Workgroup: BESS Working Group
Internet-Draft:
draft-ietf-bess-evpn-l2gw-proto-03
Published: 13 March 2023
Intended Status: Standards Track
Expires: 14 September 2023
Authors: P. Brissette
Cisco Systems
A. Sajassi
Cisco Systems
LA. Burdet, Ed.
Cisco Systems
D. Voyer
Bell Canada

EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols

Abstract

The existing EVPN multi-homing load-balancing modes defined are Single-Active and All-Active. Neither of these multi-homing mechanisms adequately represent ethernet-segments facing access networks with Layer-2 Gateway protocols such as G.8032, (M)STP, REP, MPLS-TP, etc. These loop-preventing Layer-2 protocols require a new multi-homing mechanism defined in this document.

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 https://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 14 September 2023.

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

Table of Contents

1. Introduction

Existing EVPN Single-Active and All-Active redundancy modes do not address the additional requirements of loop-preventing Layer-2 gateway protocols such as G.8032, (M)STP, REP, MPLS-TP, etc.

These Layer-2 Gateway protocols require that a given L2 flow of a VLAN be only active on one of the PEs in the multi-homing group, while another L2 flow of the same VLAN may be active on the other PE. This is in contrast with Single-Active redundancy mode where all flows of a VLAN are active on a single multi-homing PEs and it is also in contrast with All-Active redundancy mode where all flows of a VLAN are active on all PEs in the redundancy group.

This document defines a new Single-Flow-Active redundancy mode specifying that a VLAN can be active on all PEs in the redundancy group but each unique L2 flow of that VLAN can be active on only one of the PEs in the redundancy group at a time. In fact, the

Designated Forwarder election algorithm for these L2 Gateway protocols, is not per VLAN but rather for a given L2 flow. A selected PE in the redundancy group must 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, by the Layer-2 protocol.

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. The Single-Flow-Active multi-homing mechanism also requires new EVPN considerations for aliasing, mass-withdraw, fast-switchover and [RFC9135] as described in the solution section.

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 [RFC2119].

1.2. Terms and Abbreviations

AC: Attachment Circuit

BUM: Broadcast, Unknown unicast, Multicast

DF: Designated Forwarder

GW: Gateway

L2 Flow: A given flow of a VLAN, represented by (MAC-SA, MAC-DA) or customer MAC

L2GW: Layer-2 Gateway

MAC-IP: EVPN Route-Type 2 with non-zero IP field

G.8032: Ethernet Ring Protection

(M)STP: Multi-Spanning Tree Protocol

REP: Resilient Ethernet Protocol

TCN: Topology Change Notification

2. Requirements

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

* Peering PEs MUST share the same ESI.

* The Ethernet-Segment DF election MUST NOT be performed and forwarding state MUST be dictated by the L2GW protocol. In gateway mode, both PEs are usually in forwarding state. In fact, the access protocol is responsible for operationally setting the forwarding state for each VLAN.

* Split-horizon filtering is NOT needed because L2GW protocol ensures there will never be a loop in the access network. The forwarding between peering PEs MUST also be preserved. In Figure 1, CE1/CE4 device may need reachability with CE2 device. ESI-filtering capability MUST be disabled. The ESI label extended comunity advertised to other peering PEs in the redundancy group MUST NOT be applied it if received.

* ESI label BGP Extended Community MUST support a new multi-homing mode named "Single-Flow-Active" corresponding largely to the single-active behaviour of [RFC7432], applied per L2 flow rather than per VLAN.

* Upon receiving ESI label BGP Extended Community with the single-flow-active load-balancing mode, remote PE MUST:

  - Disable ESI label processing

  - Disable aliasing (at Layer-2 and Layer-3 [RFC9135])

* The Ethernet-Segment procedures in the EVPN core such as Ethernet A-D per ES and per Ethernet A-D per EVI routes advertisement/ withdraw, as well as MAC and MAC+IP advertisement, remains as explained in [RFC7432] and [RFC9135].

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

  - { PE1 active, PE2 backup }

  - { PE2 active, PE1 backup }

  The backup paths so created, operate as in Section 8.4 of [RFC7432] where the backup PE of the redundancy group MAY immediately be selected for forwarding upon detection of a specific subset of failures: Ethernet A-D per ES route withdraw, Active PE loss of reachability (via IGP detection). An Ethernet

A-D per EVI withdraw MUST NOT result in automatic switching to the backup PE as only a subset of the hosts may be changing reachability to the Backup PE, and the remote cannot determine which.

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

3. Solution

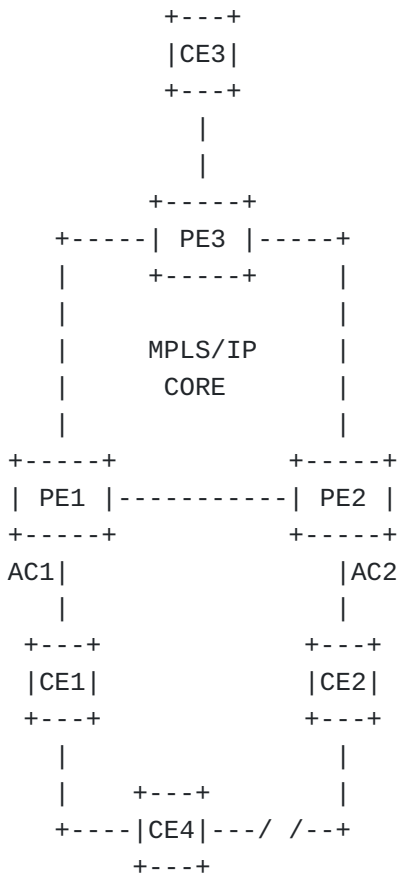

Figure 1: EVPN network with L2 access GW protocols

Figure 1 shows a typical EVPN network with an access network running a L2GW protocol, typically one of the following: G.8032, (M)STP, REP, MPLS-TP ([RFC6378]), 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.

The L2GW protocol is used for loop avoidance. In above example, the loop is broken on the right side of CE4.

In another instantiation, the L2GW protocol used for loop avoidance and splitting per-VLAN L2 flows across peering PEs could be a set of

active/backup pseudowires rooted at CE4. In such a use-case, CE4 decides which pseudowire CE4-PE1 or CE4-PE2 is active or backup, and PE1 and PE2 operate in single-flow-active mode.

3.1. Single-Flow-Active redundancy mode

PE1 and PE2 are peering PEs in a redundancy group, and sharing a same ESI. In the proposed Single-Flow-Active mode, load-balancing at PE1 and PE2 shares similarities with singular aspects of both Single-Active and All-Active. Designated Forwarder election must not compete with the L2GW protocol and must not result in blocked ports or portions of the access may become isolated. Additionally, the reachability between CE1/CE4 and CE2 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 [RFC7432] 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, the L2 flows from PE3 destined to CE1/CE4 transit via edge node PE1 and the L2 flows destined to CE2 transit via edge node PE2. A specific unicast L2 flow never goes to both peering PEs. Therefore, aliasing of [RFC7432] Section 8.4 cannot be performed by PE3. That node operates in a single-active fashion for each of the unicast L2 flows.

The backup path of [RFC7432] Section 8.4 which is also setup for single-active rapid convergence on a per-VLAN basis, is not applicable here. For example, in Figure 1, if a failure happens between CE1 and CE4 the loop-prevention at the right of CE4 is released and:

* L2 flows coming from CE3 behind PE3 destined to CE1 still transit through edge device PE1, and shall not switch to PE2 as a backup path.

* L2 flows destined to CE4 on the other hand, may be backup switched to PE2 transit node.

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.

3.2. Fast-Convergence

3.2.1. 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 REP and G.8032, TCN gets generated in both directions and thus both of the dual-homing PEs receive it. However, with (M)STP, 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 nor 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 the MAC Mobility BGP Extended Community applied to the Ethernet A-D per EVI route, detailed below. It is a slower method but has the advantage of avoiding a dedicated channel between peering PEs.

3.2.2. Propagating L2GW Protocol Events

Peering PE in Single Flow Active mode, upon receiving notification of a protocol convergence-event from access (such as TCN), MUST:

* Perform a local MAC flush on the access-facing interfaces.

* Send an ARP Probe using procedures in Section 7.2 of [RFC9135] for all hosts previously locally attached to the AC in single-flow-active mode.
  The ARP Probes are intended to re-confirm the host is still locally attached, following the convergence-event from the access, or conversely trigger a mobility event from peering PE.
  The probes are sent locally on the specific AC in single-flow-active mode on which the TCN was received, from both peering PEs.

* Advertise Ethernet A-D per EVI route along with the MAC Mobility BGP Extended Community, with incremented sequence number if

previously advertised, 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.

*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 Mobility Extended Community reception. This 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 from dataplane or control-plane (ARP/ND).

*Where an access protocol relies on TCN BPDU propagation to all participant nodes, 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 configured by user.

3.2.3. MAC Flush and Invalidation Procedure

The MAC-Flush procedure described in [RFC7623] is borrowed, and the MAC mobility BGP Extended community is signaled along with the Ethernet A-D per EVI route from a PE in Single-Flow-Active mode.

When MAC Mobility BGP Extended Community is received on the Ethernet A-D per EVI route, it indicates to all remote PEs that all MAC addresses associated with that EVI/ESI are "flushed" i.e. must be unresolved.

Remote PEs, having previously received Ethernet A-D per ES with Single Flow Active indication from an originating PE, treat the MAC Mobility indication to simply invalidate the MAC entries for that originating PE on an EVI/ESI basis, similar to [RFC7432]'s mass-withdraw mechanism.

They remain unresolved until the remote PE receives a route update (or withdraw) for those MAC addresses. Note: the MAC may be re-advertised by the same PE, but also some are expected to have moved to a multi-homing peer, within the same ESI, due to the L2 protocol's action.

The sequence number of the MAC Mobility extended community 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 per EVI/ESI.

3.2.4. MAC Mobility

When an L2 flow moves to PE2 from the PE1 L2GW peer, the MAC mobility sequence number is incremented to signal to remote peers that a 'move' has occurred and the routing tables must be updated to PE2. This is required when an Access Protocol is running where the loop is broken between two CEs in the access and the L2GWs, and the host is no longer reachable from the PE1-side but now from the PE2-side of the access network.

Frequent topology changes in the Layer 2 customer site attached to the EVPN domain via an Ethernet-Segment in Single-Flow-Active redundancy mode could result in false detection of a duplicate-MAC situation described in Section 15.1 of [I-D.ietf-bess-rfc7432bis]. It is RECOMMENDED to tune the configurable M and N parameters of the EVPN MAC Duplication detection in accordance with hold timers of the Layer 2 Control Protocol to prevent false alarms.

3.3. Backwards compatibility

3.3.1. The two-ESI solution

As a reference, an alternative solution which achieves some, but not all, of the requirements exists:

On the PE1 and PE2,

a. A single-homed (different) non-zero ESI, or zero-ESI, is used for each PE;

b. With no remote Ethernet-Segment routes received matching local ESI, each PE will be designated forwarder for all the local VLANs;

c. Each L2GW PE will send Ethernet A-D per ES and per EVI routes for its ESI if non-zero; and

d. When the L2GW PEs receive a MAC-Flush notification (STP TCN, G.8032 mac-flush, LDP MAC withdrawal etc.), they send an update of the Ethernet A-D per EVI route with the MAC Mobility extended community and a higher sequence number, using the procedure outlined in Section 3.2.3.

While this solution is feasible, it is considered to fall short of the requirements listed in Section 2, namely for all aspects meant to achieve fast-convergence.

3.3.2. RFC7432 Remote PE

A PE which receives an Ethernet A-D per ES route with the Single-Flow-Active bit set in the ESI-flags, and which does not support/ understand this bit, SHALL discard the bit and continue operating per [RFC7432] (All-Active). The operator should understand the usage of single-flow-active load-balancing mode else it is highly recommended to use the two-ESI approach as described in Section 3.3.1

The remote PE3 which does not support Single-Flow-Active redundancy mode as described, will ECMP traffic to peering PE1 and PE2 in the example topology above (Figure 1), per [RFC7432], Section 8.4 aliasing and load-balancing rules. PE1 and PE2, which support the Single-Flow-Active redundancy mode MUST setup redirections towards the PE at which the flow is currently active (sub-optimal Layer-2 forwarding and sub-optimal Layer-3 routing).

Thus, while PE3 will ECMP (on average) 50% of the traffic to the incorrect PE using [RFC7432] operation, PE1 and PE2 will handle this gracefully in Single-Flow-Active mode and redirect across peering pair of PEs appropriately.

No extra route or information is required for this. The [RFC7432] and [RFC9135] route advertisements are sufficient.

4. Multihomed site redundancy mode

In order to signal the new EVPN load-balancing mode (single-flow-active), this document claims the following value from the "EVPN ESI Multihoming Attributes" registry's "Multihomed site redundancy mode (RED)" field setup by Section 7.5 of [I-D.ietf-bess-rfc7432bis].

Multihomed site redundancy mode:

RED = 10: A value of 10 means that the multihomed site is operating in Single-Flow-Active redundancy mode.

5. EVPN Inter-subnet Forwarding

EVPN Inter-subnet forwarding procedures in [RFC9135] 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.

However, the use of a same ESI on both Single-Flow-Active L2GW PEs enables:

* the remote PE3 to create two distinct sets of active/backup paths on a per-flow basis towards each of the peering PEs.

*ARP/ND synchronization procedures which are defined for All-Active redundancy in [RFC9135]. In steady-state, on PE2 where a host is not locally-reachable the routing table will reflect PE1 as the destination. However, with ARP/ND synchronization based on a common ESI, the ARP/ND cache may be pre-populated with the local AC as destination for the host, should an AC failure occur on PE1.

These enhancements enable fast-convergence.

Host moves between PE1 and PE2 Single-Flow-Active L2GW peers are handled using the MAC mobility procedures in Section 3.2.4.

6. 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 document has the main advantage to be fully compliant with [RFC7432] and [RFC9135].

7. Security Considerations

The same Security Considerations described in [RFC7432] and [RFC9135] remain valid for this document.

8. Acknowledgements

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

9. IANA Considerations

This document solicits the allocation of the following values from the "EVPN ESI Multihoming Attributes" registry setup in [I-D.ietf-bess-rfc7432bis], and updates the listing of redundancy modes values (RED):

RED	Multihomed site redundancy mode
	00 = All-Active
	01 = Single-Active
	10 = Single-Flow-Active

10. References

10.1. Normative References

[I-D.ietf-bess-rfc7432bis] Sajassi, A., Burdet, L. A., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", Work in Progress, Internet-Draft, draft-ietf-bess-rfc7432bis-06, 5 January 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-rfc7432bis-06>.

[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>.

[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>.

[RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <https://www.rfc-editor.org/info/rfc7623>.

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <https://www.rfc-editor.org/info/rfc9135>.

10.2. Informative References

[RFC6378] Weingarten, Y., Ed., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, Ed., "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, DOI 10.17487/RFC6378, October 2011, <https://www.rfc-editor.org/info/rfc6378>.

Authors' Addresses

Patrice Brissette
Cisco Systems
Ottawa ON
Canada

Email: pbrisset@cisco.com

Ali Sajassi
Cisco Systems
United States of America

Email: sajassi@cisco.com

Luc Andre Burdet (editor)
Cisco Systems
Ottawa ON
Canada

Email: lburdet@cisco.com

Daniel Voyer
Bell Canada
Montreal QC
Canada

Email: daniel.voyer@bell.ca