

INTERNET-DRAFT
Intended Status: Proposed Standard

Patrice Brissette
Ali Sajassi
Luc Andre Burdet
Cisco Systems

Daniel Voyer
Bell Canada

Expires: April 23, 2019

October 20, 2018

EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols
draft-brissette-bess-evpn-l2gw-proto-03

Abstract

The existing EVPN multi-homing load-balancing modes defined are Single-Active and All-Active. Neither of these multi-homing mechanisms are appropriate to support access networks with Layer-2 Gateway protocols such as G.8032, MPLS-TP, STP, etc. These Layer-2 Gateway protocols require a new multi-homing mechanism defined in this draft.

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) 2018 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
1.2 Acronyms	3
2. Solution	5
3. Requirements	6
4. Handling of Topology Change Notification (TCN)	7
5. ESI-label Extended Community Extension	8
6. EVPN MAC Flush Extcomm	8
7. EVPN Inter-subnet Forwarding	9
8. Conclusion	9
9. Security Considerations	10
10. Acknowledgements	10
11. IANA Considerations	10
12. References	10
12.1 Normative References	10
12.2 Informative References	10

1. Introduction

Existing EVPN multi-homing mechanisms of Single-Active and All-Active are not sufficient to support access Layer-2 Gateway protocols such as G.8032, MPLS-TP, STP, etc.

These Layer-2 Gateway protocols require that a given flow of a VLAN (represented by {MAC-SA, MAC-DA}) to be only active on one of the PEs in the multi-homing group. This is in contrast with Single-Active redundancy mode where all flows of a VLAN are active on one of the multi-homing PEs and it is also in contrast with All-Active redundancy mode where all L2 flows of a VLAN are active on all PEs in the redundancy group.

This draft defines a new multi-homing mechanism "Single-Flow-Active" which defines that a VLAN can be active on all PEs in the redundancy group but a single given flow of that VLAN can be active on only one of the PEs in the redundancy group. In fact, the carving scheme, performed by the DF (Designated Forwarder) election algorithm for these L2 Gateway protocols, is not per VLAN but rather for a given VLAN. A selected PE in the redundancy group can be the only Designated Forwarder for a specific L2 flow but the decision is not taken by the PE. The loop-prevention blocking scheme occurs in the access network.

EVPN multi-homing procedures need to be enhanced to support Designated Forwarder election for all traffic (both known unicast and BUM) on a per L2 flow basis. This new multi-homing mechanism also requires new EVPN considerations for aliasing, mass-withdraw, fast-switchover and [EVPN-IRB] as described in the solution section.

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].

1.2 Acronyms

AC	: Attachment Circuit
BUM	: Broadcast, Unknown unicast, Multicast
DF	: Designated Forwarder
EVLAG	: EVPN LAG (equivalent to EVPN MC-LAG)
GW	: Gateway
L2 Flow	: a given flow of a VLAN, represented by (MAC-SA, MAC-DA)
L2GW	: Layer-2 Gateway
G.8032	: Ethernet Ring Protection
MST-AG	: Multi-Spanning Tree Access Gateway

INTERNET DRAFT draft-brisette-bess-evpn-l2gw-proto October 20, 2018

REP-AG : Resilient Ethernet Protocol Access Gateway
TCN : Topology Change Notification

2. Solution

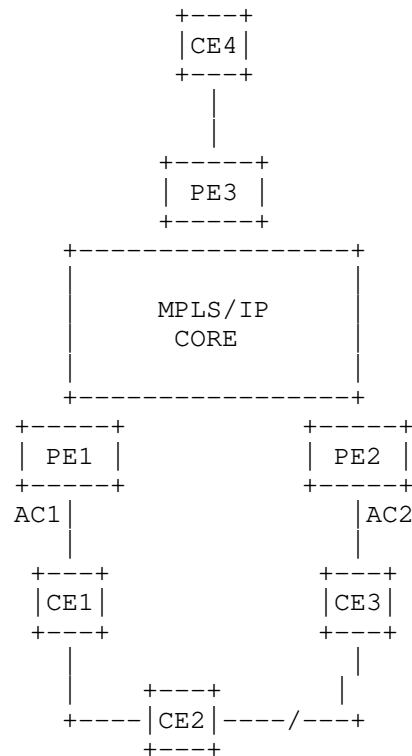


Figure 1 EVPN network with L2 access GW protocols

Figure 1. shows a typical EVPN network with an access network running a L2GW protocol; typically one of the following: G.8032, STP, MPLS-TP, etc. The L2GW protocol usually starts from AC1 (on PE1) up to AC2 (on PE2) in an open "ring" manner. AC1 and AC2 interfaces of PE1 and PE2 are participants in the access protocol. PE1 and PE2 are peering PEs (EVLAG capable) in a redundancy group sharing a same ESI. The L2GW protocol is used for loop avoidance. In above example, the loop is broken on the right side of CE2. In the proposed Single-Flow-Active mode, PE1 and PE2 'Access Gateway' load-balancing mode shares similarities with both Single-Active and All-Active. DF election must not result in blocked ports or portions of the access may become isolated. Additionally, the reachability between CE1/CE2 and CE3 is achieved with the forwarding path through the EVPN MPLS/IP core side. Thus, the ESI-Label filtering of [RFC7432] is disabled for Single-Flow-Active Ethernet segments.

Finally, PE3 behaves according to EVPN rules for traffic to/from

PE1/PE2. Peering PE, selected per L2 flow, is chosen by the L2GW protocol in the access, and is out of EVPN control. From PE3 point of view, some of the L2 flows coming from PE3 may reach CE3 via PE2 and some of the L2 flows may reach CE1/CE2 via PE1. A specific L2 flow never goes to both peering PEs. Therefore, aliasing cannot be performed by PE3. That node operates in a single-active fashion for these L2 flows. The backup path which is also setup for rapid convergence, is not applicable here. For example, in Figure 1, if a failure happens between CE1 and CE2, L2 flows coming from CE4 behind PE3 destined to CE1 still goes through PE1 and shall not switch to PE2 as a backup path. On PE3, there is no way to know which L2 flow specifically is affected. During the transition time, PE3 may flood until unicast traffic recovers properly.

3. Requirements

The EVPN L2GW framework for L2GW protocols in Access-Gateway mode, consists of the following rules:

- o Peering PEs MUST share the same ESI.
- o The Ethernet-Segment DF election MUST NOT be performed and forwarding state MUST be dictated by the L2GW protocol. In Access Gateway mode, both PEs are usually in forwarding state. In fact, access protocol guarantees drive that state.
- o Split-horizon filtering is NOT needed because L2GW protocol ensures there will never be loop in the access network. The forwarding between peering PEs MUST also be preserved. In figure 1, CE1/CE2 device may need reachability with CE3 device. ESI-filtering capability MUST be disabled. PE MUST NOT advertise corresponding ESI-label to other PEs in the redundancy group, or apply it if it is received.
- o ESI-label BGP-extcomm MUST support a new multi-homing mode named "Single-Flow-Active" corresponding to the single-active behaviour of [RFC7432], applied per flow.
- o Upon receiving ESI-label BGP-Extcomm with the single-flow-active load-balancing mode, remote PE MUST:
 - Disable ESI-Label processing
 - Disable aliasing (at Layer-2 and Layer-3 [EVPN-IRB])
- o The Ethernet-Segment procedures in the EVPN core such as per ES/EAD and per EVI/EAD routes advertisement/withdraw, as well as MAC and MAC+IP advertisement, remains as explained in [RFC7432] and [EVPN-IRB].

- o For fast-convergence, remote PE3 MAY set up two distinct backup paths on a per-flow basis:

- { PE1 active, PE2 backup }
- { PE2 active, PE1 backup }

- o MAC mobility procedures SHALL have precedence in Single-Flow-Active for tracking host reachability over backup path procedure.

4. Handling of Topology Change Notification (TCN)

In order to address rapid Layer-2 convergence requirement, topology change notification received from the L2GW protocols must be sent across the EVPN network to perform the equivalent of legacy L2VPN remote MAC flush.

The generation of TCN is done differently based on the access protocol. In the case of STP (REP-AG) and G.8032, TCN gets generated in both directions and thus both of the dual-homing PEs receive it. However, with STP (MST-AG), TCN gets generated only in one direction and thus only a single PE can receive it. That TCN is propagated to the other peering PE for local MAC flushing, and relaying back into the access.

In fact, PEs have no direct visibility on failures happening in the access network neither on the impact of those failures over the connectivity between CE devices. Hence, both peering PEs require to perform a local MAC flush on corresponding interfaces.

There are two options to relay the access protocol's TCN to the peering PE: in-band or out-of-band messaging. The first method is better for rapid convergence, and requires a dedicated channel between peering PEs. An EVPN-VPWS connection MAY be dedicated for that purpose, connecting the Untagged ACs of both PEs. The latter choice relies on a new MAC flush extended community in the Ethernet Auto-discovery per EVI route, defined below. It is a slower method but has the advantage of avoid the usage of a dedicated channel between peering PEs.

Peering PE, upon receiving TCN from access, MUST:

- o As per legacy VPLS, perform a local MAC flush on the access-facing interfaces. An ARP probe is also sent for all hosts previously locally-attached.

- o Advertise per EVI/EAD route along with a new MAC-flush BGP Extended Community in order to perform a remote MAC flush and steer L2 traffic to proper peering PE. The sequence number is incremented by one as a flushing indication to remote PEs.
- o Ensure MAC and MAC/IP route re-advertisement, with incremented sequence number when host reachability is NOT moving to peering PE. This is to ensure a re-advertisement of current MAC and MAC/IP which may have been flushed remotely upon MAC Flush extcomm reception. In theory, it should happen automatically since peering PE, receiving TCN from the access, performs local MAC flush on corresponding interface and will re-learn that local MAC or MAC/IP at ARP probe reply.
- o When MST-AG runs in the access, a dedicated EVPN-VPWS connection MAY be used as an in-band channel to relay TCN between peering PEs. That connection may be auto-generated or can simply be directly configured by user.

5. ESI-label Extended Community Extension

In order to support the new EVPN load-balancing mode (single-flow-active), the ESI-label extcomm is extended. The 1 octet flag field, as part of the ESI-label Extcomm, is updated as follow:

Each ESI Label extended community is encoded as an 8-octet value, as follows:

1																2																3																															
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1																																
Type=0x06																Sub-Type=0x01																Flags(1 octet)																Reserved=0															
Reserved=0																ESI Label																																															

Low-order bit: [7:0]

[2:0]- 000 = all-active,
 001 = single-active,
 010 = single-flow-active,
 others = unassigned

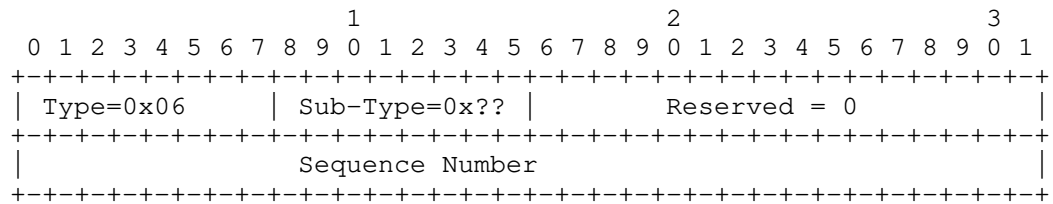
[7:3]- Reserved

6. EVPN MAC Flush Extcomm

A new BGP Extended community, similar to MAC mobility BGP-extcomm, is required by the TCN procedure. It may get advertised along with Ethernet Auto-discovery routes (per EVI/EAD) upon reception of TCN

from the access. When this extended community is used, it indicates, to all remote PEs that all MAC addresses associated with that EVI/ESI are "flushed" i.e. unresolved. They remain unresolved until remote PE receives a route update / withdraw for those MAC addresses; the MAC may be readvertised by the same PE, or by another, in the same ESI.

The sequence number used is of local significance from the originating PE, and is not used for comparison between peering PEs. Rather, it is used to signal via BGP successive MAC Flush requests from a given PE.



7. EVPN Inter-subnet Forwarding

EVPN Inter-subnet forwarding procedures in [EVPN-IRB] works with the current proposal and does not require any extension. Host routes continue to be installed at PE3 with a single remote nexthop, no aliasing.

8. Conclusion

EVPN Multi-Homing Mechanism for Layer-2 gateway Protocols solves a true problem due to the wide legacy deployment of these access L2GW protocols in Service Provider networks. The current draft has the main advantage to be fully compliant with [RFC7432] and [EVPN-IRB].

9. Security Considerations

The same Security Considerations described in [RFC7432] and [EVPN-IRB] remain valid for this document.

10. Acknowledgements

Authors would like to thank Thierry Couture for valuable review and inputs with respect to access protocol deployments related to procedures proposed in this document.

11. IANA Considerations

A new allocation of Extended Community Sub-Type for EVPN is required to support the new EVPN MAC flush mechanism.

12. References

12.1 Normative References

[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>>.

12.2 Informative References

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

Authors' Addresses

Patrice Brissette
Cisco Systems
EMail: pbrisset@cisco.com

Ali Sajassi
Cisco Systems
EMail: sajassi@cisco.com

Luc Andre Burdet
Cisco Systems
EMail: lburdet@cisco.com

INTERNET DRAFT draft-brisette-bess-evpn-l2gw-prot0 October 20, 2018

Daniel Voyer
Bell Canada
EMail: daniel.voyer@bell.ca