INTERNET-DRAFT

Intended Status: Proposed Standard

Expires: May 3, 2018

N. Malhotra, Ed.
S. Thoria
A. Sajassi
(Cisco)
A. Lingala
(AT&T)

October 30, 2017

Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing
draft-malhotra-bess-evpn-unequal-lb-00

Abstract

In an EVPN-IRB based network overlay, EVPN LAG enables all-active multi-homing for a host or CE device connected to two or more PEs via a LAG bundle, such that bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs. This document defines extensions to EVPN procedures to optimally handle unequal access bandwidth distribution across a set of multi-homing PEs in order to:

- provide greater flexibility, with respect to adding or removing individual PE-CE links within the access LAG
- handle PE-CE LAG member link failures that can result in unequal PE-CE access bandwidth across a set of multi-homing PEs

Status of this Memo

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

Table of Contents

1 Introduction 3
1.1 PE CE Link Provisioning 4
1.2 PE CE Link Failures 5
1.3 Design Requirement 6
1.1 Terminology 6
2. Solution Overview 6
3. Weighted Unicast Traffic Load-balancing 7
3.1 LOCAL PE Behavior 7
3.2 REMOTE PE Behavior 7
4. Weighted BUM Traffic Load-Sharing 8
5. Routed EVPN Overlay 8
6. EVPN-IRB Multi-homing with non-EVPN routing 9
7. References 10
7.1 Normative References 10
7.2 Informative References 10
8. Acknowledgements 10
Authors' Addresses 10

1 Introduction

In an EVPN-IRB based network overlay, with access an access CE multi-homed via a LAG interface, bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs:

- ECMP Load-balancing for bridged unicast traffic is enabled via aliasing and mass-withdraw procedures detailed in RFC 7432.
- ECMP Load-balancing for routed unicast traffic is enabled via existing L3 ECMP mechanisms.
- Load-sharing of bridged BUM traffic on local ports is enabled via EVPN DF election procedure detailed in RFC 7432

All of the above load-balancing and DF election procedures implicitly assume equal bandwidth distribution between the CE and the set of multi-homing PEs. Essentially, with this assumption of equal "access" bandwidth distribution across all PEs, ALL remote traffic is equally load balanced across the multi-homing PEs. This assumption of equal access bandwidth distribution can be restrictive with respect to adding / removing links in a multi-homed LAG interface and may also be easily broken on individual link failures. A solution to handle unequal access bandwidth distribution across a set of multi-homing EVPN PEs is proposed in this document. Primary motivation behind this proposal is to enable greater flexibility with respect to adding / removing member PE-CE links, as needed and optimally handle PE-CE link failures.

1.1 PE CE Link Provisioning

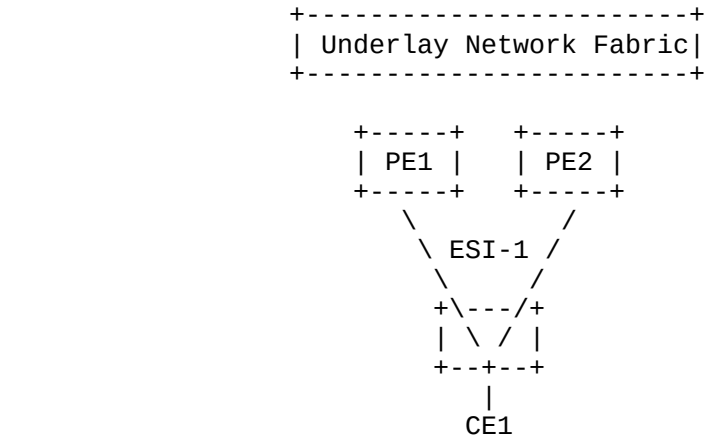


Figure 1

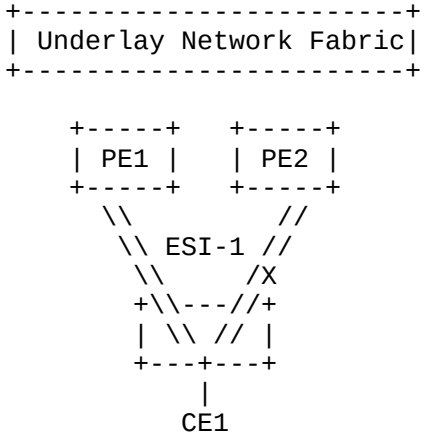
Consider a CE1 that is dual-homed to PE1 and PE2 via EVPN-LAG with single member links of equal bandwidth to each PE (aka, equal access band-width distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it MUST add a link to both PE1 and PE2 in order to maintain equal access bandwidth distribution and inter-work with EVPN ECMP load-balancing. In other words, for a dual-homed CE, total number of CE links must be provisioned in multiples of 2 (2, 4, 6, and so on). For a triple-homed CE, number of CE links must be provisioned in multiples of three (3, 6, 9, and so on). To generalize, for a CE that is multi-homed to "n" PEs, number of PE-CE physical links provisioned must be an integral multiple of "n". This is restrictive in case of dual-homing and very quickly becomes prohibitive in case of multi-homing.

Instead, a provider may wish to increase PE-CE bandwidth OR number of links in ANY link increments. As an example, for CE1 dual-homed to PE1 and PE2 in all-active mode, provider may wish to add a third link to ONLY PE1 to increase total band-width for this CE by 50%, rather than being required to increase access bandwidth by 100% by adding a link to each of the two PEs. While existing EVPN based all-active load-balancing procedures do not necessarily preclude such asymmetric access bandwidth distribution among the PEs providing redundancy, it may result in unexpected traffic loss due to congestion in the access interface towards CE. This traffic loss is due to the fact that PE1 and PE2 will continue to attract equal amount of CE1 destined traffic from remote PEs, even when PE2 only has half the bandwidth to CE1 as PE1. This may lead to congestion and traffic loss on the PE2-CE1

link. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner.

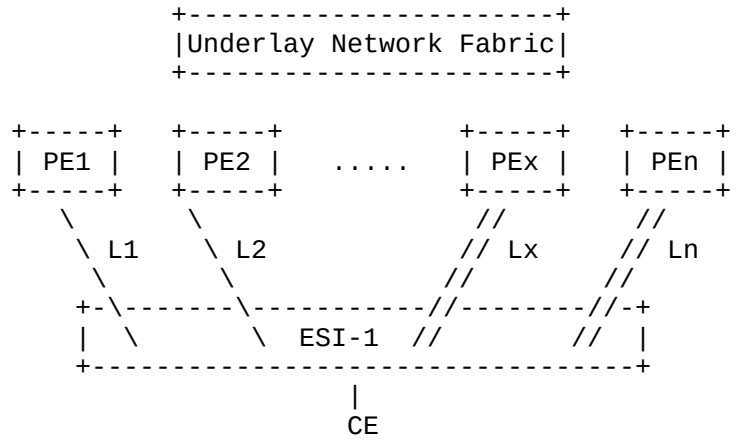
1.2 PE CE Link Failures

More importantly, unequal PE-CE bandwidth distribution described above may occur during regular operation following a link failure, even when PE-CE links were provisioned to provide equal bandwidth distribution across multi-homing PEs.



Consider a CE1 that is multi-homed to PE1 and PE2 via a link bundle with two member links to each PE. On a PE2-CE1 physical link failure, link bundle represented by ESI-1 on PE2 stays up, however, it's bandwidth is cut in half. With the existing ECMP procedures, both PE1 and PE2 will continue to attract equal amount of traffic from remote PEs, even when PE1 has double the bandwidth to CE1. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts MUST also be load-balanced across PE1 and PE2 in 2:1 manner to avoid unexpected congestion and traffic loss on PE2-CE1 links within the LAG.

1.3 Design Requirement



To generalize, if total link band-width to a CE is distributed across "n" multi-homing PEs, with Lx being the number of links / bandwidth to PEx, traffic from remote PEs to this CE MUST be load-balanced unequally across [PE1, PE2, ..., PEn] such that, the proportion of unicast and BUM flows destined for CE that are serviced by PEx is:

Lx / [L1+L2+.....+Ln]

Solution proposed below includes extensions to EVPN procedures to achieve the above.

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].

"LOCAL PE" in the context of an ESI refers to a provider edge switch OR router that physically hosts the ESI.

"REMOTE PE" in the context of an ESI refers to a provider edge switch OR router in an EVPN overlay, who’s overlay reachability to the ESI is via the LOCAL PE.

2. Solution Overview

In order to achieve weighted load balancing for overlay unicast

traffic, EVPN per-ESI EAD (Route Type 1) is leveraged to signal the ESI bandwidth to remote PEs. Using per-ESI EAD route to signal the ESI bandwidth provides a mechanism to be able to react to changes in access bandwidth in a service and host independent manner. Remote PEs computing the MAC path-lists based on global and aliasing EAD routes now have the ability to computed weighted load-balancing based on the ESI access bandwidth received from each PE that the ESI is multi-homed to. If per-ESI EAD route is also leveraged for IP path-list computation, as per [EVPN-IP-ALIASING], it would also provide a method to do weighted load-balancing for IP routed traffic.

In order to achieve weighted load-balancing of overlay BUM traffic, EVPN ES route (Route Type 4) is leveraged to signal the ESI bandwidth to PEs within an ESI's redundancy group to influence per-service DF election. PEs in an ESI redundancy group now have the ability to do per-service DF election in a manner that is proportionate to their relative ESI bandwidth.

Procedures to accomplish this are described in greater detail next.

3. Weighted Unicast Traffic Load-balancing

3.1 LOCAL PE Behavior

A PE that is part of an ESI's redundancy group would advertise a additional "link bandwidth" EXT-COMM attribute with per-ESI EAD route (EVPN Route Type 1), that represents total band-width of PE's physical links in an ESI. BGP link bandwidth EXT-COMM defined in [BGP-LINK-BW] would be re-used for this purpose.

3.2 REMOTE PE Behavior

A receiving PE should use per-ESI link band-width attribute received from each PE to compute a relative weight for each remote PE, per-ESI, as shown below.

if,

L(x,y) : link band-width advertised by PE-x for ESI-y

W(x,y) : normalized weight assigned to PE-x for ESI-y

H(y) : Highest Common Factor (HCF) of [L(1,y), L(2,y), ..., L(n,y)]

then, the normalized weight assigned to PE-x for ESI-y may be computed as follows:

W(x,y) = L(x,y) / H(y)

For a MAC+IP route (EVPN Route Type 2) received with ESI-y, receiving PE MUST compute MAC and IP forwarding path-list weighted by the above normalized weights.

As an example, for a CE dual-homed to PE-1, PE-2, PE-3 via 2, 1, and 1 GE physical links respectively, as part of a link bundle represented by ESI-10:

L(1, 10) = 2000 Mbps

L(2, 10) = 1000 Mbps

L(3, 10) = 1000 Mbps

H(10) = 1000

Normalized weights assigned to each PE for ESI-10 are as follows:

W(1, 10) = 2000 / 1000 = 2.

W(2, 10) = 1000 / 1000 = 1.

W(3, 10) = 1000 / 1000 = 1.

For a remote MAC+IP host route received with ESI-10, forwarding load-balancing path-list must now be computed as: [PE-1, PE-1, PE-2, PE-3] instead of [PE-1, PE-2, PE-3]. This now results in load-balancing of all traffic destined for ESI-10 across the three multi-homing PEs in proportion to ESI-10 band-width at each PE.

Above weighted path-list computation MUST only be done for an ESI, IF a link bandwidth attribute is received from ALL of the PE’s advertising reachability to that ESI via per-ESI EAD Route Type 1. In the event that link bandwidth attribute is not received from one or more PEs, forwarding path-list would be computed using regular ECMP semantics.

4. Weighted BUM Traffic Load-Sharing

Load sharing of per-service DF role, weighted by link-bandwidth is currently under discussion and needs to be reconciled with [EVPN-PREF-DF]. This will closed in the next revision of this draft.

5. Routed EVPN Overlay

An additional use case is possible, such that traffic to an end host

in the overlay is always IP routed. In a purely routed overlay such as this:

- o A host MAC is never advertised in EVPN overlay control plane
- o Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI
- o An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged
- o Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC 7814] for more details.

Weighted multi-path procedure described in this document may be used together with procedures described in [EVPN-IP-ALIASING] for this use case. per-ES EAD route advertised with Layer 3 VRF RTs would be used to signal ES link bandwidth attribute instead of the per-ES EAD route with Layer 2 VRF RTs. All other procedures described earlier in this document would as is.

6. EVPN-IRB Multi-homing with non-EVPN routing

EVPN-LAG based multi-homing on an IRB gateway may also be deployed together with non-EVPN routing, such as global routing or an L3VPN routing control plane. Key property that differentiates this set of use cases from EVPN IRB use cases discussed earlier is that EVPN control plane is used only to enable LAG interface based multi-homing and NOT as an overlay VPN control plane. EVPN control plane in this case enables:

- o DF election via EVPN RT-4 based procedures described in [RFC7432]
- o LOCAL MAC sync across multi-homing PEs via EVPN RT-2
- o LOCAL ARP and ND sync across multi-homing PEs via EVPN RT-2

Applicability of weighted ECMP procedures proposed in this document to these set of use cases are still under discussion and will be addressed in subsequent revisions.

7. References

7.1 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>.

[BGP-LINK-BW] Mohapatra, P., Fernando, R., "BGP Link Bandwidth Extended Community", January 2013, <https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-06>.

[EVPN-IP-ALIASING] Sajassi, A., Badoni, G., "L3 Aliasing and Mass Withdrawal Support for EVPN", July 2017, <https://tools.ietf.org/html/draft-sajassi-bess-evpn-ip-aliasing-00>.

[EVPN-PREF-DF-ELECT] Rabadan, J., et al., "Preference-based EVPN DF Election", June 2017, <https://www.ietf.org/id/draft-ietf-bess-evpn-pref-df-00.txt>.

7.2 Informative References

8. Acknowledgements

Authors' Addresses

Neeraj Malhotra
Cisco
Email: nmalhotr@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Avinash Lingala
AT&T
Email: ar977m@att.com