BESS Working Group
Internet-Draft
Intended Status: Proposed Standard

N. Malhotra, Ed.
Arrcus

A. Sajassi
S. Thoria
Cisco

J. Rabadan
Nokia

J. Drake
Juniper

A. Lingala
AT&T

Expires: Jan 23, 2020

July 22, 2019

Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing
draft-ietf-bess-evpn-unequal-lb-02

Abstract

In an EVPN-IRB based network overlay, EVPN all-active multi-homing enables multi-homing for a 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

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.

Copyright and License Notice

Table of Contents

1 Introduction

In an EVPN-IRB based network overlay, with a CE multi-homed via a EVPN all-active multi-homing, 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 to optimally handle PE-CE link failures.

1.1 PE CE Link Provisioning

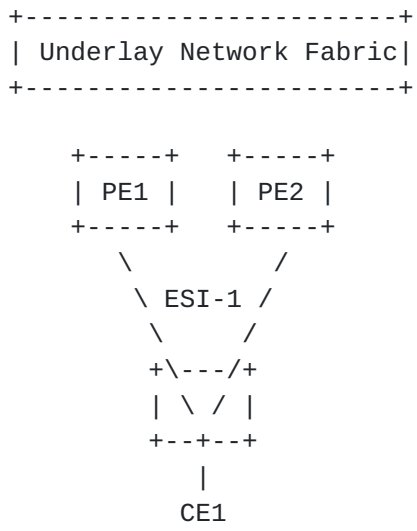


Figure 1

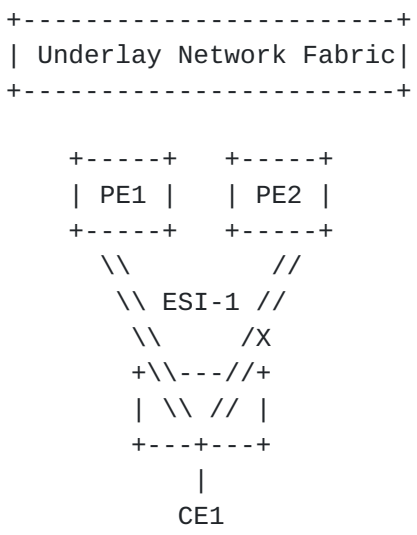
Consider a CE1 that is dual-homed to PE1 and PE2 via EVPN all-active multi-homing with single member links of equal bandwidth to each PE (aka, equal access bandwidth distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it MUST add a link to both PE1 and PE2 in order 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 bandwidth 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.

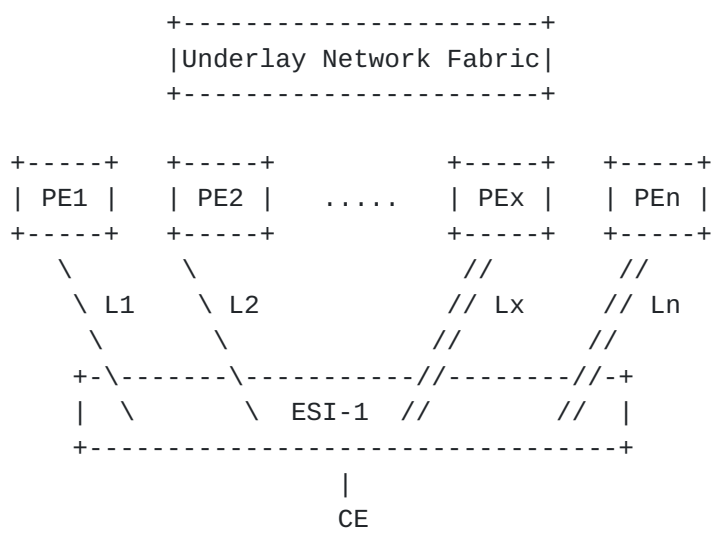
1.2 PE CE Link Failures

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 an Ethernet Segment ESI-1 on PE2 stays up, however, it's bandwidth is cut in half. With 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 bandwidth 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, fraction of total 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.

1.4 Terminology

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 BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

"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, Ethernet A-D per-ES route (EVPN Route Type 1) is leveraged to signal the Ethernet Segment bandwidth to remote PEs. Using Ethernet A-D per-ES route to signal the Ethernet Segment 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 Ethernet A-D routes now have the ability to setup weighted load-balancing path-lists based on the ESI access bandwidth received from each PE that the ESI is multi-homed to. If Ethernet A-D per-ES route is also leveraged for IP path-list computation, as per [EVPN-IP-ALIASING], it also provides 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 service carving in proportion to each PE's 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 Ethernet Segment's redundancy group would advertise a additional "link bandwidth" EXT-COMM attribute with Ethernet A-D per-ES route (EVPN Route Type 1), that represents total bandwidth of PE's physical links in an Ethernet Segment. BGP link bandwidth EXT-COMM defined in [BGP-LINK-BW] is re-used for this purpose.

3.1 Link Bandwidth Extended Community

Link bandwidth extended community described in [BGP-LINK-BW] for layer 3 VPNs is re-used here to signal local ES link bandwidth to remote PEs. link-bandwidth extended community is however defined in [BGP-LINK-BW] as optional non-transitive. In inter-AS scenarios, link-bandwidth may need to be signaled to an eBGP neighbor along with next-hop unchanged. It is work in progress with authors of [BGP-LINK-BW] to allow for this attribute to be used as transitive in inter-AS scenarios.

3.2 REMOTE PE Behavior

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

if,

L(x,y) : link bandwidth 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 bandwidth 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 Ethernet A-D per-ES 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

Optionally, load sharing of per-service DF role, weighted by individual PE's link-bandwidth share within a multi-homed ES may also be achieved.

In order to do that, a new DF Election Capability [RFC8584] called "BW" (Bandwidth Weighted DF Election) is defined. BW may be used along with some DF Election Types, as described in the following sections.

4.1 The BW Capability in the DF Election Extended Community

[RFC8584] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests a bit in the DF Election extended community Bitmap:

Bit 28: BW (Bandwidth Weighted DF Election)

ES routes advertised with the BW bit set will indicate the desire of the advertising PE to consider the link-bandwidth in the DF Election algorithm defined by the value in the "DF Type".

As per [RFC8584], all the PEs in the ES MUST advertise the same Capabilities and DF Type, otherwise the PEs will fall back to Default [RFC7432] DF Election procedure.

The BW Capability MAY be advertised with the following DF Types:

- Type 0: Default DF Election algorithm, as in [RFC7432]
- Type 1: HRW algorithm, as in [RFC8584]
- Type 2: Preference algorithm, as in [EVPN-DF-PREF]
- Type 4: HRW per-multicast flow DF Election, as in [EVPN-PER-MCAST-FLOW-DF]

The following sections describe how the DF Election procedures are modified for the above DF Types when the BW Capability is used.

4.2 BW Capability and Default DF Election algorithm

When all the PEs in the Ethernet Segment (ES) agree to use the BW Capability with DF Type 0, the Default DF Election procedure is modified as follows:

- Each PE advertises a "Link Bandwidth" EXT-COMM attribute along with the ES route to signal the PE-CE link bandwidth (LBW) for the ES.
- A receiving PE MUST use the ES link bandwidth attribute received from each PE to compute a relative weight for each remote PE.
- The DF Election procedure MUST now use this weighted list of PEs to compute the per-VLAN Designated Forwarder, such that the DF role is distributed in proportion to this normalized weight.

Considering the same example as in Section 3, the candidate PE list for DF election is:

[PE-1, PE-1, PE-2, PE-3].

The DF for a given VLAN-a on ES-10 is now computed as (VLAN-a % 4). This would result in the DF role being distributed across PE1, PE2, and PE3 in portion to each PE's normalized weight for ES-10.

4.3 BW Capability and HRW DF Election algorithm (Type 1 and 4)

[RFC8584] introduces Highest Random Weight (HRW) algorithm (DF Type 1) for DF election in order to solve potential DF election skew depending on Ethernet tag space distribution. [EVPN-PER-MCAST-FLOW-DF] further extends HRW algorithm for per-multicast flow based hash computations (DF Type 4). This section describes extensions to HRW Algorithm for EVPN DF Election specified in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] in order to achieve DF election distribution that is weighted by link bandwidth.

4.3.1 BW Increment

A new variable called "bandwidth increment" is computed for each [PE, ES] advertising the ES link bandwidth attribute as follows:

In the context of an ES,

L(i) = Link bandwidth advertised by PE(i) for this ES

L(min) = lowest link bandwidth advertised across all PEs for this ES

Bandwidth increment, "b(i)" for a given PE(i) advertising a link

bandwidth of L(i) is defined as an integer value computed as:

b(i) = L(i) / L(min)

As an example,

with PE(1) = 10, PE(2) = 10, PE(3) = 20

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 2

with PE(1) = 10, PE(2) = 10, PE(3) = 10

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 1

Note that the bandwidth increment must always be an integer, including, in an unlikely scenario of a PE's link bandwidth not being an exact multiple of L(min). If it computes to a non-integer value (including as a result of link failure), it MUST be rounded down to an integer.

4.3.2 HRW Hash Computations with BW Increment

HRW algorithm as described in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] compute a random hash value (referred to as affinity here) for each PE(i), where, (0 < i <= N), PE(i) is the PE at ordinal i, and Address(i) is the IP address of PE at ordinal i.

For 'N' PEs sharing an Ethernet segment, this results in 'N' candidate hash computations. PE that has the highest hash value is selected as the DF.

Affinity computation for each PE(i) is extended to be computed one per-bandwidth increment associated with PE(i) instead of a single affinity computation per PE(i).

PE(i) with b(i) = j, results in j affinity computations:

affinity(i, x), where 1 < x <= j

This essentially results in number of candidate HRW hash computations for each PE that is directly proportional to that PE's relative bandwidth within an ES and hence gives PE(i) a probability of being DF in proportion to it's relative bandwidth within an ES.

As an example, consider an ES that is multi-homed to two PEs, PE1 and PE2, with equal bandwidth distribution across PE1 and PE2. This would result in a total of two candidate hash computations:

affinity(PE1, 1)

affinity(PE2, 1)

Now, consider a scenario with PE1's link bandwidth as 2x that of PE2. This would result in a total of three candidate hash computations to be used for DF election:

affinity(PE1, 1)

affinity(PE1, 2)

affinity(PE2, 1)

which would give PE1 2/3 probability of getting elected as a DF, in proportion to its relative bandwidth in the ES.

Depending on the chosen HRW hash function, affinity function MUST be extended to include bandwidth increment in the computation.

For e.g.,

affinity function specified in [EVPN-PER-MCAST-FLOW-DF] MAY be extended as follows to incorporate bandwidth increment j:

affinity(S,G,V, ESI, Address(i,j)) =
(1103515245.((1103515245.Address(i).j + 12345) XOR
D(S,G,V,ESI))+12345) (mod 2^31)

affinity or random function specified in [RFC8584] MAY be extended as follows to incorporate bandwidth increment j:

affinity(v, Es, Address(i,j)) = (1103515245((1103515245.Address(i).j + 12345) XOR D(v,Es))+12345)(mod 2^31)

4.3.3 Cost-Benefit Tradeoff on Link Failures

While incorporating link bandwidth into the DF election process provides optimal BUM traffic distribution across the ES links, it also implies that affinity values for a given PE are re-computed, and DF elections are re-adjusted on changes to that PE's bandwidth increment that might result from link failures or link additions. If the operator does not wish to have this level of churn in their DF

election, then they should not advertise the BW capability. Not advertising BW capability may result in less than optimal BUM traffic distribution while still retaining the ability to allow a remote ingress PE to do weighted ECMP for its unicast traffic to a set of multi-homed PEs, as described in section 3.2.

Same also applies to use of BW capability with service carving (DF Type 0), as specified in section 4.2.

4.4 BW Capability and Weighted HRW DF Election algorithm (Type TBD)

Use of BW capability together with HRW DF election algorithm described in the previous section has a few limitations:

- While in most scenarios a change in BW for a given PE results in re-assigment of DF roles from or to that PE, in certain scenarios, a change in PE BW can result in complete re-assignment of DF roles.
- If BW advertised from a set of PEs does not have a good least common multiple, the BW set may result in a high BW increment for each PE, and hence, may result in higher order of complexity.

[WEIGHTED-HRW] document describes an alternate DF election algorithm that uses a weighted score function that is minimally disruptive such that it minimizes the probability of complete re-assignment of DF roles in a BW change scenario. It also does not require multiple BW increment based computations.

Instead of computing BW increment and an HRW hash for each [PE, BW increment], a single weighted score is computed for each PE using the proposed score function with absolute BW advertised by each PE as its weight value.

As described in section 4 of [WEIGHTED-HRW], a HRW hash computation for each PE is converted to a weighted score as follows:

Score(Oi, Sj) = -wi/log(Hash(Oi, Sj)/Hmax); where Hmax is the maximum hash value.

Oi is object being assigned, for e.g., a vlan-id in this case;

Sj is the server, for e.g., a PE IP address in this case;

wi is the weight, for e.g., BW capability in this case;

Object Oi is assigned to server Si with the highest score.

4.5 BW Capability and Preference DF Election algorithm

This section applies to ES'es where all the PEs in the ES agree use the BW Capability with DF Type 2. The BW Capability modifies the Preference DF Election procedure [EVPN-DF-PREF], by adding the LBW value as a tie-breaker as follows:

- Section 4.1, bullet (f) in [EVPN-DF-PREF] now considers the LBW value:

  f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP bit, the LBW value and the lowest IP PE in that order. For instance:

  - If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=1, LBW=2000] in PE2, PE2 would be elected due to the DP bit.
  - If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=0, LBW=2000] in PE2, PE2 would be elected due to a higher LBW, even if PE1's IP address is lower.
  - The LBW exchanged value has no impact on the Non-Revertive option described in [EVPN-DF-PREF].

5. Real-time Available Bandwidth

PE-CE link bandwidth availability may sometimes vary in real-time disproportionately across PE_CE links within a multi-homed ESI due to various factors such as flow based hashing combined with fat flows and unbalanced hashing. Reacting to real-time available bandwidth is at this time outside the scope of this document. Procedures described in this document are strictly based on static link bandwidth parameter.

6. 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. Ethernet A-D per-ES route advertised with Layer 3 VRF RTs would be used to signal ES link bandwidth attribute instead of the Ethernet A-D per-ES route with Layer 2 VRF RTs. All other procedures described earlier in this document would apply as is.

If [EVPN-IP-ALIASING] is not used for routed fast convergence, link bandwidth attribute may still be advertised with IP routes (RT-5) to achieve PE-CE link bandwidth based load-balancing as described in this document. In the absence of [EVPN-IP-ALIASING], re-balancing of traffic following changes in PE-CE link bandwidth will require all IP routes from that CE to be re-advertised in a prefix dependent manner.

7. 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:

- DF election via EVPN RT-4 based procedures described in [RFC7432]
- LOCAL MAC sync across multi-homing PEs via EVPN RT-2
- 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 is an area of further consideration.

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", March 2018, <https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-07>.

[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-DF-PREF] Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., and S. Mohanty, "Preference-based EVPN DF Election", internet-draft ietf-bess-evpn-pref-df-01.txt, April 2018.

[EVPN-PER-MCAST-FLOW-DF] Sajassi, et al., "Per multicast flow Designated Forwarder Election for EVPN", March 2018, <https://tools.ietf.org/html/draft-sajassi-bess-evpn-per-mcast-flow-df-election-00>.

[RFC8584] Rabadan, Mohanty, et al., "Framework for Ethernet VPN Designated Forwarder Election Extensibility", April 2019, <https://tools.ietf.org/html/rfc8584>.

[WEIGHTED-HRW] Mohanty, et al., "Weighted HRW and its applications", Sept. 2019, <https://tools.ietf.org/html/draft-mohanty-bess-weighted-hrw-00>.

[RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", March 1997, <https://tools.ietf.org/html/rfc2119>.

[RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", May 2017, <https://tools.ietf.org/html/rfc8174>.

7.2 Informative References

8. Acknowledgements

Authors would like to thank Satya Mohanty for valuable review and inputs with respect to HRW and weighted HRW algorithm refinements proposed in this document.

9. Contributors

Satya Ranjan Mohanty
Cisco
Email: satyamoh@cisco.com

Authors' Addresses

Neeraj Malhotra, Editor.
Arrcus
Email: neeraj.ietf@gmail.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Jorge Rabadan
Nokia
Email: jorge.rabadan@nokia.com

John Drake
Juniper
EMail: jdrake@juniper.net

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

Samir Thoria
Cisco
Email: sthoria@cisco.com