INTERNET-DRAFT
Intended Status: Proposed Standard

Patrice Brissette
Ali Sajassi
Luc Andre Burdet
Cisco Systems

Daniel Voyer
Bell Canada

Expires: August 30, 2018

February 26, 2018

EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols
draft-brissette-bess-evpn-l2gw-proto-01

Abstract

Existing EVPN multi-homing load-balancing modes are limited to Single-Active and All-Active. Neither of these multi-homing mechanisms are sufficient to support access networks with Layer-2 Gateway protocols such as MST-AG, REP-AG, and G.8032. These Layer-2 Gateway protocols require a new multi-homing mechanism defined in this draft.

Table of Contents

1. Introduction

EVPN existing multi-homing mechanisms of Single-Active and All-Active is not sufficient to support access Layer-2 Gateway protocols such as MST-AG, REP-AG, and G.8032.

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 means a given VLAN can be active on all PEs in the redundancy group but a single flow of that VLAN can only be active on only one of the PEs in the redundancy group. In fact, the carving scheme, performed by the DF election algorithm for these L2GW protocols, is not per VLAN but rather for a given VLAN. A given PE in the redundancy group can be the only Designated Forwarder for a specific L2 flow. The loop-prevention blocking scheme occurs somewhere 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 and fast-switchover 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
REP-AG  : Resilient Ethernet Protocol Access Gateway
```

TCN : Topology Change Notification

2. Solution

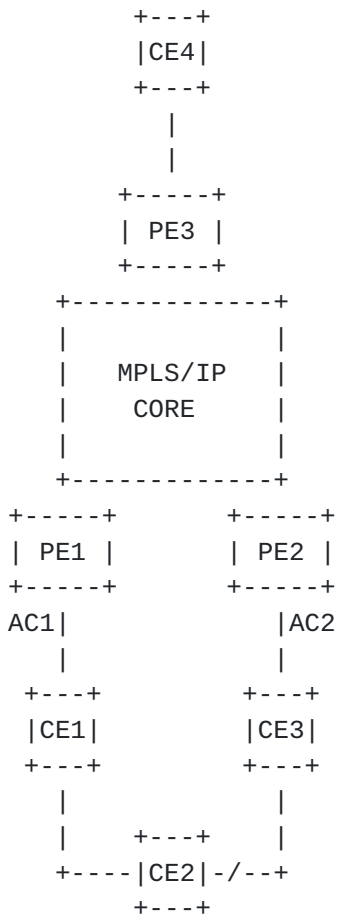


Figure 1 EVPN network with L2 access GW protocols

Figure 1. shows a typical EVPN network with an access network running a L2GW protocol; typically one of the following: MST-AG, REP-AG or G.8032. 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 participant in the access protocol. PE1 and PE2 are peering PEs in a redundancy group and EVLAG capable. The L2GW protocol is used for loop avoidance. In above example, the loop is broken on the right side of CE2. Due to them running in 'Access Gateway' mode, PE1 and PE2 EVLAG load-balancing runs in a way very-similar to all-active. Additionally, the reachability between CE1/CE2 and CE3 that is required is achieved with the forwarding path through the EVPN MPLS/IP core side.

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 flow coming from PE3 may reach CE3 via PE2 and

some of the L2 flow 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 flow. 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 flow coming from CE4 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 will 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 forwarding state MUST be dictated by the L2GW protocol.

o Split-horizon filtering capability 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 disable. PE MUST NOT advertise corresponding ESI-label to other PEs in the redundancy group.

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 aliasing (at Layer-2 and Layer-3)
- Disable ESI-label processing
- Disable backup path programming

The Ethernet-Segment DF election backend procedure such as per ES/EAD and per EVI/EAD routes advertisement/withdraw remains as explained in [RFC7432].

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 topology change notification is done differently based on the access protocol. In the case of REP-AG and G.8032, TCN gets generated in both directions and thus both of the dual-homing PEs receive it. However, with MST-AG, TCN gets generated only in one direction and thus only a single PE can receive it. That TCN ie 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 is dedicated for that purpose. The latter choice relies on the newly defined MAC flush extended community in the Ethernet Auto-discovery per EVI route. 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.

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.

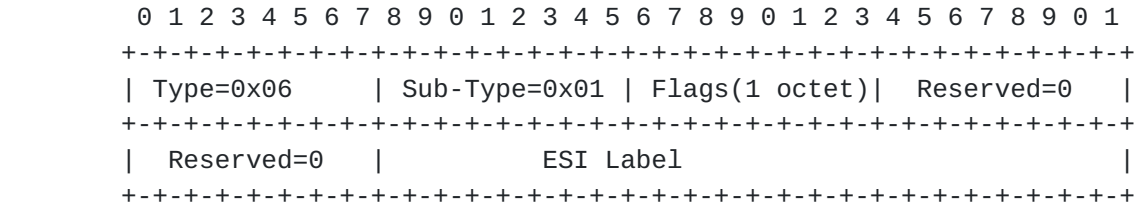
o Ensure MAC/IP route re-advertisement with sequence number is bump up when host reachability is NOT moving to peering PE. This is to ensure a re-advertisement of current MAC which may have been flushed remotely upon MAC Flush extcomm reception. In theory, it should happen automatically since peering PE, receiving TCN from the access, performs local MAC flush on corresponding interface and will re-learn that local MAC.

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:

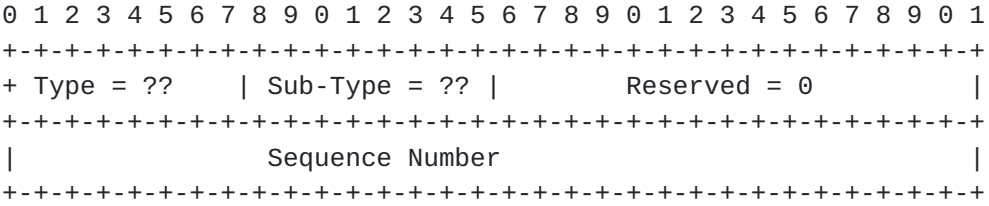


Low-order bit: [7:0]
[2:0]- 000 = all-active,
001 = single-active,
010 = single-flow-active,
others = reserved
[7:3]- Reserved

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. 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 [draft-ietf-bess-evpn-inter-subnet-forwarding]. EVPN-IRB works with the current proposal and does not require any extension.

8. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

9. IANA Considerations

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

10. References

10.1 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, <http://www.rfc-editor.org/info/rfc7432>.

10.2 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, <http://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

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