TRILL Working Group
INTERNET-DRAFT

Donald Eastlake
Zhenbin Li
Shunwan Zhuang
Haibo Wang
Huawei Technologies

Intended status: Proposed Standard
Expires: March 4, 2019

September 5, 2018

EVPN All Active Usage Enhancement
<draft-eastlake-bess-enhance-evpn-all-active-01.txt>

Abstract

A principal feature of EVPN is the ability to support multihoming from a customer equipment (CE) to multiple provider edge equipment (PE) active with all-active links. This draft specifies an improvement to load balancing such links.

1. Introduction

A principal feature of EVPN (Ethernet VPN [RFC7432]) is the ability to support multihoming from a customer equipment (CE) to multiple provider edge equipment (PE) with links used in an all-active redundancy mode. That mode is where a device is multihomed to a group of two or more PEs and where all PEs in such redundancy group can forward traffic to/from the multihomed device or network for a given VLAN [RFC7209]. This draft specifies an improvement in load balancing such PE to CE all-active multi-homing links.

In the case where a CE is multihomed to multiple PE nodes, using a Link Aggregation Group (LAG) with All-Active redundancy, it is possible that only a single PE learns a set of the MAC addresses associated with traffic transmitted by the CE. This leads to a situation where remote PE nodes receive MAC/IP Advertisement routes for these addresses from a single PE, even though multiple PEs are connected to the multihomed segment.

To address this issue, EVPN introduces the concept of "aliasing", which is the ability of a PE to signal that it has reachability to an EVPN instance (EVI) on a given Ethernet segment (ES) even when it has learned no MAC addresses from that EVI/ES. The Ethernet A-D per EVI route is used for this purpose. A remote PE that receives a MAC/IP Advertisement route with a non-reserved ESI SHOULD consider the advertised MAC address to be reachable via all PEs that have advertised reachability to that MAC address's EVI/ES via the combination of an Ethernet A-D per EVI route for that EVI/ES (and Ethernet tag, if applicable) AND Ethernet A-D per ES routes for that ES with the "Single-Active" bit in the flags of the ESI Label extended community set to 0.

1.1 Terminology and Acronyms

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

This document uses the following acronyms and terms:

A-D - Auto Discovery.

All-Active Redundancy Mode - When a device is multihomed to a group of two or more PEs and when all PEs in such redundancy group can forward traffic to/from the multihomed device or network for a

given VLAN.

CE - Customer Edge equipment.

ES - Ethernet Segment.

ESI - Ethernet Segment Identifier.

EVI - EVPN Instance.

EVPN - Ethernet VPN [RFC7432].

FRR - Fast ReRoute.

MAC - Media Access Control.

PE - Provider Edge equipment.

Single-Active Redundancy Mode - When a device or a network is multihomed to a group of two or more PEs and when only a single PE in such a redundancy group can forward traffic to/from the multihomed device or network for a given VLAN.

VPN - Virtual Private Network.

2. Improved Load Balancing

Consider the example in Figure 1. CE1 is multihomed to PE1 and PE2. CE1 typically uses a hash algorithm to determine whether to send a particular traffic to PE1 or to PE2. Thus, if such traffic from CE1 is only sent to PE1, then PE1 will learn CE1's MAC address(es) and that PE2 will not.

PE3 and PE4 can do aliasing [RFC7432] because PE1 and PE2 will be advertising the same ESI. Thus PE3 and PE4 will expect that a MAC address reachable from PE1 will also be reachable from PE2. This aliasing will cause PE3 and PE4 to load balance to CE1's MAC(s), sending some traffic to PE1 and some to PE2.

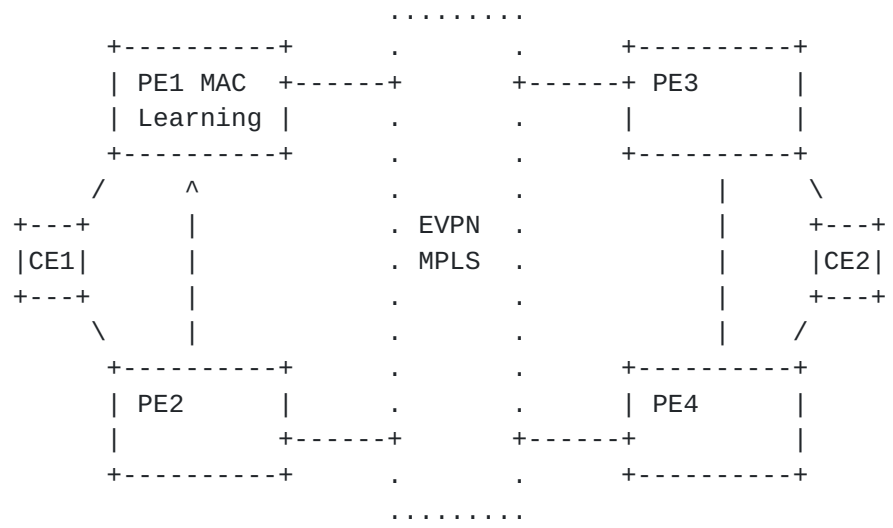

Figure 1. Current Situation

There are two problems associated with this situation that are described in the subsections below. Section 3 describes the mechanism to address these problems.

2.1 Problem 1: Traffic Bypassing

Since PE2 has not learning CE1's MAC(s), the MAC lookup at PE2 will find that MAC address associated with PE1. PE2 will then tunnel the traffic to PE1.

As an enhancement that solves this problem, PE1 can send MAC address(es) with VLAN and ESI information. PE2 will then receive the MAC address(es) and VLAN that PE1 associates with the ESI and PE2 can use this to update its forwarding tables (see Figure 2). As a result, when traffic addressed to a CE1 MAC arrives at PE2, it can send it on

the appropriate local interface and VLAN. This avoids the unnecessary

extra hop through PE1 for such traffic arriving at PE2.

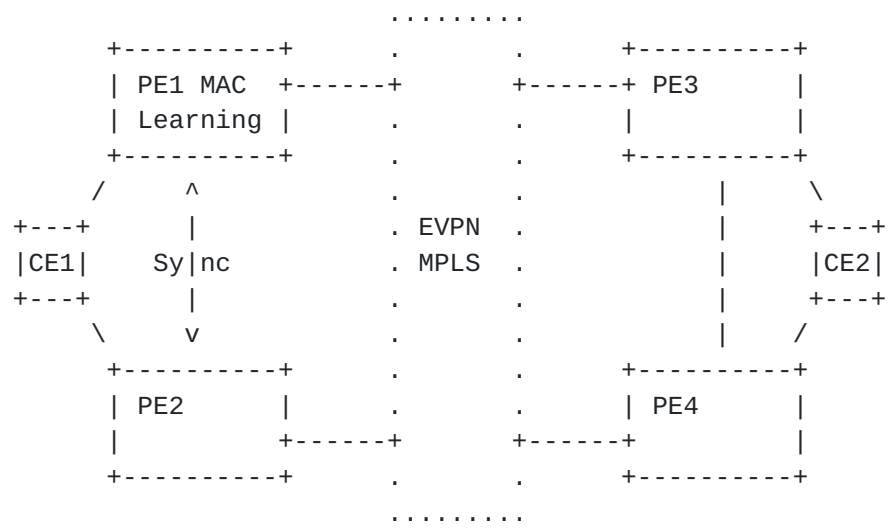

Figure 2. With Enhancement

2.2 Problem 2: VID Encapsulation Confusion

If CE1 is connected through a VLAN and has only one VLAN under the EVPN instance of PE2, the unicast traffic can be directly sent to the appropriate interface and encapsulated with the appropriate VID and forwarded to CE1.

However, there may be multiple ways for CE1 to connect to PE1 and PE2, including Ethernet Tag, Ethernet Tag termination, and Q-in-Q. PE2 cannot always obtain the appropriate VLANs and in such cases PE2 is missing the information needed to forward the unicast traffic to CE1 directly.

3. VLAN-Redirect-Extended Community Attribute

This document defines a new BGP extended community attribute called the VLAN-Redirect-Extended Community attribute as shown in Figure 3.

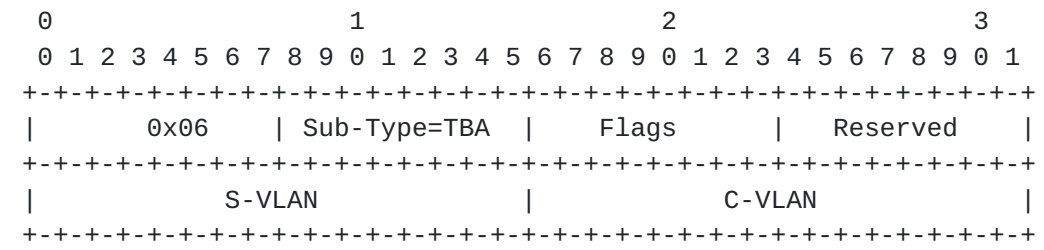

Figure 3. VLAN-Redirect-Extended Community Attribute

Where:

0x06: EVPN Extended Community Type field.

Sub-Type: Sub-Type field indicating that the extended community attribute is a VLAN-Redirect-Extended Community attribute, and the value is TBA as assigned by the IANA.

Flags: 8 bits of identification information. Bit 0 set to 0 indicates that the action is redirected to the VLANs in this community

Reserved: Not used. MUST be sent as zero and ignored on receipt.

S-VLAN: Outer VLAN information, can not be 0, 0 is illegal value

C-VLAN: Inner VLAN information. When 0, it means there is no C-VLAN.

4. Operation

Operation with the solution specified in Section 3 and the topology shown in Figure 2 is described below.

4.1 Establishment

1. PE1 learns MAC addresses from CE1, advertises them to PE2, carries the ESI value as ES1 and the next hop as PE1, and carries the VLAN- Redirect-Extended Community attributes.

2. PE2 receives the MAC route advertised by PE1 and finds the interface that connects to CE1 locally according to the ESI value. At the same time, PE2 fills in the VLAN information according to the VLAN-Redirect-Extended Community attributes

3. At the same time, PE2 generates a fast reroute (FRR) entry according to the next hop information (PE1) of the MAC route, that is, a MAC address entry on PE2, where the primary path points to the CE1 link and the standby path points to PE1

4. PE2 also sends the MAC as a local MAC route to PE1

5. PE1 receives the MAC route advertised by PE2 and generates the FRR entry with the MAC route learned by CE1, that is, the MAC address entry on PE1, with the primary path pointing to the CE1 link and the secondary path pointing to PE2

4.2 Handling Link Failure

1. When the link between PE1 and CE1 fails, PE1 withdraws the MAC address that advertised to PE2

2. PE2 receives the MAC withdrawal from PE1, does not delete the MAC immediately, but starts an aging timer, and does not withdraw the MAC address that PE1 advertised to PE2.

3. When the aging timer expires, if PE2 cannot receive the traffic from CE1, then PE2 withdraws the MAC address that was advertised to PE2 by PE1 and deletes the MAC entry. If PE2 can communicate directly with CE1, it just eliminates the FRR standby path to PE1.

5. IANA Considerations

IANA is requested to assign a new EVPN Extended Community SubType as follows:

Sub-Type Value	Name	Reference
TBA	VLAN-Redirect Extended Community	[this doc]

6. Security Considerations

TBD

For general EVPN Security Considerations, see [RFC7432].

Normative References

[RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC7432] - Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

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

Informative References

[RFC7209] - Sajassi, A., Aggarwal, R., Uttaro, J., Bitar, N., Henderickx, W., and A. Isaac, "Requirements for Ethernet VPN (EVPN)", RFC 7209, DOI 10.17487/RFC7209, May 2014, <https://www.rfc-editor.org/info/rfc7209>.

Acknowledgements

The authors of this document would like to thank the following for their comments and review of this document:

TBD

Authors' Addresses

Donald E. Eastlake, 3rd
Huawei Technologies
1424 Pro Shop Court
Davenport, FL USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Zhenbin Li
Huawei Technologies
Huawei Bldg., No. 156 Beiqing Road
Beijing 100095 China

Email: lizhenbin@huawei.com

Shunwan Zhang
Huawei Technologies
Huawei Bldg., No. 156 Beiqing Road
Beijing 100095 China

Email: zhuangshunwan@huawei.com

Haibo Wang
Huawei Technologies
Huawei Bldg., No. 156 Beiqing Road
Beijing 100095 China

Email: rainsword.wang@huawei.com

Copyright, Disclaimer, and Additional IPR Provisions