifier (MPLS label).
- c) Since an NVE can potentially advertise many Prefixes with different overlay next-hops and different VXLAN/NVGRE identifiers, it is highly desirable to be able to advertise those prefixes with their corresponding overlay next-hop and VXLAN/NVGRE identifier within the same NLRI, for a better BGP update packing. [RFC4364] does not have the capability of advertising a flexible overlay next-hop together with a prefix in the same NLRI.
- o IP prefixes must be advertised by NVE devices that have no VRF instances defined and no capability to process IP-VPN prefixes. These NVE devices just support E-VPN and advertise IP Prefixes on behalf of some connected Tenant Systems. In other words: any attempt to solve this problem by simply using [RFC4364] routes requires that any EVPN deployment must be accompanied with a concurrent IP-VPN topology, which is not possible in most of the cases.
 - o Finally, Data Center providers want to use a single BGP Subsequent Address Family (AFI/SAFI) for the advertisement of addresses within the Data Center, i.e. BGP E-VPN only, as opposed to using E-VPN and IP-VPN in a concurrent topology. This minimizes the control plane overhead in TORs and Hypervisors and simplifies the operations.

E-VPN is extended - as described in this document - to advertise IP prefixes with the flexibility required by the current and future Data Center applications.

2.3 The requirement for a new E-VPN route type

[E-VPN] defines a MAC route (or route type 2) where a MAC address can be advertised together with an IP address length (IPL) and IP address (IP). While a variable IPL might be used to indicate the presence of an IP prefix in a route type 2, there are several specific use cases in which using this route type to deliver IP Prefixes is not suitable.

One example of such use cases is the "floating IP" example described in section 2.1. In this example we need to decouple the advertisement of the prefixes from the advertisement of the floating IP (vIP23 in figure 1) and MAC associated to it, otherwise the solution gets highly inefficient and does not scale.

E.g.: if we are advertising 1k prefixes from M2 (using route type 2) and the floating IP owner changes from M2 to M3, we would need to withdraw 1k routes from M2 and re-advertise 1k routes from M3. However if we use a separate route type, we can advertise the 1k routes associated to the floating IP address (vIP23) and only one route type 2 for advertising the ownership of the floating IP, i.e. vIP23 and M2 in the route type 2. When the floating IP owner changes from M2 to M3, a single route type 2 withdraw/update is required to indicate the change. The remote DGW will not change any of the 1k prefixes associated to vIP23, but will only update the ARP resolution entry for vIP23 (now pointing at M3).

Other reasons to decouple the IP Prefix advertisement from the MAC route are listed below:

- o Clean identification, operation of troubleshooting of IP Prefixes, not subject to interpretation and independent of the IPL and the IP value. E.g.: An IP address for ARP resolution must be always clearly distinguished from an /32 IP Prefix, or a default IP route 0.0.0.0/0 must always be easily and clearly distinguished from the absence of IP information.
- o MAC address information must not be compared by BGP when selecting two IP Prefix routes. If IP Prefixes are to be advertised using MAC routes, the MAC information is always present and part of the route key.
- o IP Prefix routes must not be subject to MAC route procedures such as MAC Mobility or aliasing. Prefixes advertised from two different ESIs do not mean mobility; MACs advertised from two different ESIs do mean mobility. Similarly load balancing for IP prefixes is achieved through IP mechanisms such as ECMP, and not through MAC route mechanisms such as aliasing.
- o NVEs that do not require processing IP Prefixes must have an easy way to identify an update with an IP Prefix and ignore it, rather than processing the MAC route only to find out later that it carries a Prefix that must be ignored.

The following sections describe how E-VPN is extended with a new route type for the advertisement of prefixes and how this route is used to address the current and future inter-subnet connectivity

requirements existing in the Data Center.

3. The BGP E-VPN IP Prefix route

The current BGP E-VPN NLRI as defined in [E-VPN] is shown below:

Route Type (1 octet)
Length (1 octet)
Route Type specific (variable)

Where the route type field can contain one of the following specific values:

- + 1 - Ethernet Auto-Discovery (A-D) route
- + 2 - MAC advertisement route
- + 3 - Inclusive Multicast Route
- + 4 - Ethernet Segment Route

This document defines an additional route type that will be used for the advertisement of IP Prefixes:

- + 5 - IP Prefix Route

The support for this new route type is OPTIONAL.

By using a separate route type for IP prefix advertisements, there is a clean separation of functions between route types, i.e. route type 2 or MAC Advertisement route will be used for MAC and ARP resolution advertisement, whereas route type 5 or IP Prefix route will be used for the advertisement of prefixes. Since this new route type is OPTIONAL, an implementation not supporting it will easily ignore the route, based on the route type value.

The detailed encoding of this route and associated procedures are described in the following sections.

3.1 IP Prefix Route encoding

An IP Prefix advertisement route type specific E-VPN NLRI consists of the following fields:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
IP Address Length (1 octet)
IP Address (4 or 16 octets)
GW IP Address (4 or 16 octets)
MPLS Label (3 octets)

Where:

- o RD, Ethernet Tag ID and MPLS Label fields will be used as defined in [E-VPN] and [E-VPN-OVERLAYS].
- o The Ethernet Segment Identifier will be a non-zero 10-byte identifier if the ESI is used as an overlay next-hop. It will be zero otherwise.
- o The IP address length can be set to a value between 0 and 32 (bits) for ipv4 and between 0 and 128 for ipv6.
- o The IP address will be a 32 or 128-bit field (ipv4 or ipv6).
- o The GW IP (Gateway IP Address) will be a 32 or 128-bit field (ipv4 or ipv6), and will encode the overlay IP next-hop for the IP Prefixes. The GW IP field can be zero if it is not used as an overlay next-hop.
- o The total route length will indicate the type of prefix (ipv4 or ipv6) and the type of GW IP address (ipv4 or ipv6). Note that the IP Address + the GW IP should have a length of either 64 or 256 bits, but never 160 bits (ipv4 and ipv6 mixed values are not allowed).

The Eth-Tag ID, IP address length and IP address will be part of the route key used by BGP to compare routes. The rest of the fields will be out of the route key.

The route will contain a single overlay next-hop, i.e. if the ESI field is zero, the GW IP field will not, and vice versa. The following table shows the different inter-subnet use-cases described

in this document and the corresponding coding of the overlay next-hop in the route-type 5.

Overlay next-hop use-case	Field in the route-type 5
TS IP address	GW IP Address
Floating IP address	GW IP Address
IRB IP address	GW IP Address
"Bump in the wire"	ESI

4. Benefits of using the E-VPN IP Prefix route

This section clarifies the different functions accomplished by the E-VPN route-type 2 and route-type 5 routes, and provides a list of benefits derived from using a separate route type for the advertisement of IP Prefixes in E-VPN.

[E-VPN] describes the content of the BGP E-VPN route type 2 specific NLRI, i.e. MAC Advertisement Route, where the IP address length (IPL) and IP address (IP) of a specific advertised MAC are encoded. The subject of the MAC advertisement route is the MAC address (M) and MAC address length (ML) encoded in the route. The MAC mobility and other complex procedures are defined around that MAC address. The IP address information carries the host IP address required for the ARP resolution of the MAC.

The BGP E-VPN route type 5 defined in this document, i.e. IP Prefix Advertisement route, decouples the advertisement of IP prefixes from the advertisement of any MAC address related to it. This brings some major benefits to NVO-based networks where inter-subnet forwarding is required. Some of those benefits are:

- a) Upon receiving a route type 2 or type 5, an egress NVE can easily distinguish MACs and IPs for ARP resolution from IP Prefixes. E.g. an IP prefix with IPL=32 being advertised from two different ingress NVEs (as route type 5) can be identified as such and be imported in the designated routing context as two ECMP routes, as opposed to two ARP entries competing for the same IP.
- b) Similarly, upon receiving a route, an egress NVE not supporting processing IP Prefixes can easily ignore the update, based on the route type.
- c) A MAC route includes the ML, M, IPL and IP in the route key that is used by BGP to compare routes, whereas for IP Prefix routes, only IPL and IP (as well as Ethernet Tag ID) are part of the route

key. Advertised IP Prefixes are imported into the designated routing context, where there is no MAC information associated to IP routes. In the example illustrated in figure 1, subnet SN1 should be advertised by NVE2 and NVE3 and interpreted by DGW1 as the same route coming from two different next-hops, regardless of the MAC address associated to TS2 or TS3. This is easily accomplished in the route type 5 by including only the IP information in the route key.

- d) By decoupling the MAC from the IP Prefix advertisement procedures, we can leave the IP prefix advertisements out of the MAC mobility procedures defined in [E-VPN] for MACs. In addition, this allows us to have an indirection mechanism for IP prefixes advertised from a MAC/IP that can move between hypervisors. E.g. if there are 1,000 prefixes seating behind TS2 (figure 1), NVE2 will advertise all those prefixes in type 5 routes associated to the next-hop IP2. Should TS2 move to a different NVE, a single MAC advertisement route withdraw for the M2/IP2 route from NVE2 will invalidate the 1,000 prefixes, as opposed to have to wait for each individual prefix to be withdrawn. This may be easily accomplished by using IP Prefix routes that are not tied to a MAC address, and use a different MAC route to advertise the location and resolution of the overlay next-hop to a MAC address.

5. IP Prefix next-hop use-cases

The IP Prefix route can use a GW IP or an ESI as an overlay next-hop. This section describes some use-cases for both next-hop types.

5.1 TS IP address next-hop use-case

The following figure illustrates an example of inter-subnet forwarding for subnets seating behind Virtual Appliances (on TS2 and TS3).

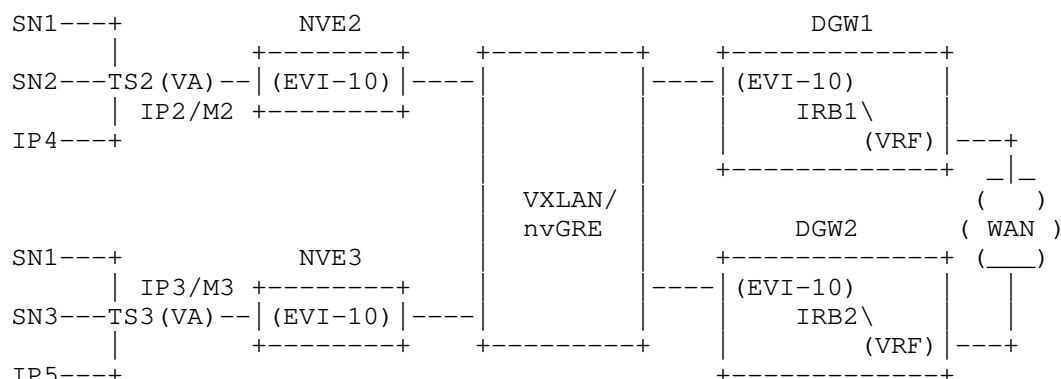


Figure 2 TS IP address use-case

An example of inter-subnet forwarding between subnet SN1/24 and a subnet seating in the WAN is described below. NVE2, NVE3, DGW1 and DGW2 are running BGP E-VPN. TS2 and TS3 do not support routing protocols, only a static route to forward the traffic to the WAN.

(1) NVE2 advertises the following BGP routes on behalf of TS2:

- o Route type 2 (MAC route) containing: ML=48, M=M2, IPL=32, IP=IP2
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP2

(2) NVE3 advertises the following BGP routes on behalf of TS3:

- o Route type 2 (MAC route) containing: ML=48, M=M3, IPL=32, IP=IP3
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP3

(3) DGW1 and DGW2 import both received routes based on the RT:

- o Based on the EVI-10 route-target in DGW1 and DGW2, the MAC route is imported and M2 is added to the EVI-10 MAC FIB along with its corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop (underlay next-hop) and VNI from the Ethernet Tag or MPLS fields (see [E-VPN-OVERLAYS]). IP2 - M2 is added to the ARP table.
- o Based on the EVI-10 route-target in DGW1 and DGW2, the IP

Prefix route is also imported and SN1/24 is added to the designated routing context with next-hop IP2 pointing at the local EVI-10. Should ECMP be enabled in the routing context, SN1/24 would also be added to the routing table with next-hop IP3.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop=IP2 is found. The tunnel information to encapsulate the packet will be derived from the route-type 2 (MAC route) received for M2/IP2.
- o IP2 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC FIB (remote VTEP and VNI for the VXLAN case).
- o The IP packet destined to IPx is encapsulated with:
 - . Source inner MAC = IRB1 MAC
 - . Destination inner MAC = M2
 - . Tunnel information provided by the MAC FIB (VNI, VTEP IPs and MACs for the VXLAN case)

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup.
- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(6) Should TS2 move from NVE2 to NVE3, MAC Mobility procedures will be applied to the MAC route IP2/M2, as defined in [EVPN]. Route type 5 prefixes are not subject to MAC mobility procedures, hence no changes in the DGW VRF routing table will occur for TS2 mobility, i.e. all the prefixes will still be pointing at IP2 as next-hop. There is an indirection for e.g. SN1/24, which still points at next-hop IP2 in the routing table, but IP2 will be simply resolved to a different tunnel, based on the outcome of the MAC mobility procedures for the MAC route IP2/M2.

Note that in the opposite direction, TS2 will send traffic based on its static-route next-hop information (IRB1 and/or IRB2), and regular

E-VPN procedures will be applied.

5.2 Floating IP next-hop use-case

Sometimes Tenant Systems (TS) work in active/standby mode where an upstream floating IP - owned by the active TS - is used as the next-hop to get to some subnets behind. This redundancy mode, already introduced in section 2.1 and 2.3, is illustrated in Figure 3.

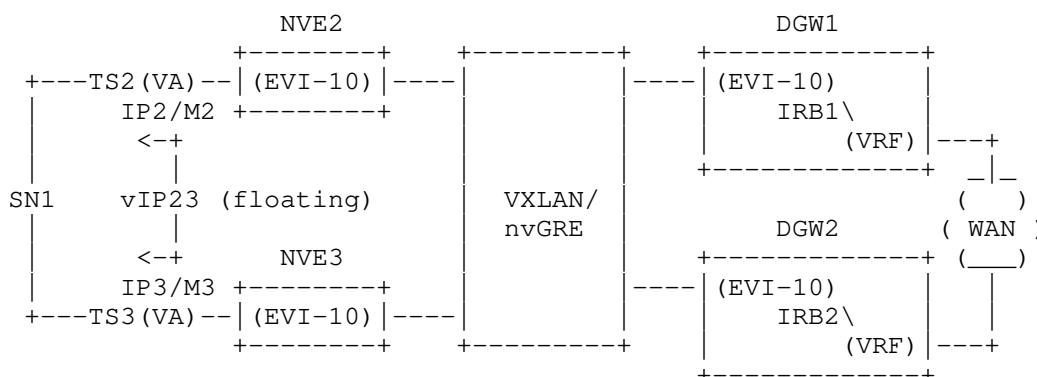


Figure 3 Floating IP next-hop for redundant TS

In this example, assuming TS2 is the active TS and owns IP23:

(1) NVE2 advertises the following BGP routes for TS2:

- o Route type 2 (MAC route) containing: ML=48, M=M2, IPL=32, IP=IP23
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP23

(2) NVE3 advertises the following BGP routes for TS3:

- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IP23

(3) DGW1 and DGW2 import both received routes based on the RT:

- o M2 is added to the EVI-10 MAC FIB along with its corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop and VNI from the Ethernet Tag or MPLS fields (see [E-VPN-OVERLAYS]). IP23 - M2 is added to the ARP table.
- o SN1/24 is added to the designated routing context in DGW1 and

DGW2 with next-hop IP23 pointing at the local EVI-10.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop=IP23 is found. The tunnel information to encapsulate the packet will be derived from the route-type 2 (MAC route) received for M2/IP23.
- o IP23 is resolved to M2 in the ARP table, and M2 is resolved to the tunnel information given by the MAC FIB (remote VTEP and VNI for the VXLAN case).
- o The IP packet destined to IPx is encapsulated with:
 - . Source inner MAC = IRB1 MAC
 - . Destination inner MAC = M2
 - . Tunnel information provided by the MAC FIB (VNI, VTEP IPs and MACs for the VXLAN case)

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup.
- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly routed.

(6) When the redundancy protocol running between TS2 and TS3 appoints TS3 as the new active TS for SN1, TS3 will now own the floating IP23 and will signal this new ownership (GARP message or similar). Upon receiving the new owner's notification, NVE3 will issue a route type 2 for M3-IP23. DGW1 and DGW2 will update their ARP tables with the new MAC resolving the floating IP. No changes are carried out in the VRF routing table.

In the DGW1/2 BGP RIB, there will be two route type 5 routes for SN1 (from NVE2 and NVE3) but only the one with the same BGP next-hop as the IP23 route type 2 BGP next-hop will be valid.

5.3 IRB IP next-hop use-case

In some other cases, the NVEs and DGWs will have just IRB interfaces as hosts in the E-VPN instance. Figure 4 illustrates an example.

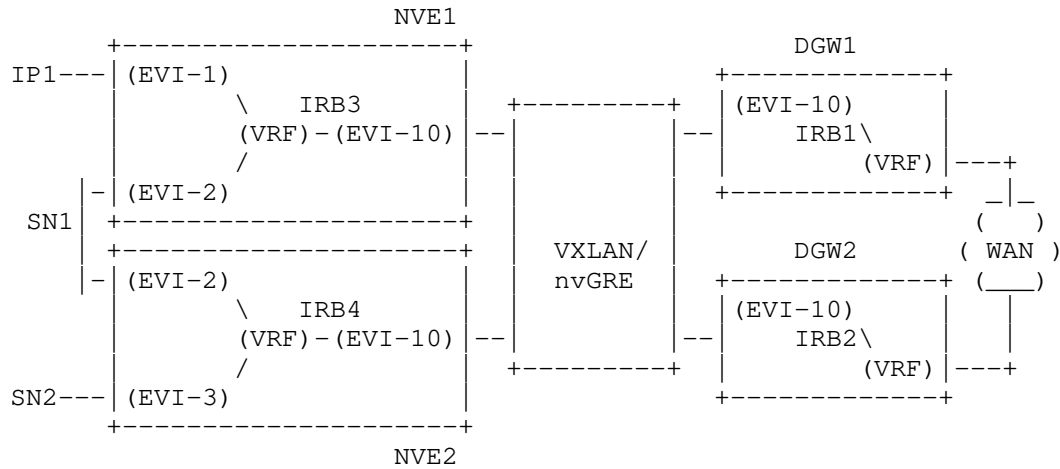


Figure 4 IRB IP next-hop use-case

In this case:

- (1) NVE1 advertises the following BGP routes for SN1 resolution:
 - o Route type 2 (MAC route) containing: ML=48, M=IRB3-MAC, IPL=32, IP=IRB3-IP
 - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IRB3-IP
- (2) NVE2 advertises the following BGP routes for SN1 resolution:
 - o Route type 2 (MAC route) containing: ML=48, M=IRB4-MAC, IPL=32, IP=IRB4-IP
 - o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=0, GW IP address=IRB4-IP
- (3) DGW1 and DGW2 import both received routes based on the RT:
 - o IRB3-MAC and IRB4-MAC are added to the EVI-10 MAC FIB along with their corresponding tunnel information. For the VXLAN use case, the VTEP will be derived from the MAC route BGP next-hop and VNI from the Ethernet Tag or MPLS fields (see [E-VPN-OVERLAYS]). IRB3-MAC - IRB3-IP and IRB4-MAC - IRB4-IP are added to the ARP table.

- o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop IRB3-IP (and/or IRB4-IP) pointing at the local EVI-10.

Similar forwarding procedures as the ones described in the previous use-cases are followed.

5.4 ESI next-hop ("Bump in the wire") use-case

The following figure illustrates an example of inter-subnet forwarding for a subnet route that uses an ESI as an overlay next-hop. In this use-case, TS2 and TS3 are layer-2 VA devices without any IP address that can be included as an overlay next-hop in the GW IP field of the IP Prefix route.

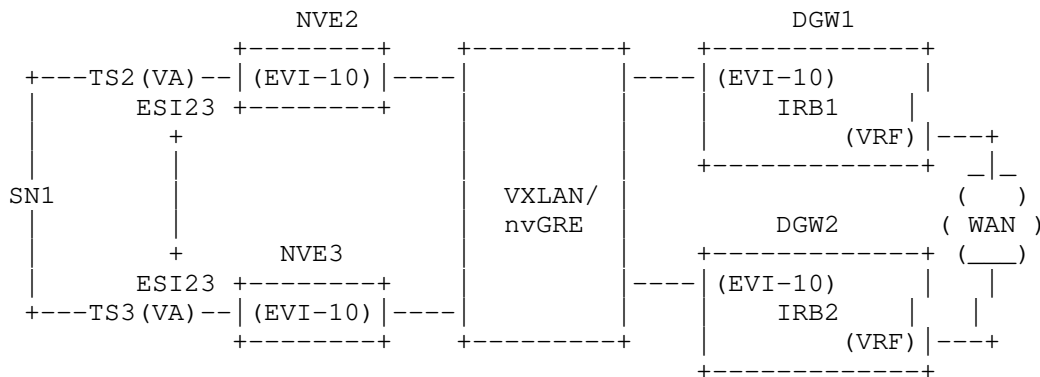


Figure 5 ESI next-hop use-case

Since neither TS2 nor TS3 can run any routing protocol and have no IP address assigned, an ESI, i.e. ESI23, will be provisioned on the attachment ports of NVE2 and NVE3. This model supports VA redundancy in a similar way as the one described in section 4.2 for the floating IP next-hop use-case, only using the E-VPN A-D route instead of the MAC advertisement route to advertise the location of the overlay next-hop. The procedure is explained below:

(1) NVE2 advertises the following BGP routes for TS2:

- o Route type 1 (A-D route for EVI-10) containing: ESI=ESI23 and the corresponding tunnel information (Ethernet Tag and/or MPLS label). Assuming the ESI is active on NVE2, NVE2 will advertise this route.
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=ESI23, GW IP address=0.

(2) NVE3 advertises the following BGP routes for TS3:

- o Route type 1 (A-D route for EVI-10) containing: ESI=ESI23 and the corresponding tunnel information (Ethernet Tag and/or MPLS label). NVE3 will advertise this route assuming the ESI is active on NVE2. Note that if the resiliency mechanism for TS2 and TS3 is in active-active mode, both NVE2 and NVE3 will send the A-D route. Otherwise, that is, the resiliency is active-standby, only the NVE owning the active ESI will advertise the A-D route for ESI23.
- o Route type 5 (IP Prefix route) containing: IPL=24, IP=SN1, ESI=23, GW IP address=0.

(3) DGW1 and DGW2 import the received routes based on the RT:

- o The tunnel information to get to ESI23 is installed in DGW1 and DGW2. For the VXLAN use case, the VTEP will be derived from the A-D route BGP next-hop and VNI from the Ethernet Tag or MPLS fields (see [E-VPN-OVERLAYS]).
- o SN1/24 is added to the designated routing context in DGW1 and DGW2 with next-hop ESI23 pointing at the local EVI-10.

(4) When DGW1 receives a packet from the WAN with destination IPx, where IPx belongs to SN1/24:

- o A destination IP lookup is performed on the DGW1 VRF routing table and next-hop=ESI23 is found. The tunnel information to encapsulate the packet will be derived from the route-type 1 (A-D route) received for ESI23.
- o The IP packet destined to IPx is encapsulated with:
 - . Source inner MAC = IRB1 MAC
 - . Destination inner MAC = M2 (this MAC will be looked up in the EVI-10 FDB using the ESI23 as the key for the lookup).
 - . Tunnel information provided by the A-D route for ESI23 (VNI, VTEP IP and MACs for the VXLAN case).

(5) When the packet arrives at NVE2:

- o Based on the tunnel information (VNI for the VXLAN case), the EVI-10 context is identified for a MAC lookup (assuming MAC disposition model).

- o Encapsulation is stripped-off and based on a MAC lookup (assuming MAC forwarding on the egress NVE), the packet is forwarded to TS2, where it will be properly forwarded.

(6) If the redundancy protocol running between TS2 and TS3 follows an active/standby model and there is a failure, appointing TS3 as the new active TS for SN1, TS3 will now own the connectivity to SN1 and will signal this new ownership (GARP message or similar). Upon receiving the new owner's notification, NVE3 will issue a route type 1 for ESI23, whereas NVE2 will withdraw its A-D route for ESI23. DGW1 and DGW2 will update their tunnel information to resolve ESI23. No changes are carried out in the VRF routing table.

In the DGW1/2 BGP RIB, there will be two route type 5 routes for SN1 (from NVE2 and NVE3) but only the one with the same BGP next-hop as the ESI23 route type 1 BGP next-hop will be valid.

6. Conclusions

A new E-VPN route type 5 for the advertisement of IP Prefixes is proposed in this document. This new route type will have a differentiated role from the route type 2, i.e. MAC advertisement route, and will address all the inter-subnet connectivity scenarios which are required in the Data Center, where the overlay next-hop can be an IP address or an ESI. As discussed throughout the document, IP-VPN cannot be used in an NVO-based DC to advertise IP Prefixes and the existing E-VPN route type 2 does not meet the requirements for all the DC use cases, therefore a new E-VPN route type is required.

This new E-VPN route type 5 decouples the IP Prefix advertisements from the MAC route advertisements in E-VPN, hence:

- a) Allows the clean and clear announcements of ipv4 or ipv6 prefixes in an NLRI with no MAC addresses in the route key, so that only IP information is used in BGP route comparisons.
- b) Since the route type is different from the MAC advertisement route, the advertisement of prefixes will be excluded from all the procedures defined for the advertisement of VM MACs, e.g. MAC Mobility or aliasing. As a result of that, the current E-VPN procedures do not need to be modified.
- c) Allows a flexible implementation where the prefix can be linked to different types of next-hops: MAC address, IP address, IRB IP address, ESI, etc. and these MAC or IP addresses do not need to reside in the advertising NVE.

- d) An E-VPN implementation not requiring IP Prefixes can simply discard them by looking at the route type value.

7. 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].

8. Security Considerations

9. IANA Considerations

10. References

10.1 Normative References

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

10.2 Informative References

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03.txt, work in progress, February, 2013

[E-VPN-OVERLAYS] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using E-VPN", draft-sd-l2vpn-evpn-overlay-01.txt, work in progress, February, 2013

11. Acknowledgments

The authors would like to thank Mukul Katiyar and Senthil Sathappan for their valuable feedback and contributions.

12. Authors' Addresses

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Wim Henderickx

Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Florin Balus
Nuage Networks
Email: florin@nuagenetworks.net

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Senad Palislamovic
Alcatel-Lucent
Email: senad.palislamovic@alcatel-lucent.com

L2VPN Workgroup
Internet Draft

Intended status: Informational

J. Uttaro
AT&T

A. Isaac
T. Boyes
Bloomberg

J. Rabadan
S. Palislamovic
W. Henderickx
F. Balus
Alcatel-Lucent

K. Patel
A. Sajassi
Cisco

Expires: April 24, 2014

October 21, 2013

Usage and applicability of BGP MPLS based Ethernet VPN
draft-rp-l2vpn-evpn-usage-01.txt

Abstract

This document discusses the usage and applicability of BGP MPLS based Ethernet VPN (E-VPN) in a simple and fairly common deployment scenario. The different E-VPN procedures will be explained on the example scenario, analyzing the benefits and trade-offs of each option. Along with [E-VPN], this document is intended to provide a simplified guide for the deployment of E-VPN in Service Provider networks.

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), 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/lid-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 December 28, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. Use-case scenario description	4
3. Provisioning Model	6
3.1. Common provisioning tasks	7
3.1.1. Non-service specific parameters	7
3.1.2. Service specific parameters	8
3.2. Service interface dependent provisioning tasks	8
3.2.1. VLAN-based service interface EVI	8
3.2.2. VLAN-bundle service interface EVI	9
3.2.3. VLAN-aware bundling service interface EVI	9
4. BGP E-VPN NLRI usage	9
5. MAC-based forwarding model use-case	10
5.1. E-VPN Network Startup procedures	10
5.2. VLAN-based service procedures	11
5.2.1. Service startup procedures	11
5.2.2. Packet walkthrough	12
5.3. VLAN-bundle service procedures	15
5.3.1. Service startup procedures	15
5.3.2. Packet Walkthrough	16
5.4. VLAN-aware bundling service procedures	16
5.4.1. Service startup procedures	17
5.4.2. Packet Walkthrough	17
6. MPLS-based forwarding model use-case	18

6.1. Impact of MPLS-based forwarding on the E-VPN network startup	19
6.2. Impact of MPLS-based forwarding on the VLAN-based service procedures	19
6.3. Impact of MPLS-based forwarding on the VLAN-bundle service procedures	20
6.4. Impact of MPLS-based forwarding on the VLAN-aware service procedures	20
7. Comparison between MAC-based and MPLS-based forwarding models	21
8. Traffic flow optimization	22
8.1. Control Plane Procedures	22
8.1.1. MAC learning options	22
8.1.2. Proxy-ARP	23
8.1.3. Unknown Unicast flooding suppression	24
8.1.4. Optimization of Inter-subnet forwarding	24
8.2. Packet Walkthrough Examples	25
8.2.1. Proxy-ARP example for CE2 to CE3 traffic	25
8.2.2. Flood suppression example for CE1 to CE3 traffic	26
8.2.3. Optimization of inter-subnet forwarding example for CE3 to CE2 traffic	26
9. Conventions used in this document	28
10. Security Considerations	28
11. IANA Considerations	28
12. References	28
12.1. Normative References	28
12.2. Informative References	29
13. Acknowledgments	29
14. Authors' Addresses	29

1. Introduction

This document complements [E-VPN] by discussing the applicability of the technology in a simple and fairly common deployment scenario, which is described in section 2.

After describing the topology of the use-case scenario and the characteristics of the service to be deployed, section 3 will describe the provisioning model, comparing the E-VPN procedures with the provisioning tasks required for other VPN technologies, such as VPLS or IP-VPN.

Once the provisioning model is analyzed, sections 4, 5 and 6 will describe the control plane and data plane procedures in the example scenario, for the two potential disposition/forwarding models: MAC-based and MPLS-based models. While both models can interoperate in the same network, each one has different trade-offs that are analyzed in section 7.

Finally, E-VPN provides some potential traffic flow optimization tools that are also described in section 8, in the context of the example scenario.

2. Use-case scenario description

The following figure depicts the scenario that will be referenced throughout the rest of the document.

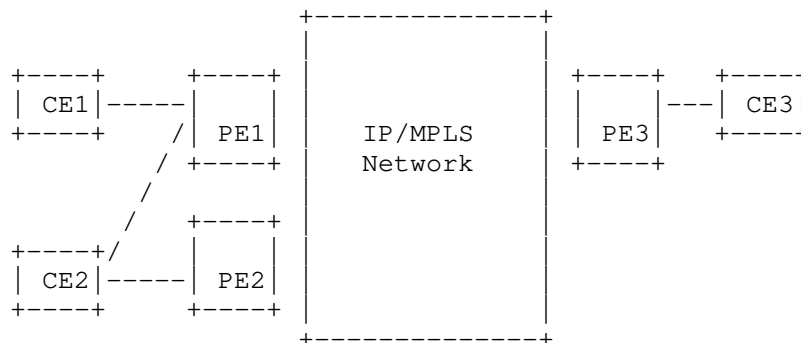


Figure 1 E-VPN use-case scenario

There are three PEs and three CEs considered in this example: PE1, PE2, PE3, as well as CE1, CE2 and CE3. Layer-2 traffic must be extended among the three CEs. The following service requirements are assumed in this scenario:

- o Redundancy requirements: CE1 and CE3 are single-homed to PE1 and PE3 respectively. CE2 requires multi-homing connectivity to PE1 and PE2, not only for redundancy purposes, but also for adding more upstream/downstream connectivity bandwidth to/from the network. If CE2 has a single CE-VID (or a few CE-VIDs) the current VPLS multi-homing solutions (based on load-balancing per CE-VID or service) do not provide the optimized link utilization required in this example. Another redundancy requirement that must be met is fast convergence. E.g.: if the link between CE2 and PE1 goes down, a fast convergence mechanism must be supported so that PE3 can immediately send the traffic to PE2, irrespectively of the number of affected services and MAC addresses. E-VPN provides the flow-based load-balancing multi-homing solution required in this scenario to optimize the upstream/downstream link utilization between CE2 and PE1-PE2. E-VPN also provides a fast convergence solution so that PE3 can immediately send the traffic to PE2 upon failure on the link between CE2 and PE1.
- o Service interface requirements: service definition must be flexible in terms of CE-VID-to-broadcast-domain assignment and service contexts in the core. The following three services are required in this example:

EVI100 - It will use VLAN-based service interfaces in the three CEs with a 1:1 mapping (VLAN-to-EVI). The CE-VIDs at the three CEs can be the same, e.g.: VID 100, or different at each CE, e.g.: VID 101 in CE1, VID 102 in CE2 and VID 103 in CE3. A single broadcast domain needs to be created for EVI100 in any case; therefore CE-VIDs will require translation at the egress PEs if they are not consistent across the three CEs. The case when the same CE-VID is used across the three CEs for EVI100 is referred in [E-VPN] as the "Unique VLAN" E-VPN case. This term will be used throughout this document too.

EVI200 - It will use VLAN-bundle service interfaces in CE1, CE2 and CE3, based on an N:1 VLAN-to-EVI mapping. In this case, the service provider just needs to assign a pre-configured number of CE-VIDs on the ingress PE to EVI200, and send the customer frames with the original CE-VIDs. The Service Provider will build a single broadcast domain for the customer. The customer will be responsible for the CE-VID handling.

EVI300 - It will use VLAN-aware bundling service interfaces in CE1, CE2 and CE3. At the ingress PE, an N:1 VLAN-to-EVI mapping will be done, however and as opposed to EVI200, a separate core broadcast domain is required per CE-VID. In addition to that, the CE-VIDs can be different (hence CE-VID translation is required). Note that, while the requirements stated for EVI100 and EVI200

might be met with the current VPLS solutions, the VLAN-aware bundling service interfaces required by EVI300 are not supported by the current VPLS tools.

NOTE: in section 3.2.1, only EVI100 is used as an example of VLAN-based service provisioning. In sections 5.2 and 6.2, 4k VLAN-based EVIs (EVI1 to EVI4k) are used so that the impact of MAC vs. MPLS disposition models in the control plane can be evaluated. In the same way, EVI200 and EVI300 will be described with a 4k:1 mapping (CE-VIDs-to-EVI mapping) in sections 5.3-4 and 6.3-4.

- o BUM (Broadcast, Unknown unicast, Multicast) optimization requirements: The solution must be able to support ingress replication, P2MP MPLS LSPs and MP2MP MPLS LSPs and the user must be able to decide what kind of provider tree will be used by each EVI service. For example, if we assume that EVI100 and EVI200 will not carry much BUM traffic, we can use ingress replication for those service instances. The benefit is that the core will not need to maintain any states for the multicast trees associated to EVI100 and EVI200. On the contrary, if EVI300 is presumably carrying a significant amount of multicast traffic, P2MP MPLS LSPs or MP2MP LSPs can be used for this service. Note that ingress replication and P2MP LSPs are supported by VPLS solutions (see [VPLS-MCAST]), however VPLS solutions do not support MP2MP LSPs, since the source of the tree must be identified for the data plane MAC learning, and that identification is challenging when using MP2MP LSPs. Since E-VPN uses the control plane for MAC learning, any type of provider multicast tree is supported in the core.

As already outlined above, the current VPLS solutions, based on [RFC4761][RFC4762][RFC6074], cannot meet all the above set of requirements and therefore a new solution is needed. The rest of the document will describe how E-VPN can be used to meet those service requirements and even optimize the network further by:

- o Providing the user with an option to reduce (and even suppress) the ARP-flooding.
- o Supporting ARP termination for inter-subnet forwarding

3. Provisioning Model

One of the requirements stated in [E-VPN-REQ] is the ease of provisioning. BGP parameters and service context parameters should be auto-provisioned so that the addition of a new MAC-VRF to the EVI requires a minimum number of single-sided provisioning touches. However this is only possible in a limited number of cases. This section describes the provisioning tasks required for the services

described in section 2, i.e. EVI100 (VLAN-based service interfaces), EVI200 (VLAN-bundle service interfaces) and EVI300 (VLAN-aware bundling service interfaces).

3.1. Common provisioning tasks

Regardless of the service interface type (VLAN-based, VLAN-bundle or VLAN-aware), the following sub-sections describe the parameters to be provisioned in the three PEs.

3.1.1. Non-service specific parameters

The multi-homing function in E-VPN requires the provisioning of certain parameters which are not service-specific and that are shared by all the MAC-VRFs in the node using the multi-homing capabilities. In our use-case, these parameters are only provisioned in PE1 and PE2, and are listed below:

- o Ethernet Segment Identifier (ESI): only the ESI associated to CE2 needs to be considered in our example. Single-homed CEs such as CE1 and CE3 do not require the provisioning of an ESI (the ESI will be coded as zero in the BGP NLRI). In our example, a LAG is used between CE2 and PE1-PE2 (since all-active multi-homing is a requirement) therefore the ESI can be auto-derived from the LACP information as described in [E-VPN]. Note that the ESI MUST be unique across all the PEs in the network, therefore the auto-provisioning of the ESI is only recommended in case the CEs are managed by the Service Provider. Otherwise the ESI should be manually provisioned in order to avoid potential conflicts.
- o ES-Import Route Target (ES-Import RT): this is the RT that will be sent by PE1 and PE2, along with the ES route. Regardless of how the ESI is provisioned in PE1 and PE2, the ES-Import RT must always be auto-derived from the 6-byte MAC address portion of the ESI value.
- o Ethernet Segment Route Distinguisher (ES RD): this is the RD to be encoded in the ES route and Ethernet Auto-Discovery (A-D) route to be sent by PE1 and PE2 for the CE2 ESI. This RD should always be auto-derived from the PE IP address, as described in [E-VPN].
- o Multi-homing type: the user must be able to provision the multi-homing type to be used in the network. In our use-case, the multi-homing type will be set to all-active for the CE2 ESI. This piece of information is encoded in the ESI Label extended community flags and sent by PE1 and PE2 along with the Ethernet A-D route for the CE2 ESI.

In our use-case, besides the above parameters, all the corresponding LAG and LACP parameters will be configured in PE1 and PE2, so that CE2 can send different flows to PE1 and PE2 for the same CE-VID as though they were forming a single system from the CE2 perspective.

3.1.2. Service specific parameters

The following parameters must be provisioned in PE1, PE2 and PE3 per EVI service:

- o EVI identifier: global identifier per EVI that is shared by all the PEs part of the EVI, i.e. PE1, PE2 and PE3 will be provisioned with EVI100, 200 and 300. The EVI identifier can be associated to (or be the same value as) the EVI default Ethernet Tag (4-byte default broadcast domain identifier for the EVI). The Ethernet Tag is different from zero in the E-VPN BGP routes only if the service interface type (of the source PE) is VLAN-aware.
- o EVI Route Distinguisher (EVI RD): This RD is a unique value across all the MAC-VRFs in a PE. Auto-derivation of this RD might be possible depending on the service interface type being used in the EVI. Next section discusses the specifics of each service interface type.
- o EVI Route Target(s) (EVI RT): one or more RTs can be provisioned per MAC-VRF. The RT(s) imported and exported can be equal or different, just as the RT(s) in IP-VPNs. Auto-derivation of this RT(s) might be possible depending on the service interface type being used in the EVI. Next section discusses the specifics of each service interface type.
- o CE-VID and port/LAG binding to EVI identifier or Ethernet Tag: see section 3.2.

3.2. Service interface dependent provisioning tasks

Depending on the service interface type being used in the EVI, a specific CE-VID binding provisioning must be specified.

3.2.1. VLAN-based service interface EVI

In our use-case, EVI100 is a VLAN-based service interface EVI.

EVI100 can be a "unique-VLAN" E-VPN if the CE-VID being used for this service in CE1, CE2 and CE3 is equal, e.g. VID 100. In that case, the VID 100 binding must be provisioned in PE1, PE2 and PE3 for EVI100 and the associated port or LAG. The MAC-VRF RD and RT can be auto-derived from the CE-VID:

- o The auto-derived MAC-VRF RD will be a Type 1 RD, as recommended in [E-VPN], and it will be comprised of [PE-IP]:[zero-padded-VID]; where PE-IP is the IP address of the PE (normally a loopback address) and [zero-padded-VID] is a 2-byte value where the low order 12 bits are the VID (VID 100 in our example) and the high order 4 bits are zero.
- o The auto-derived MAC-VRF RT will be composed of [AS]:[zero-padded-VID]; where AS is the Autonomous System that the PE belongs to and [zero-padded-VID] is a 4-byte value where the low order 12 bits are the VID (VID 100 in our example) and the high order 20 bits are zero. Note that auto-deriving the RT implies supporting a basic any-to-any topology in the EVI and using the same import and export RT in the EVI.

If EVI100 is not a "unique-VLAN" E-VPN, each individual CE-VID must be configured in each PE, and MAC-VRF RDs and RTs cannot be auto-derived, hence they must be provisioned by the user.

3.2.2. VLAN-bundle service interface EVI

Assuming EVI200 is a VLAN-bundle service interface EVI, and VIDs 200-250 are assigned to EVI200, the CE-VID bundle 200-250 must be provisioned on PE1, PE2 and PE3. Note that this model does not allow CE-VID translation and the CEs must use the same CE-VIDs for EVI200. No auto-derived EVI RDs or EVI RTs are possible.

3.2.3. VLAN-aware bundling service interface EVI

If EVI300 is a VLAN-aware bundling service interface EVI, CE-VID binding to EVI300 does not have to match on the three PEs (only on PE1 and PE2, since they are part of the same ES). E.g.: PE1 and PE2 CE-VID binding to EVI300 can be set to the range 300-310 and PE3 to 321-330. Note that each individual CE-VID will be assigned to a core broadcast domain, i.e. Ethernet Tag, which will be encoded in the BGP E-VPN routes.

Therefore, besides the CE-VID bundle range bound to EVI300 in each PE, associations between each individual CE-VID and the E-VPN Ethernet Tag must be provisioned by the user. No auto-derived EVI RDs/RTs are possible.

4. BGP E-VPN NLRI usage

[E-VPN] defines four different types of routes and four different extended communities advertised along with the different routes. However not all the PEs in a network must generate and process all the different routes and extended communities. The following table

shows the routes that must be exported and imported in the use-case described in this document. "Export", in this context, means that the PE must be capable of generating and exporting a given route, assuming there are no BGP policies to prevent it. In the same way, "Import" means the PE must be capable of importing and processing a given route, assuming the right RTs and policies. "N/A" means neither import nor export actions are required.

BGP E-VPN routes	PE1-PE2	PE3
ES	Export/import	N/A
A-D per ESI	Export/import	Import
A-D per EVI	Export/import	Import
MAC	Export/import	Export/import
Inclusive mcast	Export/import	Export/import

PE3 is only required to export MAC and Inclusive multicast routes and be able to import and process A-D routes, as well as MAC and Inclusive multicast routes. If PE3 did not support importing and processing A-D routes per ESI and per EVI, fast convergence and aliasing functions (respectively) would not be possible in this use-case.

5. MAC-based forwarding model use-case

This section describes how the BGP E-VPN routes are exported and imported by the PEs in our use-case, as well as how traffic is forwarded assuming that PE1, PE2 and PE3 support a MAC-based forwarding model. In order to compare the control and data plane impact in the two forwarding models (MAC-based and MPLS-based) and different service types, we will assume that CE1, CE2 and CE3 need to exchange traffic for up to 4k CE-VIDs.

5.1. E-VPN Network Startup procedures

Before any EVI is provisioned in the network, the following procedures are required:

- o Infrastructure setup: the proper MPLS infrastructure must be setup among PE1, PE2 and PE3 so that the E-VPN services can make use of P2P, P2MP and/or MP2MP LSPs. In addition to the MPLS transport, PE1 and PE2 must be properly configured to create a multi-chassis LAG to CE2. Details are provided in [E-VPN]. Once the LAG is properly setup, the ESI for the CE2 Ethernet Segment, e.g. ESI12, can be auto-generated by PE1 and PE2 from the LACP information exchanged with CE2, as discussed in section 3.1. Alternatively,

the ESI can also be manually provisioned on PE1 and PE2. PE1 and PE2 will auto-configure a BGP policy that will import any ES route matching the auto-derived ES-import RT for ESI12.

- o Ethernet Segment route exchange and DF election: PE1 and PE2 will advertise a BGP Ethernet Segment route for ESI12, where the ESI RD and ES-Import RT will be auto-generated as discussed in section 3.1.1. PE1 and PE2 will import the ES routes of each other and will run the DF election algorithm for any existing EVI (if any, at this point). PE3 will simply discard the route. Note that the DF election algorithm can support service carving, so that the downstream BUM traffic from the network to CE2 can be load-balanced across PE1 and PE2 on a per-service basis.

At the end of this process, the network infrastructure is ready to start deploying E-VPN services. PE1 and PE2 are aware of the existence of a shared Ethernet Segment, i.e. ESI12.

5.2. VLAN-based service procedures

Assuming that the E-VPN network must carry traffic among CE1, CE2 and CE3 for up to 4k CE-VIDs, the Service Provider can decide to implement VLAN-based service interface EVIs to accomplish it. In this case, each CE-VID will be individually mapped to a different EVI. While this means a total number of 4k MAC-VRFs is required per PE, the advantages of this approach are the auto-provisioning of most of the service parameters if no VLAN translation is needed (see section 3.2.1) and great control over each individual customer broadcast domain. We assume in this section that the range of EVIs from 1 to 4k is provisioned in the network.

5.2.1. Service startup procedures

As soon as the EVIs are created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (up to 4k routes per PE) so that the flooding tree per EVI can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use-case, since all the EVIs have the same core topology, PMSI aggregation makes sense in order to save some multicast forwarding states in the core.
- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a list of 4k RTs (one per EVI) and an ESI Label extended

community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used by the non-DF for split-horizon functions). These routes will be imported by the three PEs, since the RTs match the EVI RTs locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as discussed in [E-VPN].

- o Ethernet A-D routes per EVI (4k routes): An A-D route per EVI will be sent by PE1 and PE2 for ESI12. Each individual route includes the corresponding EVI RT and an MPLS label to be used by PE3 for the aliasing function. These routes will be imported by the three PEs.

5.2.2. Packet walkthrough

Once the services are setup, the traffic can start flowing. Assuming there are no MAC addresses learnt yet and that MAC learning at the access is performed in the data plane in our use-case, this is the process followed upon receiving packets from each CE (example for EVI1).

(1) BUM packet example from CE1:

- a) An ARP-request with CE-VID=1 is issued from source MAC CE1-MAC (MAC address coming from CE1 or from a device connected to CE1) to find the MAC address of CE3-IP.
- b) Based on the CE-VID, the packet is identified to be forwarded in the MAC-VRF-1 (EVI1) context. A source MAC lookup is done in the MAC FIB and in the proxy-ARP table within the MAC-VRF-1 (EVI1) context and if CE1-MAC is unknown in both tables, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port and CE-VID, (2) the ARP-request is snooped and the tuple CE1-MAC/CE1-IP is added to the proxy-ARP table and (3) a BGP MAC advertisement route is triggered from PE1 containing the EVI1 RD and RT, ESI=0, Ethernet-Tag=0 and CE1-MAC/CE1-IP along with an MPLS label assigned to MAC-VRF-1 from the PE1 label space. Since we assume a MAC forwarding model, a label per MAC-VRF is normally allocated and signaled by the three PEs for MAC advertisement routes. Based on the RT, the route is imported by PE2 and PE3 and the forwarding state plus ARP entry are added to their MAC-VRF-1 context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding for CE1-IP is not needed in the core.
- c) Since the ARP packet is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI1 (CE-VID=1 is kept if

translation is required). Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication and no aggregation, the packet is replicated to PE2 and PE3 with the downstream allocated labels and the P2P LSP transport labels. No other labels are added to the stack.

- d) Assuming PE1 is the DF for EVI1 on ESI12, the packet is locally replicated to CE2.
- e) The MPLS-encapsulated packet gets to PE2 and PE3. Since PE2 is non-DF for EVI1 on ESI12, and there is no other CE connected to PE2, the packet is discarded. At PE3, the packet is de-encapsulated, CE-VID translated if needed and replicated to CE3.

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet example from CE2:

- a) An ARP-request with CE-VID=1 is issued from source MAC CE2-MAC to find the MAC address of CE3-IP.
- b) CE2 will hash the packet and will forward it to e.g. PE2. Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB and in the proxy-ARP table within the MAC-VRF-1 context and if CE2-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI and CE-VID, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the proxy-ARP table and (3) a BGP MAC advertisement route is triggered from PE2 containing the EVI1 RD and RT, ESI=12, Ethernet-Tag=0 and CE2-MAC/CE2-IP along with an MPLS label assigned from the PE2 label space (one label per MAC-VRF). Note that, since PE3 is not part of ESI12, it will install a forwarding state for CE2-MAC as long as the A-D route per ESI for ESI12 is also active on PE3. On the contrary, PE1 is part of ESI12, therefore PE1 will not modify the forwarding state for CE2-MAC if it has previously learnt CE2-MAC locally attached to ESI12. Otherwise it will add forwarding state for CE2-MAC.
- c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI1 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI1, adding the ESI label for ESI12 at the bottom of the stack. The ESI label has been previously allocated and signaled by

the A-D routes for ESI12. Note that, as per [E-VPN], if the result of the CE2 hashing is different and the packet sent to PE1, PE1 MAY decide not to add the ESI label to the label stack (PE1 is the DF for EVI1 on ESI12).

- d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI identified by the ESI label. At PE3, the Inclusive multicast tree label(s) are popped and the packet forwarded to CE3. If a P2MP LSP is used as Inclusive multicast tree for EVI1, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will simply be ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

- a) A unicast packet with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE1-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE1-IP, and has added CE3-MAC/CE3-IP to its proxy-ARP table).
- b) Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB within the MAC-VRF-1 context and this time, since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and the label stack associated to the MAC CE1-MAC is found (including the label associated to MAC-VRF-1 in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1.
- c) At PE1, the packet is identified to be part of EVI1 (based on the bottom of the stack label) and a destination MAC lookup is performed in the MAC-VRF-1 context. The labels are popped and the packet forwarded to CE1 with CE-VID=1. Unicast packets from CE1 to CE3 or from CE2 to CE3 follow the same procedures described above.

(4) Unicast packet example from CE3 to CE2:

- a) A unicast packet with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE2-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE2-IP).
- b) Based on the CE-VID, the packet is identified to be forwarded in the MAC-VRF-1 context. A source MAC lookup is done in the MAC FIB within the MAC-VRF-1 context and since we assume CE3-MAC is known,

no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and PE3 finds CE2-MAC associated to PE2 on ESI12, an Ethernet Segment for which PE3 has two active A-D routes per ESI (from PE1 and PE2) and two active A-D routes for EVI1 (from PE1 and PE2). Based on a hashing function for the packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI1). Either way, the packet is encapsulated and sent to the remote PE.

- c) At PE2 (or PE1), the packet is identified to be part of EVI1 based on the bottom label, and a destination MAC lookup is performed. In particular, if the packet arrives to PE2, the bottom label is assumed to be a label per MAC-VRF, hence a MAC lookup for the MAC-VRF-1 context is done. If the packet arrives to PE1, the bottom label is assumed to be a label identifying ESI12, hence the packet is forwarded to ESI12.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

5.3. VLAN-bundle service procedures

Instead of using VLAN-based interfaces, the Service Provider can choose to implement VLAN-bundle interfaces to carry the traffic for the 4k CE-VIDs among CE1, CE2 and CE3. If that is the case, the 4k CE-VIDs can be mapped to the same EVI, e.g. EVI200, at each PE. The main advantage of this approach is the low control plane overhead (reduced number of routes and labels) and easiness of provisioning, at the expense of no control over the customer broadcast domains, i.e. a single inclusive multicast tree for all the CE-VIDs and no CE-VID translation in the Provider network.

5.3.1. Service startup procedures

As soon as the EVI200 is created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI (one route): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (hence only one route per PE) so that the flooding tree per EVI can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use-case, since all the CE-VIDs

are part of the same EVI, a single tree is created for all of them.

- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI200), an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the EVI200 RT locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [E-VPN].
- o Ethernet A-D routes per EVI (one route): An A-D route (EVI200) will be sent by PE1 and PE2 for ESI12. This route includes the EVI200 RT and an MPLS label to be used by PE3 for the aliasing function. This route will be imported by the three PEs.

5.3.2. Packet Walkthrough

The packet walkthrough for the VLAN-bundle case is similar to the one described for EVI1 in the VLAN-based case except for the way the CE-VID is handled by the ingress PE and the egress PE:

- o No VLAN translation is allowed and the CE-VIDs are kept untouched from CE to CE, i.e. the ingress CE-VID MUST be kept at the imposition PE and at the disposition PE.
- o The packet is identified to be forwarded in the MAC-VRF-200 context as long as its CE-VID belongs to the VLAN-bundle defined in the PE1/PE2/PE3 port to CE1/CE2/CE3. Our example is a special VLAN-bundle case, since the entire CE-VID range is defined in the ports, therefore any CE-VID would be part of EVI200.

Please refer to section 5.2.2 for more information about the control plane and forwarding plane interaction for BUM and unicast traffic from the different CEs.

5.4. VLAN-aware bundling service procedures

The last potential service type analyzed in this document is VLAN-aware bundling. When this type of service interface is used to carry the 4k CE-VIDs among CE1, CE2 and CE3, all the CE-VIDs will be mapped to the same EVI, e.g. EVI300. The difference, compared to the VLAN-bundle service type in the previous section, is that each incoming CE-VID will also be mapped to a different "normalized" Ethernet-Tag in addition to EVI300. If no translation is required, the Ethernet-tag will match the CE-VID. Otherwise a translation

between CE-VID and Ethernet-tag will be needed at the imposition PE and at the disposition PE. The main advantage of this approach is the ability to control customer broadcast domains while providing a single EVI to the customer.

5.4.1. Service startup procedures

As soon as the EVI300 is created in PE1, PE2 and PE3, the following control plane actions are carried out:

- o Flooding tree setup per EVI per Ethernet-Tag (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI and per Ethernet-Tag (hence 4k routes per PE) so that the flooding tree per customer broadcast domain can be setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use-case, since all the CE-VIDs and Ethernet-Tags are defined on the three PEs, multicast tree aggregation might make sense in order to save forwarding states.
- o Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI300), an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different than zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the EVI300 RT locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [E-VPN].
- o Ethernet A-D routes per EVI (one route): An A-D route (EVI300) will be sent by PE1 and PE2 for ESI12. This route includes the EVI300 RT and an MPLS label to be used by PE3 for the aliasing function. This route will be imported by the three PEs.

5.4.2. Packet Walkthrough

The packet walkthrough for the VLAN-aware case is similar to the one described before. Compared to the other two cases, VLAN-aware services allow for CE-VID translation and for an N:1 CE-VID to EVI mapping. Both things are not supported at once in either of the two other service interfaces. Note that this model requires qualified learning on the MAC FIBs. Some differences compared to the packet walkthrough described in section 5.2.2 are:

- o At the ingress PE, the packets are identified to be forwarded in the EVI300 context as long as their CE-VID belong to the range defined in the PE port to the CE. In addition to it, CE-VID=x is

mapped to a "normalized" Ethernet-Tag=y at the MAC-VRF-300 (where x and y might be equal if no translation is needed). Qualified learning is now required (a different FIB space is allocated within MAC-VRF-300 for each Ethernet-Tag). Potentially the same MAC could be learnt in two different Ethernet-Tag bridge domains of the same MAC-VRF.

- o Any new locally learnt MAC on the MAC-VRF-300/Ethernet-Tag=y interface is advertised by the ingress PE in a MAC advertisement route, using now the Ethernet-Tag field (Ethernet-Tag=y) so that the remote PE learns the MAC associated to the MAC-VRF-300/Ethernet-Tag=y FIB. Note that the Ethernet-Tag field is not used in advertisements of MACs learnt on VLAN-based or VLAN-bundle service interfaces.
- o At the ingress PE, BUM packets are sent to the corresponding flooding tree for the particular Ethernet-Tag they are mapped to. Although aggregated trees are supported, each individual Ethernet-Tag can have a different flooding tree within the same EVI300. For instance, Ethernet-Tag=y can use ingress replication to get to the remote PEs whereas Ethernet-Tag=z can use a p2mp LSP.
- o At the egress PE, Ethernet-Tag=y, for a given broadcast domain within MAC-VRF-300, can be translated to egress CE-VID=x. That is not possible for VLAN-bundle interfaces. It is possible for VLAN-based interfaces, but it requires a separate EVI per CE-VID.

6. MPLS-based forwarding model use-case

E-VPN supports an alternative forwarding model, usually referred to as MPLS-based forwarding or disposition model as opposed to the MAC-based forwarding or disposition model described in section 5. Using MPLS-based forwarding model instead of the MAC-based one might have an impact on:

- o The number of forwarding states required
- o The FIB where the forwarding states are handled: MAC FIB or MPLS LFIB.

The MPLS-based forwarding model avoids the destination MAC lookup at the egress PE MAC FIB, at the expense of increasing the number of next-hop forwarding states at the egress MPLS LFIB. This also has an impact on the control plane and the label allocation model, since an MPLS-based disposition PE MUST send as many routes and labels as required next-hops in the egress MAC-VRF. This concept is equivalent to the forwarding models supported in IP-VPNs at the egress PE, where an IP lookup in the IP-VPN FIB might be necessary or not depending on

the available next-hop forwarding states in the LFIB.

The following sub-sections highlight the impact on the control and data plane procedures described in section 5 when and MPLS-based forwarding model is used.

Note that both forwarding models are compatible and interoperable in the same network. The implementation of either model in each PE is a decision local to the PE node.

6.1. Impact of MPLS-based forwarding on the E-VPN network startup

The MPLS-based forwarding model has no impact on the procedures explained in section 5.1.

6.2. Impact of MPLS-based forwarding on the VLAN-based service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of number of routes, when all the service interfaces are VLAN-based. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (4k routes per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per EVI (4k routes per PE/ESI): no impact compared to the MAC-based model.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same MAC-VRF, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. E.g. if CE2 sends traffic from two different MACs to PE1, CE2-MAC1 and CE2-MAC2, the same MPLS label=x can be re-used for both MAC advertisements since they both share the same source ESI12. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment (even if only one label per ESI is enough)

- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

6.3. Impact of MPLS-based forwarding on the VLAN-bundle service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of number of routes when all the service interfaces are VLAN-bundle type. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (one route): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per EVI (one route per PE/ESI): no impact compared to the MAC-based model since no VLAN translation is required.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same MAC-VRF, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. E.g. if CE2 sends traffic from two different MACs to PE1, CE2-MAC1 and CE2-MAC2, the same MPLS label=x can be re-used for both MAC advertisements since they both share the same source ESI12. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment (even if only one label per ESI is enough).
- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

6.4. Impact of MPLS-based forwarding on the VLAN-aware service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has definitively an impact in terms of number of A-D routes

when all the service interfaces are VLAN-aware bundle type. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (4k routes per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.
- o Ethernet A-D routes per EVI (4k routes per PE/ESI): PE1 and PE2 will send 4k routes for EVI300, one per <ESI, Ethernet-Tag ID> tuple. This will allow the egress PE to find out all the forwarding information in the MPLS LFIB and even support Ethernet-Tag to CE-VID translation at the egress. The MAC-based forwarding model would allow the PEs to send a single route per PE/ESI for EVI300, since the packet with the embedded Ethernet-Tag would be used to perform a MAC lookup and find out the egress CE-VID.
- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same MAC-VRF, a different label MUST be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label at least per CE must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence less scalability) is required in terms of forwarding states. E.g. if CE2 sends traffic from two different MACs to PE1, CE2-MAC1 and CE2-MAC2, the same MPLS label=x can be re-used for both MAC advertisements since they both share the same source ESI12. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment (even if only one label per ESI is enough). Note that, in this model, the Ethernet-Tag will be set to a non-zero value for the MAC-advertisement routes. The same MAC address can be announced with different Ethernet-Tag value. This will make the advertising PE install two different forwarding states in the MPLS LFIB.
- o PE1, PE2 and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

7. Comparison between MAC-based and MPLS-based forwarding models

Both forwarding models are possible in a network deployment and each

one has its own trade-offs.

The MAC-based forwarding model can save A-D routes per EVI when VLAN-aware bundling services are deployed and therefore reduce the control plane overhead. A MAC FIB lookup at the egress PE is required in order to do so.

The MPLS-based forwarding model can save forwarding states at the egress PEs if labels per next hop CE (as opposed to per MAC) are implemented. No egress MAC lookup is required. An A-D route per <EVI, Ethernet-Tag> is required for VLAN-aware services, as opposed to an A-D route per EVI.

The following table summarizes the implementation details of both models for the VLAN-aware bundling service type.

4k CE-VID VLANs	MAC-based Model	MPLS-based Model
A-D routes/EVI	1 per ESI/EVI	4k per ESI/EVI
Egress PE Forwarding states	1 per MAC	1 per next-hop
Egress PE Lookups	2 (MPLS+MAC)	1 (MPLS)

The egress forwarding model is an implementation local to the egress PE and is independent of the model supported on the rest of the PEs, i.e. in our use-case, PE1, PE2 and PE3 could have either egress forwarding model without any dependencies.

8. Traffic flow optimization

In addition to the procedures described across sections 1 through 7, E-VPN [E-VPN] procedures allow for optimized traffic handling in order to minimize unnecessary flooding across the entire infrastructure. Optimization is provided through specific ARP termination and the ability to block unknown unicast flooding. Additionally, E-VPN procedures allow for intelligent, closest to the source, inter-subnet forwarding and solves the commonly known sub-optimal routing problem. Besides the traffic efficiency, ingress based inter-subnet forwarding also optimizes packet forwarding rules and implementation at the egress nodes as well. Details of these procedures are outlined in sections 8.1 and 8.2.

8.1. Control Plane Procedures

8.1.1. MAC learning options

The fundamental premise of [E-VPN] is the notion of a different approach to MAC address learning compared to traditional IEEE 802.1 bridge learning methods; specifically E-VPN differentiates between data and control plane driven learning mechanisms.

Data driven learning implies that there is no separate communication channel used to advertise and propagate MAC addresses. Rather, MAC addresses are learned through IEEE defined bridge-learning procedures as well as by snooping on DHCP and ARP requests. As different MAC addresses show up on different ports, the L2 FIB is populated with the appropriate MAC addresses.

Control plane driven learning implies that there is a communication channel could be either a control-plane protocol or a management-plane mechanism. In the context of E-VPN, two different learning procedures are defined, i.e. local and remote procedures:

- o Local learning defines the procedures used for learning the MAC addresses of network elements locally connected to a MAC-VRF. Local learning could be implemented through all three learning procedures: control plane, management plane as well as data plane. However, the expectation is that for most of the use cases, local learning through data plane should be sufficient.
- o Remote learning defines the procedures used for learning MAC addresses of network elements remotely connected to a MAC-VRF, i.e. far-end PEs. Remote learning procedures defined in [E-VPN] advocate using only control plane learning; specifically BGP. Through the use of BGP E-VPN NLRI, the remote PE has the capability of advertising all the MAC addresses present in its local FIB.

8.1.2. Proxy-ARP

In E-VPN, MAC addresses are advertised via the MAC Advertisement Route, as discussed in [E-VPN]. Optionally an IP address can be advertised along with the MAC address announcement. However, there are certain rules put in place in terms of IP address usage: if the MAC Advertisement Route contains an IP address, and the IP Address Length is 32 bits (or 128 in the IPv6 case), this particular IP address correlates directly with the advertised MAC address. Such advertisement allows us to build a proxy-ARP table populated with the IP<>MAC bindings received from all the remote nodes.

Furthermore, based on these bindings, a local MAC-VRF can now provide Proxy-ARP functionality for all ARP requests directed to the IP address pool learned through BGP. Therefore, the amount of unnecessary L2 flooding, ARP requests in this case, can be further

reduced by the introduction of Proxy-ARP functionality across all EVI MAC-VRFs.

8.1.3. Unknown Unicast flooding suppression

Given that all locally learned MAC addresses are advertised through BGP to all remote PEs, suppressing flooding of any Unknown Unicast traffic towards the remote PEs is a feasible network optimization.

The assumption in the use case is made that any network device that appears on a remote MAC-VRF will somehow signal its presence to the network. This signaling can be done through e.g. gratuitous ARPs. Once the remote PE acknowledges the presence of the node in the MAC-VRF, it will do two things: install its MAC address in its local FIB and advertise this MAC address to all other BGP speakers via E-VPN NLRI. Therefore, we can assume that any active MAC address is propagated and learnt through the entire EVI. Given that MAC addresses become pre-populated - once nodes are alive on the network - there is no need to flood any unknown unicast towards the remote PEs. If the owner of a given destination MAC is active, the BGP route will be present in the local RIB and FIB, assuming that the BGP import policies are successfully applied; otherwise, the owner of such destination MAC is not present on the network.

It is worth noting that unless: a) control or management plane learning is performed through the entire EVI or b) all the EVI-attached devices signal their presence when they come up (GARPs or similar), unknown unicast flooding MUST be enabled.

8.1.4. Optimization of Inter-subnet forwarding

In a scenario in which both L2 and L3 services are needed over the same physical topology, some interaction between E-VPN and IP-VPN is required. A common way of stitching the two service planes is through the use of an IRB interface, which allows for traffic to be either routed or bridged depending on its destination MAC address. If the destination MAC address is the one of the IRB interface, traffic needs to be passed through a routing module and potentially be either routed to a remote PE or forwarded to a local subnet. If the destination MAC address is not the one of the IRB, the MAC-VRF follows standard bridging procedures.

A typical example of E-VPN inter-subnet forwarding would be a scenario in which multiple IP subnets are part of a single or multiple EVIs, and they all belong to a single IP-VPN. In such topologies, it is desired that inter-subnet traffic can be efficiently routed without any tromboning effects in the network. Due to the overlapping physical and service topology in such scenarios,

all inter-subnet connectivity will be locally routed through the IRB interface.

In addition to optimizing the traffic patterns in the network, local inter-subnet forwarding also optimizes greatly the amount of processing needed to cross the subnets: standard VPLS to IP-VPN stitching through IRB interfaces forces the traffic to pass through IRB interfaces twice, once locally, as the traffic gets into the routing domain for a given IP VPN, and once remotely as the traffic exits the routing domain and enters the remote VPLS instance at the egress PE.

Through E-VPN MAC advertisements, the local PE learns the real destination MAC address associated with the remote IP address and the inter-subnet forwarding can happen locally. When the packet is received at the egress PE, it is directly mapped to an egress MAC-VRF, bypassing any egress IP-VPN processing.

Please refer to [EVPN-INTERSUBNET] for more information about the IP inter-subnet forwarding procedures in E-VPN.

8.2. Packet Walkthrough Examples

Assuming that the services are setup according to figure 1 in section 2, the following flow optimization processes will take place in terms of creating, receiving and forwarding packets across the network.

8.2.1. Proxy-ARP example for CE2 to CE3 traffic

Using figure 1 in section 2, consider EVI 400 residing on PE1, PE2 and PE3 connecting CE2 and CE3 networks. Also, consider that PE1 and PE2 are part of the all-active multi-homing ES for CE2, and that PE2 is elected designated-forwarder for EVI400. We assume that all the PEs implement the proxy-ARP functionality in the MAC-VRF-400 context.

In this scenario, PE3 will not only advertise the MAC addresses through the E-VPN MAC Advertisement Route but also IP addresses of individual hosts, i.e. /32 prefixes, behind CE3. Upon receiving the E-VPN routes, PE1 and PE2 will install the MAC addresses in the MAC-VRF-400 FIB and based on the associated received IP addresses, PE1 and PE2 can now build a proxy-ARP table within the context of MAC-VRF-400.

From the forwarding perspective, when a node behind CE2 sends a packet destined to a node behind CE3, it will first send an ARP request to e.g. PE2 (based on the result of the CE2 hashing). Assuming that PE2 has populated its proxy-ARP table for all active nodes behind the CE3, and that the IP address in the ARP message

matches the entry in the table, PE2 will respond to the ARP request with the actual MAC address on behalf of the node behind CE3.

Once the nodes behind CE2 learn the actual MAC address of the nodes behind CE3, all the MAC-to-MAC communications between the two networks will be unicast.

8.2.2. Flood suppression example for CE1 to CE3 traffic

Using figure 1 in section 2, consider EVI 500 residing on PE1 and PE3 connecting CE1 and CE3 networks. Consider that both PE1 and PE3 have disabled unknown unicast flooding for this specific EVI context. Once the network devices behind CE3 come online they will learn their MAC addresses and create local FIB entries for these devices. Note that local FIB entries could also be created through either a control or management plane between PE and CE as well. Consequently, PE3 will automatically create E-VPN Type 2 MAC Advertisement Routes and advertise all locally learned MAC addresses. The routes will also include the MPLS label associated with the corresponding egress MAC-VRF or egress next-hop, depending on the forwarding model scheme being used by PE3.

Given that PE1 automatically learns and installs all MAC addresses behind CE3, its MAC-VRF FIB will already be pre-populated with the respective next-hops and label assignments associated with the MAC addresses behind CE3. As such, as soon as the traffic sent by CE1 to nodes behind CE3 is received into the context of EVI 500, PE1 will push the MPLS Label(s) onto the original Ethernet frame and send the packet to the MPLS network. As usual, once PE3 receives this packet, and depending on the forwarding model, PE3 will either do a next-hop lookup in the EVI 500 context, or will just forward the traffic directly to the CE3. In the case that PE1 MAC-VRF-500 does not have a MAC entry for a specific destination that CE1 is trying to reach, PE1 will drop the packet since unknown unicast flooding is disabled.

Based on the assumption that all the MAC entries behind the CEs are pre-populated through gratuitous-ARP and/or DHCP requests, if one specific MAC entry is not present in the MAC-VRF-500 FIB on PE1, the owner of that MAC is not alive on the network behind the CE3, hence the traffic can be dropped at PE1 instead of be flooded and consume network bandwidth.

8.2.3. Optimization of inter-subnet forwarding example for CE3 to CE2 traffic

Using figure 1 in section 2 consider that there is an IP-VPN 666 context residing on PE1, PE2 and PE3 which connects CE1, CE2 and CE3 into a single IP-VPN domain. Also consider that there are two EVIs

present on the PEs, EVI 600 and EVI 60. Each IP subnet is associated to a different MAC-VRF context. Thus there is a single subnet, subnet 600, between CE1 and CE3 that is established through EVI 600. Similarly, there is another subnet, subnet 60, between CE2 and CE3 that is established through EVI 60. Since both subnets are part of the same IP VPN, there is a mapping of each EVI (or individual subnet) to a local IRB interface on the three PEs.

If a node behind CE2 wants to communicate with a node on the same subnet seating behind CE3, the communication flow will follow the standard E-VPN procedures, i.e. FIB lookup within the PE1 (or PE2) after adding the corresponding E-VPN label to the MPLS label stack (downstream label allocation from PE3 for EVI 60).

When it comes to crossing the subnet boundaries, the ingress PE implements local inter-subnet forwarding. For example, when a node behind CE2 (EVI 60) sends a packet to a node behind CE1 (EVI 600) the destination IP address will be in the subnet 600, but the destination MAC address will be the address of source node's default gateway, which in this case will be an IRB interface on PE1 (connecting EVI 60 to IP-VPN 666). Once PE1 sees the traffic destined to its own MAC address, it will route the packet to EVI 600, i.e. it will change the source MAC address to the one of the IRB interface in EVI 600 and change the destination MAC address to the address belonging to the node behind CE1, which is already populated in the MAC-VRF-600 FIB, either through data or control plane learning.

An important optimization to be noted is the local inter-subnet forwarding in lieu of IP VPN routing. If the node from subnet 60 (behind CE2) is sending a packet to the remote end node on subnet 600 (behind CE3), the mechanism in place still honors the local inter-subnet (inter-EVI) forwarding. In a typical IP-VPN-to-VPLS scenario, once the packet leaves the L2 domain on PE1, it would be routed through the IP-VPN procedures and consequently, through a remote PE3 IRB interface, routed back into the remote VPLS domain for further processing. However, in the E-VPN case, traffic locally routed and forwarded to the egress PE within the MAC-VRF context.

In our use-case, therefore, when node from subnet 60 behind CE2 sends traffic to the node on subnet 600 behind CE3, the destination MAC address is the PE1 MAC-VRF-60 IRB MAC address. However, once the traffic locally crosses EVIs, to EVI 600, via the IRB interface on PE1, the source MAC address is changed to that of the IRB interface and the destination MAC address is changed to the one advertised by PE3 via E-VPN and already installed in MAC-VRF-600. The rest of the forwarding through PE1 is using the MAC-VRF-600 forwarding context and label space.

Another very relevant optimization is due to the fact that traffic between PEs is forwarded through E-VPN, rather than through IP-VPN. In the example described above for traffic from EVI 60 on CE2 to EVI 600 on CE3, there is no need for IP-VPN processing on the egress PE3. Traffic is forwarded either to the EVI 600 context in PE3 for further MAC lookup and next-hop processing, or directly to the node behind CE3, depending on the egress forwarding model being used.

9. Conventions used in this document

In the examples, the following conventions are used:

- o CE-VIDs refer to the VLAN tag identifiers being used at CE1, CE2 and CE3 to tag customer traffic sent to the Service Provider E-VPN network
- o CE1-MAC, CE2-MAC and CE3-MAC refer to source MAC addresses "behind" each CE respectively. Those MAC addresses can belong to the CEs themselves or to devices connected to the CEs.
- o CE1-IP, CE2-IP and CE3-IP refer to IP addresses associated to the above MAC addresses.

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.

10. Security Considerations

11. IANA Considerations

12. References

12.1. Normative References

[RFC4761]Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762]Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private

LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC6074]Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

[RFC4364]Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

12.2. Informative References

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt, work in progress, July, 2013

[EVPN-REQ] A. Sajassi, R. Aggarwal et al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-02.txt

[VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et al., draft-ietf-l2vpn-vpls-mcast-13.txt

[EVPN-INTERSUBNET] Sajassi et al., "IP Inter-subnet forwarding in EVPN", draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-02.txt

13. Acknowledgments

The authors want to thank Giles Heron for his detailed review of the document.

This document was prepared using 2-Word-v2.0.template.dot.

14. Authors' Addresses

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Senad Palislamovic
Alcatel-Lucent
Email: senad.palislamovic@alcatel-lucent.com

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

Florin Balus
Alcatel-Lucent
Email: Florin.Balus@alcatel-lucent.com

Keyur Patel
Cisco
Email: keyupate@cisco.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

James Uttaro
AT&T
Email: uttaro@att.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Truman Boyes
Bloomberg
Email: tboyes@bloomberg.net

L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Cisco

Wim Henderickx
Alcatel-Lucent

Jim Uttaro
AT&T

Aldrin Isaac
Bloomberg

Expires: April 21, 2014

October 21, 2013

E-TREE Support in EVPN & PBB-EVPN
draft-sajassi-l2vpn-evpn-etree-02

Abstract

The Metro Ethernet Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet Tree (E-Tree). [ETREE-FMWK] proposes a solution framework for supporting this service in MPLS networks. This document discusses how those functional requirements can be easily met with EVPN.

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	3
2	E-Tree Scenarios and EVPN / PBB-EVPN Support	3
2.1	Scenario 1: Leaf OR Root site(s) per PE	3
2.2	Scenario 2: Leaf AND Root site(s) per PE	4
2.3	Scenario 3: Leaf AND Root site(s) per Ethernet Segment	4
3	Operation for EVPN	5
3.1	Known Unicast Traffic	5
3.2	BUM Traffic	6
3.3	E-TREE Traffic Flows for EVPN	7
3.3.1	E-Tree with MAC Learning	7
3.3.2	E-Tree without MAC Learning	8
4	Operation for PBB-EVPN	8
4.1	Known Unicast Traffic	9
4.2	BUM Traffic	9
5	Acknowledgement	10
6	Security Considerations	10
7	IANA Considerations	10
8	References	10
8.1	Normative References	10
8.2	Informative References	10
	Authors' Addresses	10

1 Introduction

The Metro Ethernet Forum (MEF) has defined a rooted-multipoint Ethernet service known as Ethernet Tree (E-Tree). In an E-Tree service, endpoints are labeled as either Root or Leaf sites. Root sites can communicate with all other sites. Leaf sites can communicate with Root sites but not with other Leaf sites.

[ETREE-FMWK] proposes the solution framework for supporting E-Tree service in MPLS networks. The document identifies the functional components of the overall solution to emulate E-Tree services in addition to Ethernet LAN (E-LAN) services on an existing MPLS network.

[EVPN] is a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reach-ability information over the MPLS/IP network. [PBB-EVPN] combines the functionality of EVPN with [802.1ah] Provider Backbone Bridging for MAC address scalability.

This document discusses how the functional requirements for E-Tree service can be easily met with EVPN and PBB-EVPN.

1.1 Terminology

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

2 E-Tree Scenarios and EVPN / PBB-EVPN Support

In this section, we will categorize support for E-Tree into three different scenarios, depending on the nature of the site association (Root/Leaf) per PE or per Ethernet Segment:

- Leaf OR Root site(s) per PE
- Leaf AND Root site(s) per PE
- Leaf AND Root site(s) per Ethernet Segment

2.1 Scenario 1: Leaf OR Root site(s) per PE

In this scenario, a PE may have Root sites OR Leaf sites for a given VPN instance, but not both concurrently. The PE may have both Root and Leaf sites albeit for different VPNs. Every Ethernet Segment

connected to the PE is uniquely identified as either a Root or a Leaf site.

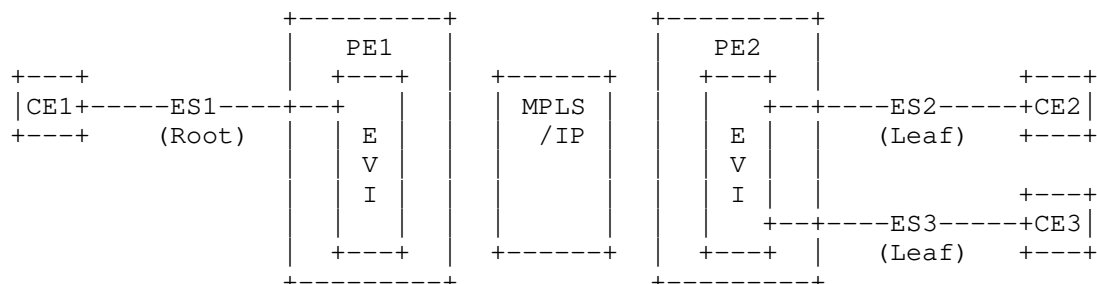


Figure 1: Scenario 1

2.2 Scenario 2: Leaf AND Root site(s) per PE

In this scenario, a PE may have a set of one or more Root sites AND a set of one or more Leaf sites for a given VPN instance. Every Ethernet Segment connected to the PE is uniquely identified as either a Root or a Leaf site.

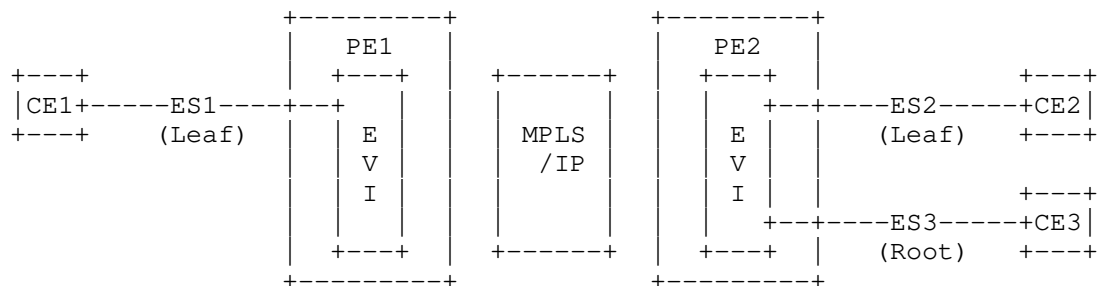


Figure 2: Scenario 2

2.3 Scenario 3: Leaf AND Root site(s) per Ethernet Segment

In this scenario, a PE may have a set of one or more Root sites AND a set of one or more Leaf sites for a given VPN instance. An Ethernet Segment connected to the PE may be identified as both a Root and a Leaf site concurrently.

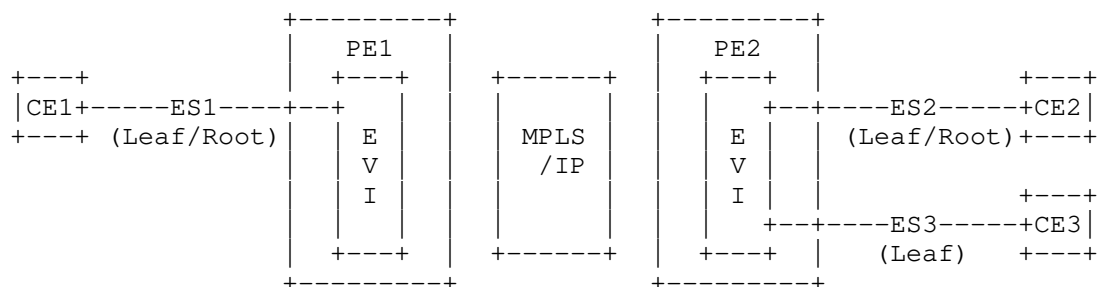


Figure 3: Scenario 3

3 Operation for EVPN

[EVPN] defines the notion of an Ethernet Segment which can be readily used to identify a Root and/or Leaf site in E-TREE services. In other words, [EVPN] has inherent capability to support E-TREE services without defining any new BGP routes. It only requires a minor modification to the existing procedures and a modification to a BGP attribute as shown in this section.

The following procedures are used consistently for all the scenarios highlighted in the previous section.

3.1 Known Unicast Traffic

For known unicast traffic, the PE must advertise a Root/Leaf indication along with each MAC Advertisement route, to indicate whether the associated MAC address was learnt from a Root or a Leaf. This enables remote PEs to perform ingress filtering for known unicast traffic: On the ingress PE, the MAC destination address lookup yields, in addition to the forwarding adjacency, a flag which indicates whether the target MAC is associated with a Root or a Leaf site. The ingress PE cross-checks this flag with the status of the originating site, and if both are a Leaf, then the packet is not forwarded.

The PE places all Leaf Ethernet Segments of a given bridge domain in a single split-horizon group in order to prevent intra-PE forwarding among Leaf segments. This split-horizon function applies to BUM traffic as well.

To support the above ingress filtering functionality, a new Root/Leaf indication flag will be added to the Tunnel Encapsulation Type Extended Community [RFC5512]. This extended community will be advertised with each EVPN MAC Advertisement route.

3.2 BUM Traffic

For BUM traffic, it is not possible to perform filtering on the ingress PE, as is the case with known unicast, because of the multi-destination nature of the traffic. As such, the solution relies on egress filtering. In order to apply the proper egress filtering, which varies based on whether a packet is sent from a Root or a Leaf, the MPLS-encapsulated frames MUST be tagged with an indication of whether they originated from a Root or a Leaf Ethernet Segment. This can be achieved in EVPN through the use of the ESI MPLS label, since this label identifies the Ethernet Segment of origin of a given frame. The egress PE determines whether or not to forward a particular frame to an Ethernet Segment depending on the split-horizon rule defined in [EVPN]:

- If the ESI Label indicates that the source Ethernet Segment is a Root, then the frame can be forwarded on a segment granted that it passes the split-horizon check.
- If the ESI Label indicates that the source Ethernet Segment is a Leaf, then the frame can be forwarded only on a Root segment, granted that it passes the split-horizon check.

When advertising the ESI MPLS label for a given Ethernet Segment, a PE must indicate whether the corresponding ESI is a Root or a Leaf site. This can be done by encoding the Root or Leaf indication in the Flags field of the ESI MPLS label Extended Community attribute ([EVPN] Section 8) to indicate Root/Leaf status.

In the case where a multi-homed Ethernet Segment has both Root and Leaf sites attached, two ESI MPLS labels are allocated and advertised: one ESI MPLS label denotes Root and the other denotes Leaf. The ingress PE imposes the right ESI MPLS label depending on whether the Ethernet frame originated from the Root or Leaf site on that Ethernet Segment. The mechanism by which the PE identifies whether a given frame originated from a Root or Leaf site on the segment is based on the Ethernet Tag associated with the frame. Other mechanisms of identification, beyond the Ethernet Tag, are outside the scope of this document. It should be noted that support for both Root and Leaf sites on a single Ethernet Segment requires that the PE performs the Ethernet Segment split-horizon check on a per Ethernet Tag basis. In the case where a multi-homed Ethernet Segment has either Root or Leaf sites attached, then a single ESI MPL label is allocated and advertised.

Furthermore, a PE advertises two special ESI MPLS labels: one for Root and another for Leaf. These are used by remote PEs for traffic originating from single-homed segments and for multi-homed segments

that are not connected to the advertising PE. Note that these special labels are advertised on a per PE basis (i.e. each PE advertises only two such special labels).

In addition to egress filtering (which is a MUST requirement), an EVPN PE implementation MAY provide topology constraint among the PEs belonging to the same EVI associated with an E-TREE service. The purpose of this topology constraint is to avoid having PEs with only host Leaf sites importing and processing BGP MAC routes from each other, thereby unnecessarily exhausting their RIB tables. However, as soon as a Root site is added to a Leaf PE, then that PE needs to process MAC routes from all other Leaf PEs and add them to its forwarding table. To support such topology constrain in EVPN, two BGP Route-Targets (RTs) are used for every EVPN Instance (EVI): one RT is associated with the Root sites and the other is associated with the Leaf sites. On a per EVI basis, every PE exports the single RT associated with its type of site(s). Furthermore, a PE with Root site(s) imports both Root and Leaf RTs, whereas a PE with Leaf site(s) only imports the Root RT. If for a given EVI, the PEs can eventually have both Leaf and Root sites attached, even though they may start as Root-only or Leaf-only PEs, then it is recommended to use a single RT per EVI and avoid additional configuration and operational overhead. If the number of EVIs is very large (e.g., more than 32K or 64K), then RT type 0 as defined in [RFC4360] SHOULD be used; otherwise, RT type 2 is sufficient.

3.3 E-TREE Traffic Flows for EVPN

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

- Ethernet Unicast from Root to Roots & Leaf
- Ethernet Unicast from Leaf to Root
- Ethernet Broadcast/Multicast from Root to Roots & Leafs
- Ethernet Broadcast/Multicast from Leaf to Roots

A particular E-Tree service may need to support all of the above types of flows or only a select subset, depending on the target application. In the case where unicast flows need not be supported, the L2VPN PEs can avoid performing any MAC learning function.

In the subsections that follow, we will describe the operation of EVPN to support E-Tree service with and without MAC learning.

3.3.1 E-Tree with MAC Learning

The PEs implementing an E-Tree service must perform MAC learning when

unicast traffic flows must be supported from Root to Leaf or from Leaf to Root sites. In this case, the PE with Root sites performs MAC learning in the data-path over the Ethernet Segments, and advertises reachability in EVPN MAC Advertisement routes. These routes will be imported by PEs that have Leaf sites as well as by PEs that have Root sites, in a given EVI. Similarly, the PEs with Leaf sites perform MAC learning in the data-path over their Ethernet Segments, and advertise reachability in EVPN MAC Advertisement routes which are imported only by PEs with at least one Root site in the EVI. A PE with only Leaf sites will not import these routes. PEs with Root and/or Leaf sites may use the Ethernet A-D routes for aliasing (in the case of multi-homed segments) and for mass MAC withdrawal.

To support multicast/broadcast from Root to Leaf sites, either a P2MP tree rooted at the PE(s) with the Root site(s) or ingress replication can be used. The multicast tunnels are set up through the exchange of the EVPN Inclusive Multicast route, as defined in [EVPN].

To support multicast/broadcast from Leaf to Root sites, ingress replication should be sufficient for most scenarios where there is a single Root or few Roots. If the number of Roots is large, a P2MP tree rooted at the PEs with Leaf sites may be used.

3.3.2 E-Tree without MAC Learning

The PEs implementing an E-Tree service need not perform MAC learning when the traffic flows between Root and Leaf sites are multicast or broadcast. In this case, the PEs do not exchange EVPN MAC Advertisement routes. Instead, the Ethernet A-D routes are used to exchange the EVPN labels.

The fields of the Ethernet A-D route are populated per the procedures defined in [EVPN], and the route import rules are as described in previous sections.

4 Operation for PBB-EVPN

In PBB-EVPN, the PE must advertise a Root/Leaf indication along with each MAC Advertisement route, to indicate whether the associated B-MAC address corresponds to a Root or a Leaf site. Similar to the EVPN case, this flag will be added to the Tunnel Encapsulation Type Extended Community [RFC5512], and advertised with each MAC Advertisement route.

In the case where a multi-homed Ethernet Segment has both Root and Leaf sites attached, two B-MAC addresses are allocated and advertised: one B-MAC address denotes Root and the other denotes Leaf. The ingress PE uses the right B-MAC source address depending on

whether the Ethernet frame originated from the Root or Leaf site on that Ethernet Segment. The mechanism by which the PE identifies whether a given frame originated from a Root or Leaf site on the segment is based on the Ethernet Tag associated with the frame. Other mechanisms of identification, beyond the Ethernet Tag, are outside the scope of this document. It should be noted that support for both Root and Leaf sites on a single Ethernet Segment requires that the PE performs the Ethernet Segment split-horizon check on a per Ethernet Tag basis. In the case where a multi-homed Ethernet Segment has either Root or Leaf sites attached, then a single B-MAC address is allocated and advertised per segment.

Furthermore, a PE advertises two global B-MAC addresses: one for Root and another for Leaf, and tags them as such in the MAC Advertisement routes. These B-MAC addresses are used as source addresses for traffic originating from single-homed segments.

4.1 Known Unicast Traffic

For known unicast traffic, the PEs perform ingress filtering: On the ingress PE, the C-MAC destination address lookup yields, in addition to the target B-MAC address and forwarding adjacency, a flag which indicates whether the target B-MAC is associated with a Root or a Leaf site. The ingress PE cross-checks this flag with the status of the originating site, and if both are a Leaf, then the packet is not forwarded.

The PE places all Leaf Ethernet Segments of a given bridge domain in a single split-horizon group in order to prevent intra-PE forwarding among Leaf segments. This split-horizon function applies to BUM traffic as well.

4.2 BUM Traffic

For BUM traffic, the PEs must perform egress filtering. When a PE receives a MAC advertisement route, it updates its Ethernet Segment egress filtering function (based on the B-MAC source address), as follows:

- If the MAC Advertisement route indicates that the advertised B-MAC is a Leaf, and the local Ethernet Segment is a Leaf as well, then the source B-MAC address is added to the B-MAC filtering list.
- Otherwise, the B-MAC filtering list is not updated.

When the egress PE receives the packet, it examines the B-MAC source address to check whether it should filter or forward the frame. Note that this uses the same filtering logic as baseline [PBB-EVPN] and

does not require any additional flags in the data-plane.

5 Acknowledgement

We would like to thank Sami Boutros and Dennis Cai for their comments.

6 Security Considerations

Same security considerations as [EVPN].

7 IANA Considerations

Allocation of Extended Community Type and Sub-Type for EVPN.

8 References

8.1 Normative References

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

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

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

8.2 Informative References

[ETREE-FMWK] Key et al., "A Framework for E-Tree Service over MPLS Network", draft-ietf-l2vpn-etree-frwk-03, work in progress, September 2013.

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt, work in progress, July, 2013.

[PBB-EVPN] Sajassi et al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-05.txt, work in progress, October, 2013.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

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

Jim Uttaro
AT&T
Email: jul738@att.com

Aldrin
Bloomberg Issac
Email: aisaac71@bloomberg.net

Sami Boutros
Cisco
Email: sboutros@cisco.com

INTERNET-DRAFT
Intended Status: Standard Track

Ali Sajassi
Samer Salam
Cisco

Nick Del Regno
Verizon

Expires: April 21, 2014

October 21, 2013

(PBB-)EVPN Seamless Integration with (PBB-)VPLS
draft-sajassi-l2vpn-evpn-vpls-integration-00

Abstract

This draft discusses the backward compatibility of the (PBB-)EVPN solution with (PBB-)VPLS and provides mechanisms for seamless integration of the two technologies in the same MPLS/IP network on a per-VPN-instance basis.

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	3
2.	Requirements	3
3	PBB-VPLS Integration with PBB-EVPN	4
3.1	Capability Discovery	4
3.2	Forwarding Setup and Unicast Operation	5
3.3	Multicast Operation	6
3.3.1	Ingress Replication	6
3.3.2	LSM	6
4	VPLS Integration with EVPN	6
4.1	Capability Discovery	6
4.2	Forwarding Setup and Unicast Operation	6
4.3	Multicast Operation	7
4.3.1	Ingress Replication	7
4.3.2	LSM	7
5	VPLS Integration with PBB-EVPN	7
5.1	Capability Discovery	7
5.2	Forwarding Setup and Unicast Operation	7
5.3	Multicast Operation	7
5.3.1	Ingress Replication	7
5.3.2	LSM	7
6	Solution Advantages	7
7	Security Considerations	8
8	IANA Considerations	8
9	References	8
9.1	Normative References	8
9.2	Informative References	8
	Authors' Addresses	9

1 Introduction

VPLS and PBB-VPLS are widely-deployed L2VPN technologies. Many SPs who are looking at adopting EVPN and PBB-EVPN want to preserve their investment in the (PBB-)VPLS networks. Hence, it is required to provide mechanisms by which (PBB-)EVPN technology can be introduced into existing L2VPN networks without requiring a fork-lift upgrade. This document discusses mechanisms for the seamless integration of the two technologies in the same MPLS/IP network.

Section 2 provides the details of the requirements. Section 3 discusses PBB-VPLS integration with PBB-EVPN. Section 4 discusses the integration of VPLS and EVPN. Section 5 discusses the integration of VPLS and PBB-EVPN, and finally Section 6 discusses the solution advantages.

It is worth noting that the scenario where PBB-VPLS is integrated with EVPN, is for future study and upon market validation. The reason for that is that deployments which employ PBB-VPLS typically require PBB encapsulation for various reasons. Hence, it is expected that for those deployments the evolution path would be from PBB-VPLS towards PBB-EVPN, rather than EVPN.

1.1 Terminology

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

2. Requirements

Following are the key requirements for backward compatibility between (PBB-)EVPN and (PBB-)VPLS:

1. The solution MUST allow for staged migration towards (PBB-)EVPN on a site-by-site basis per VPN instance - e.g., new EVPN sites to be provisioned on (PBB-)EVPN PEs.
2. The solution MUST require no changes to existing VPLS or PBB-VPLS PEs, not even a software upgrade.
3. The solution MUST allow for the coexistence of PE nodes running (PBB-)EVPN and (PBB-)VPLS for the same VPN instance and single-homed segments.
4. The solution MUST support single-active redundancy of multi-homed networks and multi-homed devices for (PBB-)EVPN PEs.

5. In case of single-active redundancy, the participant VPN instances MAY span across both (PBB-)EVPN PEs and (PBB-)VPLS PEs as long as single-active redundancy is employed by (PBB-)EVPN PEs.

6. The solution SHOULD support all-active redundancy of multi-homed networks and multi-homed devices for (PBB-)EVPN PEs.

7. In case of all-active redundancy, the participant VPN instances SHOULD be confined to (PBB-)EVPN PEs only.

8. In case of all-active redundancy, the participant VPN instances MAY span across (PBB-)EVPN and (PBB-)VPLS PEs.

These requirements collectively allow for the seamless insertion of the (PBB-)EVPN technology into brown-field (PBB-)VPLS deployments.

3 PBB-VPLS Integration with PBB-EVPN

In order to support seamless integration with (PBB-)VPLS, the (PBB-)EVPN PEs MUST support EVPN BGP routes (EVPN SAFI) and SHOULD support VPLS AD route (VPLS SAFI). All the logic for the integration will reside on the (PBB-)EVPN PEs side. However, if a VPLS instance is setup without the use of BGP auto-discovery, it is still possible (but cumbersome) for (PBB-)EVPN PEs to integrate into that VPLS instance.

3.1 Capability Discovery

The (PBB-)EVPN PEs must advertise both the BGP VPLS auto-discovery (AD) route as well as the BGP EVPN Inclusive Multicast route for a given VPN instance. The (PBB-)VPLS PEs only advertise the BGP VPLS AD route, per current standard procedures specified in [RFC4761] and [RFC6074]. The operator may decide to use the same BGP RT for both (PBB-)EVPN and (PBB-)VPLS. In this case, when a (PBB-)VPLS PE receives the EVPN Inclusive Multicast route, it will ignore it on the basis that it belongs to an unknown SAFI. However, the operator may use two RTs (one for (PBB-)VPLS and another for (PBB-)EVPN) and employ RT-constraint in order to prevent EVPN BGP routes from reaching the (PBB-)VPLS PEs. This provides an optimization in case required by the scale of the network.

When a (PBB-)EVPN PE receives both a VPLS AD route as well as an EVPN Inclusive Multicast route from a given remote PE for the same VPN instance, it MUST give preference to the EVPN route for the purpose of discovery. This ensures that, at the end of the route exchanges, all (PBB-)EVPN capable PEs discover other (PBB-)EVPN capable PEs as well as the (PBB-)VPLS-only PEs for that VPN instance. Furthermore, all the (PBB-)VPLS-only PEs would discover the (PBB-)EVPN PEs as if

they were standard (PBB-)VPLS nodes. In other words, when the discovery phase is complete, the (PBB-)EVPN PE would have discovered all the PEs in the VPN instance, and their associated capability: (PBB-)EVPN or VPLS-only. Whereas the (PBB-)VPLS PE would have discovered all the PEs in the VPN instance, as if they were all VPLS-only nodes.

3.2 Forwarding Setup and Unicast Operation

The procedures for forwarding setup and unicast operation on the (PBB-)VPLS PE are per [RFC4447] and [PBB-VPLS].

The procedures for forwarding state setup and unicast operation on the (PBB-)EVPN PE are as follows:

- The (PBB-)EVPN PE must establish a pseudowire to a remote PE from which it has received only a VPLS AD route, for the VPN instance in question, and set up the label stack corresponding to the pseudowire FEC. This PW is between B-components of PBB-EVPN PE and PBB-VPLS PE per section 4 of [PBB-VPLS-PE-MODEL].
- The (PBB-)EVPN PE must set up the label stack corresponding to the MP2P (PBB-)VPN unicast FEC to any remote PE that has advertised EVPN AD route.
- If a (PBB-)EVPN PE receives a VPLS AD route followed by an EVPN AD route from the same PE and a pseudowire is setup to that PE, then the (PBB-)EVPN PE MUST deactivate that pseudowire.
- If a (PBB-)EVPN PE receives an EVPN AD route followed by a VPLS AD route from the same PE, then the (PBB-)EVPN PE MUST ignore the VPLS AD route.

When the (PBB-)EVPN PE receives traffic over the pseudowires, it learns the associated MAC addresses in the data-plane. This is analogous to dynamic learning in IEEE bridges. The (PBB-)EVPN PE learns MAC addresses in the control plane, via the EVPN MAC Advertisement routes sent by remote (PBB-)EVPN PEs, and updates its MAC forwarding table accordingly. This is analogous to static learning in IEEE bridges. In PBB-EVPN, a given B-MAC address can be learnt either over the BGP control-plane from a remote PBB-EVPN PE, or in the data-plane over a pseudowire from a remote PBB-VPLS PE. There is no mobility associated with B-MAC addresses in this context. Hence, when the same B-MAC address shows up behind both a remote PBB-VPLS PE as well as a PBB-EVPN PE, the local PE can deduce that there is an anomaly in the network.

3.3 Multicast Operation

3.3.1 Ingress Replication The procedures for multicast operation on the (PBB-)VPLS PE, using ingress replication, are per [RFC4447] and [PBB-VPLS].

The procedures for multicast operation on the PBB-EVPN PE, for ingress replication, are as follows:

- The PBB-EVPN PE builds a replication sub-list per I-SID to all the remote PBB-EVPN PEs in a given VPN instance, as a result of the exchange of the EVPN Inclusive multicast routes, as described in [PBB-EVPN]. This will be referred to as sub-list A. It comprises MP2P tunnels used for delivering PBB-EVPN BUM traffic [EVPN].
- The PBB-EVPN PE builds a replication sub-list per VPN instance to all the remote PBB-VPLS PEs, as a result of the exchange of the VPLS AD routes. This will be referred to as sub-list B. It comprises pseudowires from the PBB-EVPN PE in question to all the remote PBB-VPLS PEs in the same VPN instance.
- The PBB-EVPN PE may further prune sub-list B, on a per I-SID basis, if [MMRP] is run over the PBB-VPLS network. This will be referred to as sub-list C. This list comprises a pruned set of the pseudowires in sub-list B.

The replication list, maintained per I-SID, on a given PBB-EVPN PE will be the union of sub-list A and sub-list B if [MMRP] is NOT used, and the union of sub-list A and sub-list C if [MMRP] is used. Note that the PE must enable split-horizon over all the entries in the replication list, across both pseudowires and MP2P tunnels.

3.3.2 LSM Will be covered in a future revision of this document.

4 VPLS Integration with EVPN

4.1 Capability Discovery

The procedures for capability discovery are per Section 3.1 above.

4.2 Forwarding Setup and Unicast Operation

The operation here is largely similar to that of PBB-EVPN integration with PBB-VPLS, with the exception of the need to handle MAC mobility, the details of which will be covered in a future revision of this document.

4.3 Multicast Operation

4.3.1 Ingress Replication

The operation is per the procedures of Section 3.3.1 above for the scenario WITHOUT [MMRP]. The replication list is maintained per VPN instance, rather than per I-SID.

4.3.2 LSM Will be covered in a future revision of this document.

5 VPLS Integration with PBB-EVPN

5.1 Capability Discovery

The procedures for capability discovery are per Section 3.1 above.

5.2 Forwarding Setup and Unicast Operation

The operation here is largely similar to that of PBB-EVPN integration with PBB-VPLS, with a few exceptions listed below:

- When a PW is setup between a PBB-EVPN PE and a VPLS PE, it gets setup between the I-component of PBB-EVPN PE and the bridge component of VPLS PE.
- The MAC mobility needs to be handled. The details of which will be covered in a future revision of this document.

5.3 Multicast Operation

5.3.1 Ingress Replication

The operation is per the procedures of Section 3.3.1 above for the scenario WITHOUT [MMRP]. The replication list is maintained per I-SID on the PBB-EVPN PEs and per VPN instance on the VPLS PEs.

5.3.2 LSM Will be covered in a future revision of this document.

6 Solution Advantages

The solution for seamless integration of (PBB-)EVPN with (PBB-)VPLS has the following advantages:

- When ingress replication is used for multi-destination traffic delivery, the solution reduces the scope of [MMRP] (which is a soft-

state protocol) to only that of existing VPLS PEs, and uses the more robust BGP-based mechanism for multicast pruning among new EVPN PEs.

- It is completely backward compatible.
- New PEs can leverage the extensive multi-homing mechanisms and provisioning simplifications of PBB-EVPN:
 1. Auto-sensing of MHN / MHD
 2. Auto-discovery of redundancy group
 3. Auto-provisioning of DF election and VLAN carving

7 Security Considerations

No new security considerations beyond those for VPLS and EVPN.

8 IANA Considerations

This document has no actions for IANA.

9 References

9.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.
- [RFC4447] Martini, et al., "Pseudowire Setup and Maintenance using the Label Distribution Protocol", draft-ietf-pwe3-rfc4447bis-02.txt, October 2013.
- [EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt, work in progress, July, 2013.
- [PBB-EVPN] Sajassi et al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-05.txt, work in progress, October, 2013.

9.2 Informative References

- [MMRP] Clause 10 of "IEEE Standard for Local and metropolitan area

networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q, 2013.

[PBB-VPLS-PE-MODEL] Balus, F., Sajassi, S., Bitar, N., "Extensions to VPLS PE model for Provider Backbone Bridging", RFC xxxx, June 2013.

[PBB-VPLS] Sajassi, et al., "VPLS Interoperability with Provider Backbone Bridges", draft-ietf-l2vpn-pbb-vpls-interop-05.txt, work in progress, October, 2013.

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

Nick Del Regno
Verizon
400 International Pkwy
Richardson, TX 75089, US
Email: nick.delregno@verizon.com

L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

A. Sajassi (Editor)
Cisco

Y. Rekhter
R. Shekhar
B. Schliesser
Juniper

J. Drake (Editor)
Juniper

Nabil Bitar
Verizon

S. Salam
K. Patel
D. Rao
S. Thoria
Cisco

Aldrin Isaac
Bloomberg

James Uttaro
AT&T

L. Yong
Huawei

W. Henderickx
Alcatel-Lucent

Expires: April 21, 2014

October 21, 2013

A Network Virtualization Overlay Solution using EVPN
draft-sd-l2vpn-evpn-overlay-02

Abstract

This document describes how EVPN can be used as an NVO solution and explores the various tunnel encapsulation options over IP and their impact on the EVPN control-plane and procedures. In particular, the following encapsulation options are analyzed: MPLS over GRE, VXLAN, and NVGRE.

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Terminology	5
2	EVPN Features	5
3	Encapsulation Options for EVPN Overlays	6
3.1	VXLAN/NVGRE Encapsulation	6
3.1.1	Virtual Identifiers Scope	7
3.1.1.1	Data Center Interconnect with Gateway	7
3.1.1.2	Data Center Interconnect without Gateway	8
3.1.2	Virtual Identifiers to EVI Mapping	8
3.1.2.1	Auto Derivation of RT & RD	9
3.1.3	Constructing EVPN BGP Routes	10
3.1.3.1	Constructing E-VPN MAC Address Advertisement Route	11
3.2	MPLS over GRE	11
4	EVPN with Multiple Data Plane Encapsulations	12
5	NVE Residing in Hypervisor	12
5.1	Impact on EVPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation	12
5.2	Impact on EVPN Procedures for VXLAN/NVGRE Encapsulation	13
6	NVE Residing in ToR Switch	14
6.1	EVPN Multi-Homing Features	14
6.1.1	Multi-homed Ethernet Segment Auto-Discovery	14

6.1.2 Fast Convergence and Mass Withdraw	14
6.1.3 Split-Horizon	15
6.1.4 Aliasing and Backup-Path	15
6.1.5 DF Election	16
6.2 Impact on EVPN BGP Routes & Attributes	16
6.3 Impact on EVPN Procedures	16
6.3.1 Split Horizon	17
6.3.2 Aliasing and Backup-Path	18
7 Support for Multicast	18
8 Support for NVEs with data plane MAC learning	19
8.1 Advertising NVE capabilities	20
8.2 Advertising flood lists for ingress replication	20
9 Inter-AS	21
10 Acknowledgement	22
11 Security Considerations	22
12 IANA Considerations	22
13 References	22
11.1 Normative References	22
11.2 Informative References	23
Authors' Addresses	23

1 Introduction

In the context of this document, a Network Virtualization Overlay (NVO) is a solution to address the requirements of a multi-tenant data center, especially one with virtualized hosts, e.g., Virtual Machines (VMs). The key requirements of such a solution, as described in [Problem-Statement], are:

- Isolation of network traffic per tenant
- Support for a large number of tenants (tens or hundreds of thousands)
- Extending L2 connectivity among different VMs belonging to a given tenant segment (subnet) across different PODs within a data center or between different data centers
- Allowing a given VM to move between different physical points of attachment within a given L2 segment

The underlay network for NVO solutions is assumed to provide IP connectivity between NVO endpoints (NVEs).

This document describes how EVPN can be used as an NVO solution and explores applicability of EVPN functions and procedures. In particular, it describes the various tunnel encapsulation options for EVPN over IP, and their impact on the EVPN control-plane and procedures for two main scenarios:

- a) when the NVE resides in the hypervisor, and
- b) when the NVE resides in a ToR device

Note that the use of EVPN as an NVO solution does not necessarily mandate that the BGP control-plane be running on the NVE. For such scenarios, it is still possible to leverage the EVPN solution by using XMPP, or alternative mechanisms, to extend the control-plane to the NVE as discussed in [L3VPN-ENDSYSTEMS].

The possible encapsulation options for EVPN overlays that are analyzed in this document are:

- VXLAN and NVGRE
- MPLS over GRE

Before getting into the description of the different encapsulation options for EVPN over IP, it is important to highlight the EVPN solution's main features, how those features are currently supported,

and any impact that the encapsulation has on those features.

1.1 Terminology

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

NVE: Network Virtualization Endpoint

Virtual Identifier: refers to a VXLAN VNI or NVGRE VSID

2 EVPN Features

EVPN was originally designed to support the requirements detailed in [EVPN-REQ] and therefore has the following attributes which directly address control plane scaling and ease of deployment issues.

- 1) Control plane traffic is distributed with BGP and Broadcast and Multicast traffic is sent using a shared multicast tree or with ingress replication.
- 2) Control plane learning is used for MAC (and IP) addresses instead of data plane learning. The latter requires the flooding of unknown unicast and ARP frames; whereas, the former does not require any flooding.
- 3) Route Reflector is used to reduce a full mesh of BGP sessions among PE devices to a single BGP session between a PE and the RR. Furthermore, RR hierarchy can be leveraged to scale the number BGP routes on the RR.
- 4) Auto-discovery via BGP is used to discover PE devices participating in a given VPN, PE devices participating in a given redundancy group, tunnel encapsulation types, multicast tunnel type, multicast members, etc.
- 5) All-Active multi-homing is used. This allows a given customer device (CE) to have multiple links to multiple PEs, and traffic to/from that CE fully utilizes all of these links. This set of links is termed an Ethernet Segment (ES).
- 6) Mass withdraw is used. When a link between a CE and a PE fails, the PEs in all EVPNs associated with that failed link are notified via the withdrawal of a single EVPN route regardless of how many MAC addresses are located at the CE.
- 7) Route filtering and constrained route distribution are used to

ensure that the control plane traffic for a given EVPN is only distributed to the PEs in that EVPN.

8) The internal identifier of a broadcast domain, the Ethernet Tag, is a 32 bit number, which is mapped into whatever broadcast domain identifier, e.g., VLAN ID, is understood by the attaching CE device. This means that when 802.1q interfaces are used, there are up to 4096 distinct VLAN IDs for each attaching CE device in a given EVPN.

9) VM Mobility mechanisms ensure that all PEs in a given EVPN know the ES with which a given VM, as identified by its MAC and IP addresses, is currently associated.

10) Route Targets are used to allow the operator (or customer) to define a spectrum of logical network topologies including mesh, hub & spoke, and extranets (e.g., a VPN whose sites are owned by different enterprises), without the need for proprietary software or the aid of other virtual or physical devices.

11) Because the design goal for NVO is millions of instances per common physical infrastructure, the scaling properties of the control plane for NVO are extremely important. EVPN and the extensions described herein, are designed with this level of scalability in mind.

3 Encapsulation Options for EVPN Overlays

3.1 VXLAN/NVGRE Encapsulation

Both VXLAN and NVGRE are examples of technologies that provide a data plane encapsulation which is used to transport a packet over the common physical infrastructure between NVEs, VXLAN Tunnel End Point (VTEPs) in VXLAN and Network Virtualization Endpoint (NVEs) in NVGRE. Both of these technologies include the identifier of the specific NVO instance, Virtual Network Identifier (VNI) in VXLAN and Virtual Subnet Identifier (VSID), NVGRE, in each packet.

Note that a Provider Edge (PE) is equivalent to a VTEP/NVE.

[VXLAN] encapsulation is based on UDP, with an 8-byte header following the UDP header. VXLAN provides a 24-bit VNI, which typically provides a one-to-one mapping to the tenant VLAN ID, as described in [VXLAN]. In this scenario, the VTEP does not include an inner VLAN tag on frame encapsulation, and discards decapsulated frames with an inner VLAN tag. This mode of operation in [VXLAN] maps to VLAN Based Service in [EVPN], where a tenant VLAN ID gets mapped to an EVPN instance (EVI).

[VXLAN] also provides an option of including an inner VLAN tag in the encapsulated frame, if explicitly configured at the VTEP. This mode of operation maps to VLAN Bundle Service in [EVPN], where the VLANs of a given tenant get mapped to an EVI.

[NVGRE] encapsulation is based on [GRE] and it mandates the inclusion of the optional GRE Key field which carries the VSID. There is a one-to-one mapping between the VSID and the tenant VLAN ID, as described in [NVGRE] and the inclusion of an inner VLAN tag is prohibited. This mode of operation in [NVGRE] maps to VLAN Based Service in [EVPN]. In other words, [NVGRE] prohibits the application of VLAN Bundle Service in [EVPN] and it only requires VLAN Based Service in [EVPN].

As described in the next section there is no change to the encoding of EVPN routes to support VXLAN or NVGRE encapsulation except for the use of BGP Encapsulation extended community. However, there is potential impact to the EVPN procedures depending on where the NVE is located (i.e., in hypervisor or TOR) and whether multi-homing capabilities are required.

3.1.1 Virtual Identifiers Scope

Although VNI or VSID are defined as 24-bit globally unique values, there are scenarios in which it is desirable to use a locally significant value for VNI or VSID, especially in the context of data center interconnect:

3.1.1.1 Data Center Interconnect with Gateway

In the case where NVEs in different data centers need to be interconnected, and a Gateway is employed at the edge of the data center network, the NVEs should treat the VNI or VSID as a globally unique identifier within a data center. This is because the Gateway will provide the functionality of translating the VNI or VSID when crossing network boundaries, which may align with operator span of control boundaries. As an example, consider the network of Figure 1 below. Assume there are three network operators: one for each of the DC1, DC2 and WAN networks. The Gateways at the edge of the data centers are responsible for translating the VNIs / VSIDs between the values used in each of the data center networks and the values used in the WAN.

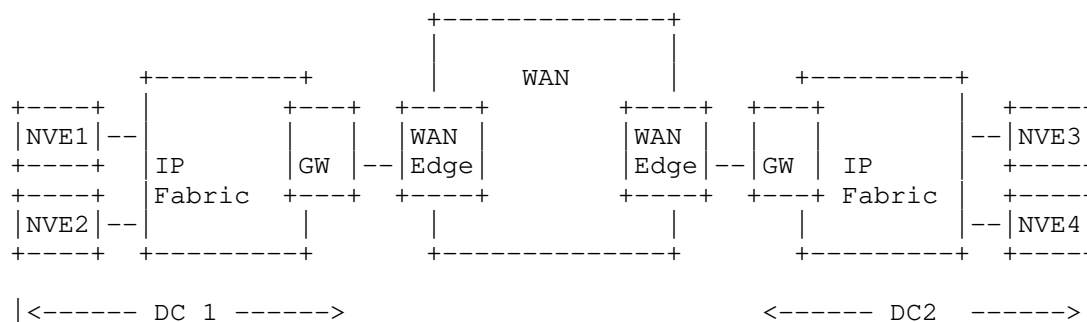


Figure 1: Data Center Interconnect with Gateway

3.1.1.2 Data Center Interconnect without Gateway

In the case where NVEs in different data centers need to be interconnected, and Gateways are not employed at the edge of the data center network, it is useful to treat the VNIs or VSIDs as locally significant identifiers (e.g., as an MPLS label). More specifically, the VNI or VSID value that is used by the transmitting NVE is allocated by the NVE that is receiving the traffic (in other words, this is a "downstream assigned" MPLS label). This allows the VNI or VSID space to be decoupled between different data center networks without the need for a dedicated Gateway at the edge of the data centers.

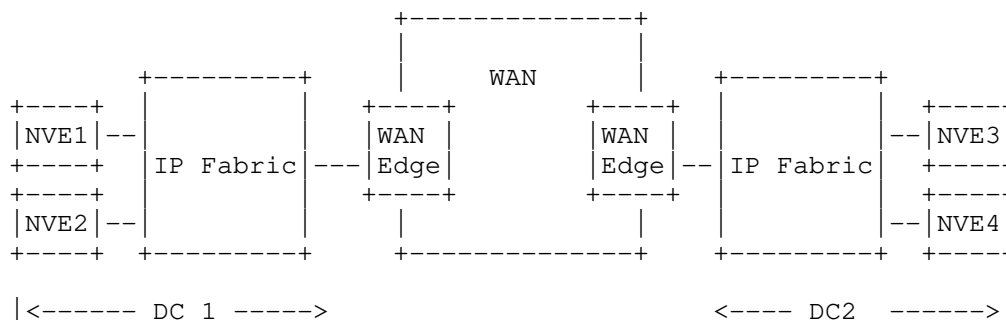


Figure 2: Data Center Interconnect without Gateway

3.1.2 Virtual Identifiers to EVI Mapping

When the EVPN control plane is used in conjunction with VXLAN or NVGRE, two options for mapping the VXLAN VNI or NVGRE VSID to an EVPN Instance (EVI) are possible:

1. Option 1: Single Virtual Identifier per EVI

In this option, every VNI or VSID is mapped to a unique EVI. As such, a BGP RD and RT is needed per VNI / VSID on every VTEP. The advantage of this model is that it allows the BGP RT constraint mechanisms to be used in order to limit the propagation and import of routes to only the VTEPs that are interested in a given VNI (or VSID). The disadvantage of this model may be the provisioning overhead if RD and RT are not derived automatically from VNI (for VSID).

In this option, the MAC-VRF table is identified by the RT in the control plane (because Ethernet Tag field in the MAC route is set to zero) and by the VNI (or VSID) in the data-plane.

2. Option 2: Multiple Virtual Identifiers per EVI

In this option, multiple VNIs or VSIDs are mapped to a unique EVI. For example, if a tenant has multiple segments/subnets each represented by a VNI (or VSID), then all the VNIs (or VSIDs) for that tenant are mapped to a single EVI - e.g., the EVI in this case represents the tenant and not a subnet. The advantage of this model is that it doesn't require the provisioning of RD/RT per VNI or VSID which is a moot point if auto-derivation is used. The disadvantage of this model is that routes would be imported by VTEPs that may not be interested in a given VNI (or VSID).

In this option the MAC-VRF table is identified by the VNI (or VSID) in both the control plane and the data-plane.

3.1.2.1 Auto Derivation of RT & RD

When the option of a single VNI (or VSID) per EVI is used, it is important to auto-derive RD and RT for EVPN BGP routes in order to simplify configuration for data center operations. RD can be auto-derive as described in [EVPN] and RT can be auto-derived as described next.

Since a gateway PE as depicted in figure-1 participates in both the DCN and WAN BGP sessions, it is important that when RT values are auto-derived for VNIs (or VSIDs), there is no conflict in RT spaces between DCN and WAN networks assuming that both are operating within the same AS. Also, there can be scenarios where both VXLAN and NVGRE encapsulations may be needed within the same DCN and their corresponding VNIs and VSIDs are administered independently which means VNI and VSID spaces can overlap. In order to ensure that no such conflict in RT spaces arises, RT values for DCNs are auto-derived as follow:

- 2 bytes of global admin field of the RT is set to the AS number.
- Three least significant bytes of the local admin field of the RT is set to the VNI or VSID, I-SID, or VID. The most significant bit of the local admin field of the RT is set as follow:
 - 0: auto-derived
 - 1: manually-derived
- The remaining 7 bits of the most significant byte of the local admin field of the RT identifies the space in which the other 3 bytes are defined. The following spaces are defined:
 - 0 : EVI
 - 1 : VXLAN
 - 2 : NVGRE
 - 3 : I-SID
 - 4 : VID

3.1.3 Constructing EVPN BGP Routes

In EVPN, an MPLS label distributed by the egress PE via the EVPN control plane and placed in the MPLS header of a given packet by the ingress PE. This label is used upon receipt of that packet by the egress PE to disposition that packet. This is very similar to the use of the VNI or VSID by the egress VTEP or NVE, respectively, with the difference being that an MPLS label has local significance and is distributed by the EVPN control plane, while a VNI or VSID typically has global significance.

As discussed in Section 3.1.1 above, there are scenarios in which it is desirable to use a locally significant value for VNI or VSID and in such such scenarios, MPLS label is advertised in EVPN BGP routes and it is used in VXLAN or NVGRE encapsulation as a 20-bit value for VNI or VSID.

This memo specifies that when EVPN is used with a VXLAN or NVGRE data plane and when a globally significant VNI or VSID is desirable, then for VNI-based mode (single VNI per EVI), the Ethernet Tag field of EVPN BGP routes (which is a 4-octet field) MUST be set to zero just like the VLAN-based mode in baseline EVPN. If VNI-aware bundle mode (multiple VNIs per EVI) is desired, then the Ethernet Tag field of EVPN BGP routes MUST be set to VNI (or VSID) accordingly just like VLAN-aware bundle mode in baseline EVPN. In both cases, the MPLS label field of the EVPN BGP routes MUST be set to zero.

This memo also specifies that when EVPN is used with a VXLAN or NVGRE data plane and when a locally significant VNI or VSID is desirable, then MPLS field of EVPN BGP routes (which is a 3-octet field) MUST be

used and Ethernet Tag field MUST be set to zero. In such scenarios, only VNI-based mode (single VNI per EVI) is supported.

In order to indicate that a VXLAN or NVGRE data plane encapsulation rather than MPLS label stack encapsulation is to be used, the BGP Encapsulation extended community defined in [RFC5512] is included with EVPN MAC route, Inclusive Multicast route, or per EVI Ethernet AD route advertised by an egress PE. Two new values, one for VXLAN and one for NVGRE, will be defined to extend the list of encapsulation types defined in [RFC5512]:

- + 3 - VXLAN Encapsulation
- + 4 - NVGRE Encapsulation

If BGP Encapsulation extended community is not present, then the default encapsulation MPLS encapsulation (or statically configured encapsulation) is used.

3.1.3.1 Constructing E-VPN MAC Address Advertisement Route

In EVPN, unicast MAC addresses are advertised via MAC Advertisement route. The Ethernet Tag field in this route is set zero for the VNI-based mode and set to VNI (or VSID) for VNI-aware bundle mode. The MPLS label field is set to zero. The encapsulation is set via the BGP Encapsulation extended community as described in section 3.1.3.

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the NVE. The remaining fields in the route are set as per EVPN.

3.2 MPLS over GRE

The EVPN data-plane is modeled as an EVPN MPLS client layer sitting over an MPLS PSN tunnel. Some of the EVPN functions (split-horizon, aliasing and repair-path) are tied to the MPLS client layer. If MPLS over GRE encapsulation is used, then the EVPN MPLS client layer can be carried over an IP PSN tunnel transparently. Therefore, there is no impact to the EVPN procedures and associated data-plane operation.

The existing standards for MPLS over GRE encapsulation as defined by [RFC4023] can be used for this purpose; however, when it is used in conjunction with EVPN the key field MUST be present, and SHOULD be used to provide a 32-bit entropy field. The Checksum and Sequence Number fields are not needed and their corresponding C and S bits MUST be set to zero.

4 EVPN with Multiple Data Plane Encapsulations

The use of the BGP Encapsulation extended community allows each PE in a given EVPN to know whether the other PEs in that EVPN support MPLS label stack, VXLAN, and/or NVGRE data plane encapsulations. I.e., PEs in a given EVPN may support multiple data plane encapsulations.

If BGP Encapsulation extended community is not present, then the default MPLS encapsulation (or statically configured encapsulation) is used. However, if this attribute is present, then an ingress PE can send a frame to an egress PE only if the set of encapsulations advertised by the egress PE in the subject MAC Advertisement or Per EVI Ethernet AD route, forms a non-empty intersection with the set of encapsulations supported by the ingress PE, and it is at the discretion of the ingress PE which encapsulation to choose from this intersection.

An ingress node that uses shared multicast trees for sending broadcast or multicast frames MUST maintain distinct trees for each different encapsulation type.

It is the responsibility of the operator of a given EVPN to ensure that all of the PEs in that EVPN support at least one common encapsulation. If this condition is violated, it could result in service disruption or failure. The use of the BGP Encapsulation extended community provides a method to detect when this condition is violated but the actions to be taken are at the discretion of the operator and are outside the scope of this document.

5 NVE Residing in Hypervisor

When a PE and its CEs are co-located in the same physical device, e.g., when the PE resides in a server and the CEs are its VMs, the links between them are virtual and they typically share fate; i.e., the subject CEs are typically not multi-homed or if they are multi-homed, the multi-homing is a purely local matter to the server hosting the VM, and need not be "visible" to any other PEs, and thus does not require any specific protocol mechanisms. The most common case of this is when the NVE resides in the hypervisor.

In the sub-sections that follow, we will discuss the impact on EVPN procedures for the case when the NVE resides on the hypervisor and the VXLAN or NVGRE encapsulation is used.

5.1 Impact on EVPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation

As discussed above, both [NVGRE] and [VXLAN] do not require the tenant VLAN tag to be sent in BGP routes. Therefore, the 4-octet

Ethernet Tag field in the EVPN BGP routes can be used to represent the globally significant value for VXLAN VNI or NVGRE VSID and MPLS field can be used to represent the locally significant value for VNI or VSID.

When the VXLAN VNI or NVGRE VSID is assumed to be a global value, one might question the need for the Route Distinguisher (RD) in the EVPN routes. In the scenario where all data centers are under a single administrative domain, and there is a single global VNI/VSID space, the RD MAY be set to zero in the EVPN routes. However, in the scenario where different groups of data centers are under different administrative domains, and these data centers are connected via one or more backbone core providers as described in [NOV3-Framework], the RD must be a unique value per EVI or per NVE as described in [EVPN]. In other words, whenever there is more than one administrative domain for global VNI or VSID, then a non-zero RD MUST be used, or whenever the VNI or VSID value have local significance, then a non-zero RD MUST be used. It is recommend to use a non-zero RD at all time.

When the NVEs reside on the hypervisor, the EVPN BGP routes and attributes associated with multi-homing are no longer required. This reduces the required routes and attributes to the following subset of five out of the set of eight :

- MAC Advertisement Route
- Inclusive Multicast Ethernet Tag Route
- MAC Mobility Extended Community
- Default Gateway Extended Community

As mentioned in section 3.1.1, BGP Encapsulation extended community as defined in [RFC5512] SHOULD be used along with MAC Advertisement Route or Ethernet AD Route to indicate the supported encapsulations.

5.2 Impact on EVPN Procedures for VXLAN/NVGRE Encapsulation

When the NVEs reside on the hypervisors, the EVPN procedures associated with multi-homing are no longer required. This limits the procedures on the NVE to the following subset of the EVPN procedures:

1. Local learning of MAC addresses received from the VMs per section 10.1 of [EVPN].
2. Advertising locally learned MAC addresses in BGP using the MAC Advertisement routes.
3. Performing remote learning using BGP per Section 10.2 of [EVPN].

4. Discovering other NVEs and constructing the multicast tunnels using the Inclusive Multicast Ethernet Tag routes.
5. Handling MAC address mobility events per the procedures of Section 16 in [EVPN].

6 NVE Residing in ToR Switch

In this section, we discuss the scenario where the NVEs reside in the Top of Rack (ToR) switches AND the servers (where VMs are residing) are multi-homed to these ToR switches. The multi-homing may operate in All-Active or Active/Standby redundancy mode. If the servers are single-homed to the ToR switches, then the scenario becomes similar to that where the NVE resides in the hypervisor, as discussed in Section 5, as far as the required EVPN functionality.

[EVPN] defines a set of BGP routes, attributes and procedures to support multi-homing. We first describe these functions and procedures, then discuss which of these are impacted by the encapsulation (such as VXLAN or NVGRE) and what modifications are required.

6.1 EVPN Multi-Homing Features

In this section, we will recap the multi-homing features of EVPN to highlight the encapsulation dependencies. The section only describes the features and functions at a high-level. For more details, the reader is to refer to [EVPN].

6.1.1 Multi-homed Ethernet Segment Auto-Discovery

EVPN NVEs (or PEs) connected to the same Ethernet Segment (e.g. the same server via LAG) can automatically discover each other with minimal to no configuration through the exchange of BGP routes.

6.1.2 Fast Convergence and Mass Withdraw

EVPN defines a mechanism to efficiently and quickly signal, to remote NVEs, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet segment (e.g., a link or a port failure). This is done by having each NVE advertise an Ethernet A-D Route per Ethernet segment for each locally attached segment. Upon a failure in connectivity to the attached segment, the NVE withdraws the corresponding Ethernet A-D route. This triggers all NVEs that receive the withdrawal to update their next-hop adjacencies for all MAC addresses associated with the Ethernet segment in question. If no other NVE had advertised an Ethernet A-D route for

the same segment, then the NVE that received the withdrawal simply invalidates the MAC entries for that segment. Otherwise, the NVE updates the next-hop adjacencies to point to the backup NVE(s).

6.1.3 Split-Horizon

Consider a station that is multi-homed to two or more NVEs on an Ethernet segment ESI, with all-active redundancy. If the station sends a multicast, broadcast or unknown unicast packet to a particular NVE, say NE1, then NE1 will forward that packet to all or subset of the other NVEs in the EVPN instance. In this case the NVEs, other than NE1, that the station is multi-homed to MUST drop the packet and not forward back to the station. This is referred to as "split horizon" filtering.

6.1.4 Aliasing and Backup-Path

In the case where a station is multi-homed to multiple NVEs, it is possible that only a single NVE learns a set of the MAC addresses associated with traffic transmitted by the station. This leads to a situation where remote NVEs receive MAC advertisement routes, for these addresses, from a single NVE even though multiple NVEs are connected to the multi-homed station. As a result, the remote NVEs are not able to effectively load-balance traffic among the NVEs connected to the multi-homed Ethernet segment. This could be the case, for e.g. when the NVEs perform data-path learning on the access, and the load-balancing function on the station hashes traffic from a given source MAC address to a single NVE. Another scenario where this occurs is when the NVEs rely on control plane learning on the access (e.g. using ARP), since ARP traffic will be hashed to a single link in the LAG.

To alleviate this issue, EVPN introduces the concept of Aliasing. This refers to the ability of an NVE to signal that it has reachability to a given locally attached Ethernet segment, even when it has learnt no MAC addresses from that segment. The Ethernet A-D route per EVI is used to that end. Remote NVEs which receive MAC advertisement routes with non-zero ESI SHOULD consider the MAC address as reachable via all NVEs that advertise reachability to the relevant Segment using Ethernet A-D routes with the same ESI and with the Active-Standby flag reset.

Backup-Path is a closely related function, albeit it applies to the case where the redundancy mode is Active/Standby. In this case, the NVE signals that it has reachability to a given locally attached Ethernet Segment using the Ethernet A-D route as well. Remote NVEs which receive the MAC advertisement routes, with non-zero ESI, SHOULD consider the MAC address as reachable via the advertising NVE.

Furthermore, the remote NVEs SHOULD install a Backup-Path, for said MAC, to the NVE which had advertised reachability to the relevant Segment using an Ethernet A-D route with the same ESI and with the Active-Standby flag set.

6.1.5 DF Election

Consider a station that is a host or a VM that is multi-homed directly to more than one NVE in an EVPN on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

- Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular Ethernet segment, to the station.
- Flooding unknown unicast traffic (i.e. traffic for which an NVE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the station, if the environment requires flooding of unknown unicast traffic.

This is required in order to prevent duplicate delivery of multi-destination frames to a multi-homed host or VM, in case of all-active redundancy.

6.2 Impact on EVPN BGP Routes & Attributes

Since multi-homing is supported in this scenario, then the entire set of BGP routes and attributes defined in [EVPN] are used. As discussed in Section 3.1, the VSID or VNI is encoded in the Ethernet Tag field of the routes if globally significant or in the MPLS label field if locally significant.

As mentioned in section 3.1.1, BGP Encapsulation extended community as defined in [RFC5512] SHOULD be used along with MAC Advertisement Route or Ethernet AD Route to indicate the supported encapsulations.

6.3 Impact on EVPN Procedures

Two cases need to be examined here, depending on whether the NVEs are operating in Active/Standby or in All-Active redundancy.

First, let's consider the case of Active/Standby redundancy, where the hosts are multi-homed to a set of NVEs, however, only a single NVE is active at a given point of time for a given VNI or VSID. In this case, the Split-Horizon and Aliasing functions are not required

but other functions such as multi-homed Ethernet segment auto-discovery, fast convergence and mass withdraw, repair path, and DF election are required. In this case, the impact of the use of the VXLAN/NVGRE encapsulation on the EVPN procedures is when the Backup-Path function is supported, as discussed next:

In EVPN, the NVEs connected to a multi-homed site using Active/Standby redundancy optionally advertise a VPN label, in the Ethernet A-D Route per EVI, used to send traffic to the backup NVE in the case where the primary NVE fails. In the case where VXLAN or NVGRE encapsulation is used, some alternative means that does not rely on MPLS labels is required to support Backup-Path. This is discussed in Section 4.3.2 below. If the Backup-Path function is not used, then the VXLAN/NVGRE encapsulation would have no impact on the EVPN procedures.

Second, let's consider the case of All-Active redundancy. In this case, out of the EVPN multi-homing features listed in section 4.1, the use of the VXLAN or NVGRE encapsulation impacts the Split-Horizon and Aliasing features, since those two rely on the MPLS client layer. Given that this MPLS client layer is absent with these types of encapsulations, alternative procedures and mechanisms are needed to provide the required functions. Those are discussed in detail next.

6.3.1 Split Horizon

In EVPN, an MPLS label is used for split-horizon filtering to support active/active multi-homing where an ingress ToR switch adds a label corresponding to the site of origin (aka ESI MPLS Label) when encapsulating the packet. The egress ToR switch checks the ESI MPLS label when attempting to forward a multi-destination frame out an interface, and if the label corresponds to the same site identifier (ESI) associated with that interface, the packet gets dropped. This prevents the occurrence of forwarding loops.

Since the VXLAN or NVGRE encapsulation does not include this ESI MPLS label, other means of performing the split-horizon filtering function MUST be devised. The following approach is recommended for split-horizon filtering when VXLAN or NVGRE encapsulation is used.

Every NVE track the IP address(es) associated with the other NVE(s) with which it has shared multi-homed Ethernet Segments. When the NVE receives a multi-destination frame from the overlay network, it examines the source IP address in the tunnel header (which corresponds to the ingress NVE) and filters out the frame on all local interfaces connected to Ethernet Segments that are shared with the ingress NVE. With this approach, it is required that the ingress NVE performs replication locally to all directly attached Ethernet

Segments (regardless of the DF Election state) for all flooded traffic ingress from the access interfaces (i.e. from the hosts). This approach is referred to as "Local Bias", and has the advantage that only a single IP address needs to be used per NVE for split-horizon filtering, as opposed to requiring an IP address per Ethernet Segment per NVE.

In order to prevent unhealthy interactions between the split horizon procedures defined in [EVPN] and the local bias procedures described in this memo, a mix of MPLS over GRE encapsulations on the one hand and VXLAN/NVGRE encapsulations on the other on a given Ethernet Segment is prohibited. The use of the BGP Encapsulation extended community provides a method to detect when this condition is violated but the actions to be taken are at the discretion of the operator and are outside the scope of this document.

6.3.2 Aliasing and Backup-Path

The Aliasing and the Backup-Path procedures for VXLAN/NVGRE encapsulation is very similar to the ones for MPLS. In case of MPLS, two different Ethernet AD routes are used for this purpose. The one used for Aliasing has a VPN scope and carries a VPN label but the one used for Backup-Path has Ethernet segment scope and doesn't carry any VPN specific info (e.g., Ethernet Tag and MPLS label are set to zero). The same two routes are used when VXLAN or NVGRE encapsulation is used with the difference that when Ethernet AD route is used for Aliasing with VPN scope, the Ethernet Tag field is set to VNI or VSID to indicate VPN scope (and MPLS field may be set to a VPN label if needed).

7 Support for Multicast

The E-VPN Inclusive Multicast BGP route is used to discover the multicast tunnels among the endpoints associated with a given VXLAN VNI or NVGRE VSID. The Ethernet Tag field of this route is used to encode the VNI for VLXAN or VSID for NVGRE. The Originating router's IP address field is set to the NVE's IP address. This route is tagged with the PMSI Tunnel attribute, which is used to encode the type of multicast tunnel to be used as well as the multicast tunnel identifier. The tunnel encapsulation is encoded by adding the BGP Encapsulation extended community as per section 3.1.1. The following tunnel types as defined in [RFC6514] can be used in the PMSI tunnel attribute for VXLAN/NVGRE:

- + 3 - PIM-SSM Tree
- + 4 - PIM-SM Tree
- + 5 - BIDIR-PIM Tree
- + 6 - Ingress Replication

Except for Ingress Replication, this multicast tunnel is used by the PE originating the route for sending multicast traffic to other PEs, and is used by PEs that receive this route for receiving the traffic originated by CEs connected to the PE that originated the route.

In the scenario where the multicast tunnel is a tree, both the Inclusive as well as the Aggregate Inclusive variants may be used. In the former case, a multicast tree is dedicated to a VNI or VSID. Whereas, in the latter, a multicast tree is shared among multiple VNIs or VSIDs. This is done by having the NVEs advertise multiple Inclusive Multicast routes with different VNI or VSID encoded in the Ethernet Tag field, but with the same tunnel identifier encoded in the PMSI Tunnel attribute.

8 Support for NVEs with data plane MAC learning

In an overlay network it possible to have a mix of NVEs, such that only a subset of the NVEs are capable of participating in control plane MAC learning via EVPN. The other subset of NVEs would perform conversational MAC learning in data plane. It must be possible for NVEs with this mixed capability to still be part of the same overlay network.

If the administrative policy of an EVPN NVE requires for flooding of unknown unicast, then the following procedures are not needed; however, if the administrative policy of the EVPN NVE requires no flooding of unknown unicast, then for such a mixed overlay network to operate correctly, the following requirements MUST be met:

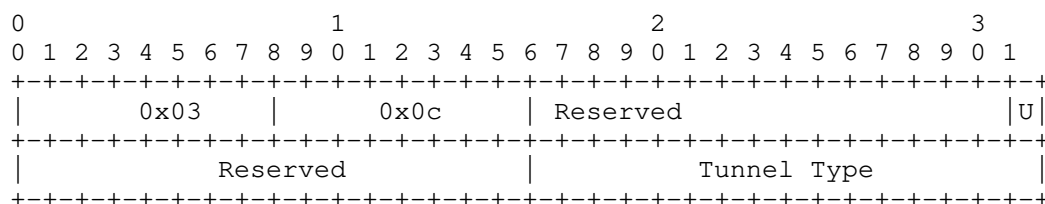
- When an NVE capable of doing control-plane MAC learning via EVPN wants to send an unknown unicast frame, it MUST send it to a subset of NVEs in the VNI that only have data plane MAC learning capability. This can be achieved by creating a flood list for each VNI to carry unknown unicast traffic, which only the subset of NVEs with data plane MAC learning are part of. Section 8.2 describes the procedure to accomplish this.

- Broadcast traffic MUST be sent to all NVEs in the VNI regardless of the MAC learning capability. A separate flood list for each VNI to carry broadcast traffic can be created for this, and all NVEs in the VNI would be part of this flood list. Section 8.2 describes the procedure to accomplish this.

- When an NVE capable of only data plane MAC learning wants to send an unknown unicast frame, it MUST send it to all NVEs in the VNI. This can be achieved by flooding the unknown unicast frame in the broadcast flood list (as described earlier).

8.1 Advertising NVE capabilities

BGP Encapsulation extended community is used to signal NVE capabilities. NVE capabilities are used to build different types of flood lists in the broadcast domain for optimal forwarding in the case of NVEs with mixed capabilities, as described in section 8. The reserved field of the BGP Encapsulation extended community is repurposed to indicate NVE capabilities as following:



U bit indicates that the NVE must be included in unknown unicast flood list

The Reserved fields must be set to zero and ignored on receipt.

8.2 Advertising flood lists for ingress replication

Flooding of unknown unicast, broadcast and multicast can either be achieved by using multicast trees in the underlay or using ingress replication. If IP multicast is used for flooding, separate flood lists, as described in section 8, can be created by using separate IP multicast groups for different flood lists. If ingress replication is used for flooding, then the EVPN capable NVEs must maintain separate flood lists depending on advertised NVE capability. Either way, there is a need to signal which NVEs are part of which flood lists. This section describes enhancements to BGP signaling required to achieve this.

The EVPN Inclusive Multicast Route along with NVE capabilities as described in section 8.1 can be used to build different flood lists. The Inclusive Multicast Route is encoded as follows: The Ethernet Tag field is set to the VNI for VXLAN and VSID for NVGRE. The Originator's IP address field is set to the NVE's IP address. The Next Hop field of the MP_REACH_NLRI attribute of the route is set to NVE's IP address. The Inclusive Multicast route is tagged with the PMSI tunnel attribute. The BGP Encapsulation extended community is included with U, B or K bit set as described in section 8.1 to enable an NVE to be part of a specific flood list depending on its

capabilities.

9 Inter-AS

For inter-AS operation, two scenarios must be considered:

- Scenario 1: The tunnel endpoint IP addresses are public
- Scenario 2: The tunnel endpoint IP addresses are private

In the first scenario, inter-AS operation is straight-forward and follows existing BGP inter-AS procedures. However, in the first scenario where the tunnel endpoint IP addresses are public, there may be security concern regarding the distribution of these addresses among different ASes. This security concern is one of the main reasons for having the so called inter-AS "option-B" in MPLS VPN solutions such as EVPN.

The second scenario is more challenging, because the absence of the MPLS client layer from the VXLAN encapsulation creates a situation where the ASBR has no fully qualified indication within the tunnel header as to where the tunnel endpoint resides. To elaborate on this, recall that with MPLS, the client layer labels (i.e. the VPN labels) are downstream assigned. As such, this label implicitly has a connotation of the tunnel endpoint, and it is sufficient for the ASBR to look up the client layer label in order to identify the label translation required as well as the tunnel endpoint to which a given packet is being destined. With the VXLAN encapsulation, the VNI is globally assigned and hence is shared among all endpoints. The destination IP address is the only field which identifies the tunnel endpoint in the tunnel header, and this address is privately managed by every data center network. Since the tunnel address is allocated out of a private address pool, then we either need to do a lookup based on VTEP IP address in context of a VRF (e.g., use IP-VPN) or terminate the VXLAN tunnel and do a lookup based on the tenant's MAC address to identify the egress tunnel on the ASBR. This effectively mandates that the ASBR to either run another overlay solution such as IP-VPN over MPLS/IP core network or to be aware of the MAC addresses of all VMs in its local AS, at the very least.

If VNIs/VSIDs have local significance, then the inter-AS operation can be simplified to that of MPLS and thus MPLS inter-AS option B and C can be leveraged in here. That's why the use of local significance VNIs/VSIDs (e.g., MPLS labels) are recommended for inter-AS operation of DC networks without gateways.

10 Acknowledgement

The authors would like to thank David Smith, John Mullooly, Thomas Nadeau for their valuable comments and feedback.

11 Security Considerations

This document uses IP-based tunnel technologies to support data plane transport. Consequently, the security considerations of those tunnel technologies apply. This document defines support for [VXLAN] and [NVGRE]. The security considerations from those documents as well as [RFC4301] apply to the data plane aspects of this document.

As with [RFC5512], any modification of the information that is used to form encapsulation headers, to choose a tunnel type, or to choose a particular tunnel for a particular payload type may lead to user data packets getting misrouted, misdelivered, and/or dropped.

More broadly, the security considerations for the transport of IP reachability information using BGP are discussed in [RFC4271] and [RFC4272], and are equally applicable for the extensions described in this document.

If the integrity of the BGP session is not itself protected, then an imposter could mount a denial-of-service attack by establishing numerous BGP sessions and forcing an IPsec SA to be created for each one. However, as such an imposter could wreak havoc on the entire routing system, this particular sort of attack is probably not of any special importance.

It should be noted that a BGP session may itself be transported over an IPsec tunnel. Such IPsec tunnels can provide additional security to a BGP session. The management of such IPsec tunnels is outside the scope of this document.

12 IANA Considerations

13 References

13.1 Normative References

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

[RFC4271] Y. Rekhter, Ed., T. Li, Ed., S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", January 2006.

- [RFC4272] S. Murphy, "BGP Security Vulnerabilities Analysis.", January 2006.
- [RFC4301] S. Kent, K. Seo., "Security Architecture for the Internet Protocol.", December 2005.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

11.2 Informative References

- [EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (EVPN)", draft-ietf-l2vpn-evpn-req-01.txt, work in progress, October 21, 2012.
- [NVGRE] Sridhavan, M., et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-01.txt, July 8, 2012.
- [VXLAN] Dutt, D., et al, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-02.txt, August 22, 2012.
- [EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-02.txt, work in progress, February, 2012.
- [Problem-Statement] Narten et al., "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement-01, September 2012.
- [L3VPN-ENDSYSTEMS] Marques et al., "BGP-signaled End-system IP/VPNs", draft-ietf-l3vpn-end-system, work in progress, October 2012.
- [NOV3-FRWK] Lasserre et al., "Framework for DC Network Virtualization", draft-ietf-nvo3-framework-01.txt, work in progress, October 2012.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

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

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

James Uttaro
AT&T
Email: uttaro@att.com

Wim Henderickx
Alcatel-Lucent
e-mail: wim.henderickx@alcatel-lucent.com

Ravi Shekhar
Juniper Networks
Email: rshekhar@juniper.net

Samer Salam
Cisco
Email: ssalam@cisco.com

Keyur Patel
Cisco
Email: Keyupate@cisco.com

Dhananjaya Rao
Cisco
Email: dhrao@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com

L2VPN Working Group
INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: April 24, 2014

R. Singh
K. Kompella
Juniper Networks
October 21, 2013

Updated processing of control flags for BGP VPLS
draft-singh-l2vpn-bgp-vpls-control-flags-00

Abstract

This document updates the meaning of the "control flags" fields inside the "layer2 info extended community" used for BGP-VPLS NLRI.

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	3
2	Problem	3
3	Updated meaning of control flags in the layer2 info extended community	4
4	Using p2mp LSP as transport for CW-marked VPLS frames	4
5	Illustrative diagram	5
6	Security Considerations	5
7	IANA Considerations	6
8	References	6
8.1	Normative References	6
	Authors' Addresses	6

1 Introduction

The use of control word (CW) helps prevent mis-ordering of IPv4 or IPv6 PW traffic over ECMP-paths/LAG-bundles. [RFC4385] describes the format for control-word that may be used over point-2-point PWs and over a VPLS.

[RFC4761] describes the concepts and signaling for using BGP to bring up a VPLS. It specifies as part of its BGP VPLS NLRI that a PE may require other PEs in the same VPLS to include (or not) control-word and sequencing information in VPLS frames sent to this PE.

However, [RFC4761] does not describe the behavior of PEs in a mixed environment where some PEs support control-word/sequencing and others do not.

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 Problem

[RFC4761] uses a VPLS BGP NLRI to specify the required behavior off multiple PEs. The behavior required off the multiple PEs identified by the NLRI indicates the VPLS label they should use in the VPLS traffic being forwarded to this PE. Additionally, by using the "control flags" it specifies whether the other PEs (in the same VPLS) should use control-word or sequenced-delivery for packets forwarded to this PE. These are respectively indicated by the C and the S bits in the "control flags" as specified in section 3.2.4 in [RFC4761].

[RFC4761] requires that if the advertising PE sets the C and S bits, the receiving PE MUST honor the same by inserting control word (CW) and by including sequence numbers respectively.

However, in a BGP VPLS deployment there would often be cases where a PE receiving the VPLS BGP NLRI may not have the ability to insert a CW or include sequencing information inside PW frames. In that case the behavior of BGP VPLS needs to be further specified.

This document enhances the meaning of the control flags in layer2 extended community in the BGP VPLS NLRI for an environment where not every PE in a VPLS has the ability or the configuration to honor the control flags received from the PE advertising the BGP NLRI.

3 Updated meaning of control flags in the layer2 info extended community

Currently, the CW setting is not negotiated. Rather, if a PE sets the C-bit, it expects to receive VPLS packets with a control word, and will send packets the same way. If the PEs at both ends of a pseudowire don't agree on the setting of the C-bit, the PW doesn't come up. Similarly for the S-bit.

This memo changes the meaning of the C-bit and the S-bit in the control flags. If a PE sets the C-bit in its NLRI, it means that the PE can send and receive packets with a control word. If the PEs at both ends of a PW set the C-bit, control words are used in both directions of the PW. If both PEs send a C-bit of 0, control words are not used on the PW. These two cases behave as before.

However, if the PEs don't agree on the setting of the C-bit, control words are not used on that PW. This behavior is new; the old behavior is that the PW doesn't come up.

The behavior for the S-bit is similar.

4 Using p2mp LSP as transport for CW-marked VPLS frames

BGP VPLS can be used over either point-2-point LSPs acting as transport between the VPLS PEs. Alternately, BGP VPLS may also be used over p2mp LSPs with the source of the p2mp LSPs rooted at the PE advertising the VPLS BGP NLRI.

In a network that uses p2mp LSPs as transport for BGP VPLS, in a given VPLS there may be some PEs that support control-word while others do not. In such a setup, a source PE that supports control-word / sequenced-delivery should setup 2 different p2mp trees - one which has as its leaves those VPLS PEs that are advertising the C/S-bits as 1, and another p2mp LSP whose leaves are PEs that are not advertising C/S-bits as 1.

5 Illustrative diagram

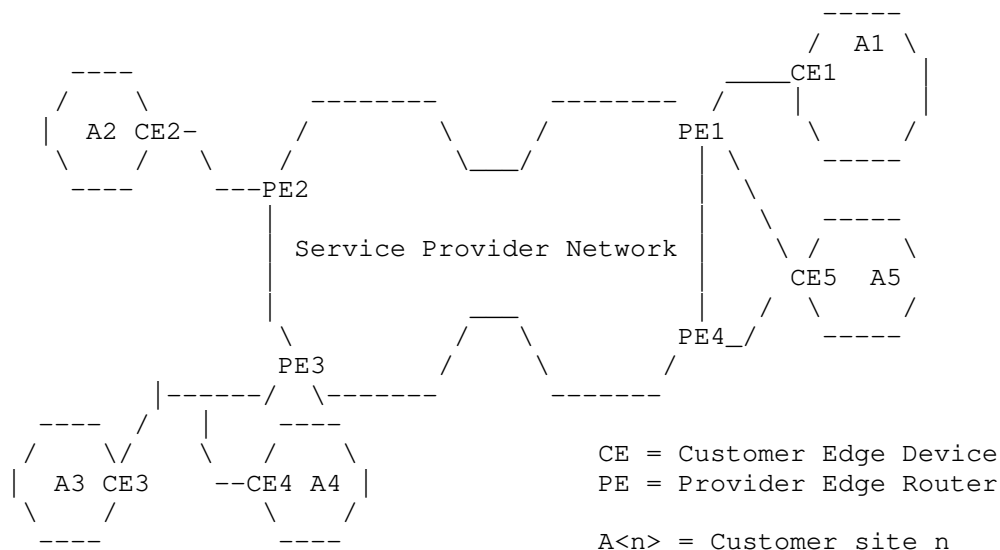


Figure 1: Example of a VPLS

In the above topology, let there be a VPLS configured with the PEs as displayed. Let PE1 be the PE under consideration that is CW enabled. Let PE2 and PE3 also be CW enabled. Let PE4 not be CW enabled. PE1 will advertise a VPLS BGP NLRI, containing the C/S bits marked as 1. PE2 and PE3 on learning of NLRI from PE1, shall include the control word in VPLS frames being forwarded to PE1. However, PE4 which does not have the ability to include control-word.

As per [RFC4761], PE1 would have an expectation that all other PEs forward traffic to it by including CW.

However, to support the mixed-CW environment as above, PE1 will bring up the PW with PE4 despite the CW mismatch. Additionally, it will setup its data-plane such that it will strip the control-word only for those VPLS packets that are received from PEs that are themselves indicating their desire to receive CW marked packets. So, PE1 will setup its data plane to strip-off the CW only for VPLs frames received from PEs PE2 and PE3. PE1 will setup its data plane to not strip CW from frames received from PE4.

6 Security Considerations

No new security issues.

7 IANA Considerations

None.

8 References

8.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4761] Kompella, K., Y. Rekhter, Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling, RFC 4761, January 2007.
- [RFC4385] Bryant, S., Swallow G., Martini L., D. McPherson, Pseudowire Emulation Edge-to-Edge (PWE3) Control Word, RFC 4385, February 2006.

Authors' Addresses

Ravi Singh
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

EMail: ravis@juniper.net

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

EMail: kireeti@juniper.net

INTERNET-DRAFT
Intended Status: Standards Track
Expires: April 22, 2014

Vengada Prasad Govindan
Samer Salam
Ali Sajassi
Cisco
October 19, 2013

Proactive fault detection in E-VPN
draft-vgovindan-l2vpn-evpn-bfd-00

Abstract

This document proposes a proactive, in-band network OAM mechanism to detect connectivity faults that affect unicast and multi-destination paths in an E-VPN network. The multi-destination paths are used by Broadcast, unknown Unicast and Multicast (BUM) traffic. The mechanisms proposed in the draft use the principles of the widely adopted Bidirectional Forwarding Detection (BFD) protocol.

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents
 (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	2
1.1	Terminology	3
2.	Scope of fault detection mechanisms proposed in this document	3
2.1	Fault Detection of BUM traffic using ingress replication (MP2P)	3
2.1.1	Bootstrapping BFD sessions at the head of the MP2P tunnel	4
2.1.2	Bootstrapping BFD sessions at the tail nodes of the MP2P tunnel	4
2.2	Fault Detection of BUM traffic using P2MP tunnels (LSM)	4
2.2.1	Bootstrapping BFD sessions at the root of the P2MP tunnel	5
2.2.2	Bootstrapping BFD sessions at the tail nodes of the P2MP tunnel	5
2.3	Fault Detection of unicast traffic (P2P)	5
3	BFD packet encapsulation	5
3.1	Encapsulation using IP headers	5
3.1.1	Ingress replication	5
3.1.2	LSM	5
3.1.3	Unicast	6
3.2	Using GAL/G-Ach encapsulation without IP headers	6
3.2.1	Ingress replication	6
3.2.2	LSM	6
3.2.3	Unicast	6
4.	Scalability Considerations	6
5.	Security Considerations	7
6	IANA Considerations	7
7.	Acknowledgments	7
8	References	7
8.1	Normative References	7
8.2	Informative References	8
	Authors' Addresses	8

1 Introduction

[EVPNOAM] outlines the OAM requirements of Ethernet VPN networks [EVPN]. This document proposes mechanisms for proactive fault detection at the network OAM layer of E-VPN. These mechanisms could either be deployed for periodic and proactive monitoring, or be triggered by specific events to aid troubleshooting. E-VPN fault detection mechanisms need to consider unicast and BUM traffic separately since they map to different FECs in E-VPN. Since BUM traffic can be transported using MP2P or P2MP tunnels, this document proposes slightly different fault detection mechanisms to suit each type using the principles of Point-to-multipoint BFD[P2MPBFD]. Please note that this document uses the term E-VPN loosely to include [E-VPN], [PBB-EVPN] as well as [TRILL-EVPN].

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. Scope of fault detection mechanisms proposed in this document

This section proposes proactive fault detection using BFD mechanisms for:

- a. BUM traffic using MP2P tunnels (ingress replication).
- b. BUM traffic using P2MP tunnels (LSM).
- c. Unicast traffic

The approach takes advantage of the inclusive multicast route used in E-VPN to advertise the multi-destination FEC for bootstrapping the BFD sessions. Earlier approaches for P2MP BFD [MPLSCVBFD] have used periodic MPLS ping requests to bootstrap P2MP BFD sessions over MPLS.

2.1 Fault Detection of BUM traffic using ingress replication (MP2P)

Ingress replication uses separate MP2P tunnels for transporting BUM traffic from the ingress PE (head) to a set of one or more egress PEs (tails). The fault detection mechanism proposed by this document takes advantage of the fact that a unique copy is made by the head for each tail. Another key aspect to be considered in E-VPN is the advertisement of the inclusive multicast route. The BUM traffic flows from a head node to a particular tail only after the head receives the inclusive multicast route containing the BUM EVPN label (downstream allocated) corresponding to the MP2P tunnel.

2.1.1 Bootstrapping BFD sessions at the head of the MP2P tunnel

There could exist multiple BFD sessions between a head of the multi-destination tunnel and an individual tail due to the usage of entropy labels [MPLSEL] for an inclusive multicast FEC. A P2MP BFD session could be identified at the tail node using the source IP (if present in the BFD packet) and the attributes of the inclusive multicast FEC (the BUM label, label stack and optionally the entropy label). To simplify such a lookup, we take advantage of the fact that the head replicates a BUM packet for each tail by inserting unique discriminators in each copy of the (replicated) BFD packet. These discriminators are exchanged out-of-band using MPLS ping [BFD-MPLS] before the start of the BFD session. The head PE performing ingress replication MUST initiate the LSP ping using the inclusive multicast FEC [EVPNPING] upon receiving an inclusive multicast route from a tail. The receiving tail MUST generate a unique discriminator to identify the tuple of inclusive multicast FEC and the source IP address. This discriminator MUST be communicated back to the head using the response to the MPLS ping. The head MUST insert the corresponding discriminator generated by the tail for each copy of the BFD packet in the yourDisc field of the BFD header. It is obvious that the MPLS ping to perform discriminator exchange is required only once to bootstrap the session.

2.1.2 Bootstrapping BFD sessions at the tail nodes of the MP2P tunnel

The tail nodes MUST bootstrap a BFD session based on the incoming MPLS ping initiated by the head [P2MPBFD].

2.2 Fault Detection of BUM traffic using P2MP tunnels (LSM)

The case of using P2MP tunnels for distributing BUM traffic presents a different challenge for using BFD. Clearly, the yourDisc of the BFD packet MUST be zero [P2MPBFD] as the packet is multicast from the root unlike ingress replication where individual copies are made from the head. However the MPLS label that identifies the P-Tunnel used for forwarding the multi-destination traffic provides a convenient method of identifying the source and the FEC (multi-destination tree) being tracked by the BFD session. The tails of the multi-destination tree MUST use the MPLS label identifying the P-tunnel to de-multiplex the BFD packet. In the case of Aggregate Inclusive trees, where the root of the multi-destination tree reuses the same LSP label for traffic of various EVIs, the tail node MUST use the MPLS labels of the P-Tunnel and the upstream assigned label which the PE has bound uniquely to the EVI. The myDisc of the BFD packet is filled with an unique value allocated by the root to identify the multipath session.

2.2.1 Bootstrapping BFD sessions at the root of the P2MP tunnel

The P2MP BFD sessions MUST be bootstrapped at the head [P2MPBFD] as soon as there is one receiver for the MDT traffic.

2.2.2 Bootstrapping BFD sessions at the tail nodes of the P2MP tunnel

The PMP BFD sessions MUST be bootstrapped at the tail upon reception of the P2MP BFD packets from the head. The tail MUST use the P2MP MDT label to de-multiplex the incoming BFD packet.

2.3 Fault Detection of unicast traffic (P2P)

The mechanisms specified in BFD for MPLS LSPs [BFD-MPLS] can be applied to bootstrap and maintain BFD sessions for unicast EVPN traffic. The discriminators required for de-multiplexing the BFD sessions can be exchanged using MPLS ping before starting the BFD session. This is needed since the MPLS label stack does not contain enough information to disambiguate the sender of the packet. The usage of MPLS entropy labels take care of addressing the requirement of monitoring faults of the various paths of the multi-path server layer network [MPLSEL].

3 BFD packet encapsulation

Two BFD encapsulations are possible [VCCV-BFD].

3.1 Encapsulation using IP headers

3.1.1 Ingress replication

The packet contains the following labels: LSP label (transport) when not using PHP, BUM label, SH label (where applicable) and the optional entropy label. The source IP of the packet is set to that of the originating PE. The destination IP MUST be 127/8 for IPv4 or 0:0:0:0:0:FFFF: to 127.0.0.0/104 for IPv6. The UDP port indicates the type of the payload as BFD control packet. The discriminator values of BFD are obtained through negotiation through the out-of-band MPLS ping.

3.1.2 LSM

The packet contains the following labels: P-Tunnel label, upstream allocated label which the PE has bound uniquely to the EVI (aggregate inclusive tress only). The source IP of the packet is set to that of the originating PE. The destination IP MUST be 127/8 for IPv4 or 0:0:0:0:0:FFFF: to 127.0.0.0/104 for IPv6. The UDP port

indicates the type of the payload as BFD control packet. The yourDisc value is set to 0 and the myDisc value is uniquely generated by the root.

3.1.3 Unicast

The packet contains the following labels: LSP label (transport) when not using PHP, E-VPN Unicast label and the optional entropy label. The source IP of the packet is set to that of the originating PE. The destination IP MUST be 127/8 for IPv4 or 0:0:0:0:0:FFFF: to 127.0.0.0/104 for IPv6. The UDP port indicates the type of the payload as BFD control packet. The discriminator values of BFD are obtained through negotiation through the out-of-band MPLS ping.

3.2 Using GAL/G-Ach encapsulation without IP headers

3.2.1 Ingress replication

The packet contains the following labels: LSP label (transport) when not using PHP, BUM label and the SH label (where applicable). The G-Ach type is set to 0x0007 [VCCV-BFD]. The discriminator values of BFD are obtained through negotiation through the out-of-band MPLS ping.

3.2.2 LSM

The packet contains the following labels: label identifying the P-Tunnel, upstream label which the PE has bound uniquely to the EVI (for aggregate inclusive trees only). The G-Ach type is set to 0x0007 [VCCV-BFD]. The yourDisc value is set to 0 and the myDisc value is uniquely generated by the root.

3.2.3 Unicast

The packet contains the following labels: LSP label (transport) when not using PHP, E-VPN Unicast label and the optional entropy label. The G-Ach type is set to 0x0007 [VCCV-BFD]. The discriminator values of BFD are obtained through negotiation through the out-of-band MPLS ping.

4. Scalability Considerations

The mechanisms proposed in this draft could either be run periodically to monitor the network proactively or in response to specific network events to aid troubleshooting in a demand-driven

manner. Since the mechanisms proposed could affect the load of the network elements as networks scale to support a large number of EVIs, caution needs to be exercised in choosing the parameters and the interfaces on which these mechanisms are enabled.

5. Security Considerations

This document does not introduce any new security issues, the security considerations defined in [BFD] and [P2MPBFD] apply in this document.

6 IANA Considerations

No new requests are made to IANA by this document.

7. Acknowledgments

We thank Tina Lam, Jose Liste and Mudigonda Mallik for their valuable input, discussions and comments.

8 References

8.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [BFD] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [BFD-MPLS] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.
- [VCCV-BFD] Nadeau, T., Ed., and C. Pignataro, Ed., "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, June 2010.
- [P2MPBFD] Katz, D. and D. Ward, "BFD for multipoint networks", draft-ietf-multipoint-01.txt, work in progress, June 2013.
- [EVPNOAM] Salam et al, "E-VPN OAM Requirements and Framework", draft-salam-l2vpn-evpn-oam-req-frmwrk-00.txt, work in progress, October 2012.

[EVPNPING] Jain et al, "LSP-Ping Mechanisms for E-VPN and PBB-EVPN" ,
draft-jain-l2vpn-evpn-lsp-ping-01,
work in progress, February 2013.

[MPLSCVBFD] Swallow et al, "Connectivity Verification for Multicast
Label Switched Paths",
draft-ietf-mpls-mcast-cv-00
work in progress, April 2007.

8.2 Informative References

[EVPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN",
draft-ietf-l2vpn-evpn-02.txt, work in progress,
October 2012.

[PBB-EVPN] Sajassi et al, "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-05,
July 2013.

[TRILL-EVPN]
Sajassi et al., "TRILL-EVPN",
draft-ietf-l2vpn-trill-evpn-00.txt,
work in progress, June 2012.

[MPLSEL] Kompella et al,
"The Use of Entropy Labels in MPLS Forwarding",
RFC 6790, November 2012.

Authors' Addresses

Vengada Prasad Govindan
EMail: venggovi@cisco.com

Samer Salam
EMail: ssalam@cisco.com

Ali Sajassi
EMail:sajassi@cisco.com