BESS WorkGroup
Internet-Draft
Updates: RFC8584 (if approved)
Intended status: Standards Track
Expires: 6 July 2026

N. Malhotra, Ed.
A. Sajassi
Cisco Systems
J. Rabadan
Nokia
J. Drake
Independent
A. Lingala
ATT
S. Thoria
Cisco Systems

2 January 2026

Weighted Multi-Path Procedures for EVPN Multi-Homing
draft-ietf-bess-evpn-unequal-lb-30

Abstract

Ethernet VPN (EVPN) provides all-active multi-homing for Customer Equipment (CE) devices connected to multiple Provider Edge (PE) devices, enabling equal cost load balancing of both bridged and routed traffic across the set of multi-homing PEs. However, existing procedures implicitly assume equal access bandwidth distribution among the multi-homing PEs, which can constrain link additions or removals and may not handle unequal PE-CE link bandwidth following link failures. This document specifies extensions to EVPN procedures to support weighted multi-pathing in proportion to PE-CE link bandwidth or operator-defined weights, thereby providing greater flexibility and resilience in multi-homing deployments. The extensions include signaling mechanisms to distribute traffic across egress PEs based on relative bandwidth or weight, and enhancements to Broadcast, Unknown Unicast, and Multicast (BUM) designated forwarder (DF) election to achieve weighted DF distribution across the multi-homing PE set. The document updates RFC 8584 and related EVPN DF election extensions (i.e. draft-ietf-bess-evpn-per-mcast-flow-df-election and draft-ietf-bess-evpn-pref-df) to enable weighted load balancing across different DF election algorithms.

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/.

Copyright Notice

Table of Contents

1. Introduction

In an Ethernet VPN (EVPN) Integrated Routing and Bridging (IRB) overlay network, as described in [RFC9135], a Customer Edge (CE) device may be multi-homed to multiple Provider Edge (PE) devices using EVPN all-active multi-homing. In such deployments, both bridged and routed traffic from ingress PEs can be equally load balanced across the set of egress PEs, as follows:

* Equal-Cost Multipath (ECMP) load balancing for bridged unicast traffic is provided through the aliasing and mass-withdraw procedures defined in [RFC7432].

* ECMP load balancing for routed unicast traffic is provided through existing Layer 3 ECMP mechanisms.

* Load sharing of bridged Broadcast, Unknown Unicast, and Multicast (BUM) traffic on local ports is provided through the EVPN Designated Forwarder (DF) election procedures defined in [RFC7432].

These load-balancing and DF election procedures implicitly assume an equal distribution of access bandwidth between the CE and each of the egress PEs. Under this assumption, remote traffic is evenly distributed across all egress PEs. However, this assumption can be restrictive in operational environments, particularly when adding or

removing member links in a multi-homed Link Aggregation Group (LAG), and can be violated in the presence of individual PE窶電E link failures.

This document specifies procedures to support unequal access bandwidth distribution across a set of egress PEs. The objective is to provide greater operational flexibility when modifying PE窶電E connectivity and to enable more efficient traffic distribution following PE窶電E link failures.

1.1. PE-CE Link Provisioning

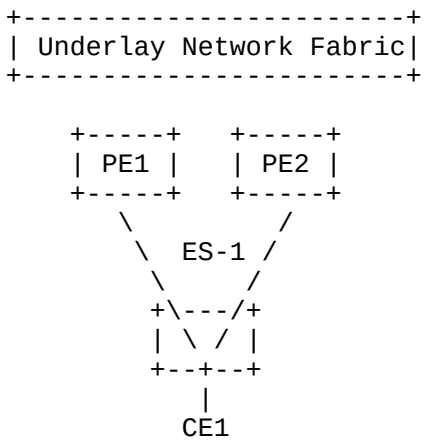

Figure 1

Figure 1 illustrates an EVPN all-active multi-homing topology in which CE1 is dual-homed to egress PE1 and egress PE2. Each PE is connected to CE1 by a single member link of equal bandwidth, resulting in an equal access bandwidth distribution across the two PEs. As described in [RFC7432], [RFC8584], and [RFC9135], existing EVPN procedures enable equal-cost load balancing of both bridged and routed traffic across the set of egress PEs under this assumption.

When equal access bandwidth distribution is maintained, increasing the aggregate PE窶電E bandwidth requires symmetric provisioning of additional links to each egress PE. Consequently, for a dual-homed CE, the total number of PE窶電E links must be provisioned in multiples of two (e.g., 2, 4, 6). Similarly, for a CE multi-homed to "n" PEs, the total number of PE窶電E physical links must be an integer multiple of "n". While this approach is feasible for small values of "n", it can become operationally restrictive as the number of multi-homing PEs increases.

Operationally, a provider may wish to increase PE窶ｦE bandwidth in arbitrary increments. For example, in the topology shown in Figure 1, a provider may choose to add an additional link to PE1 only, thereby increasing the aggregate access bandwidth to CE1 without symmetrically augmenting connectivity to PE2. Although existing EVPN all-active procedures do not explicitly prohibit such asymmetric access bandwidth distributions, they assume equal-cost forwarding toward the multi-homed CE. As a result, remote PEs may continue to distribute traffic approximately evenly across PE1 and PE2, despite the reduced access bandwidth toward CE1 via PE2.

This mismatch between traffic distribution and available access bandwidth may lead to congestion and packet loss on the PE2窶ｦE1 link. To avoid such conditions, traffic destined for a multi-homed CE needs to be distributed across egress PEs in proportion to their respective access bandwidths.

1.2. PE-CE Link Failures

Unequal PE窶ｦE access bandwidth distribution may also arise during normal operation following a PE窶ｦE link failure, even when PE窶ｦE links are initially provisioned to provide equal bandwidth distribution across the multi-homing PEs.

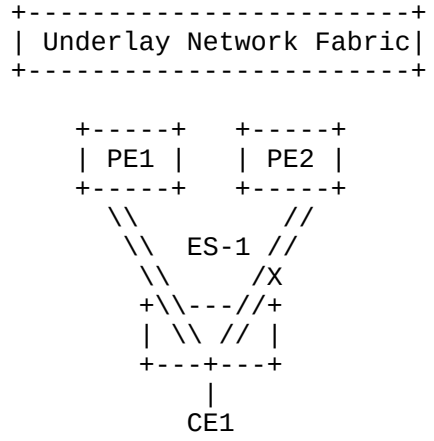

Figure 2

Figure 2 illustrates a scenario in which CE1 is multi-homed to egress PE1 and egress PE2 using a LAG, with two member links connecting CE1 to each PE. Upon failure of a single PE2窶ｦE1 physical link, the Ethernet Segment (ES-1) on PE2 remains operational; however, the effective access bandwidth toward CE1 via PE2 is reduced by half.

With existing EVPN ECMP procedures, both PE1 and PE2 may continue to attract approximately equal amounts of traffic from remote PEs, despite PE1 having twice the available access bandwidth toward CE1. In this case, where the effective access bandwidth distribution between PE1 and PE2 is 2:1, traffic destined for CE1 should be distributed across the two PEs in the same proportion in order to avoid congestion and packet loss on the PE2宴鼎E1 links within the LAG.

As a mitigation, some deployments use the LAG minimum-link feature to force the LAG interface down upon member link failure. While this approach avoids asymmetric bandwidth conditions, it also results in unnecessary loss of available bandwidth and is therefore operationally suboptimal.

1.3. Design Requirement

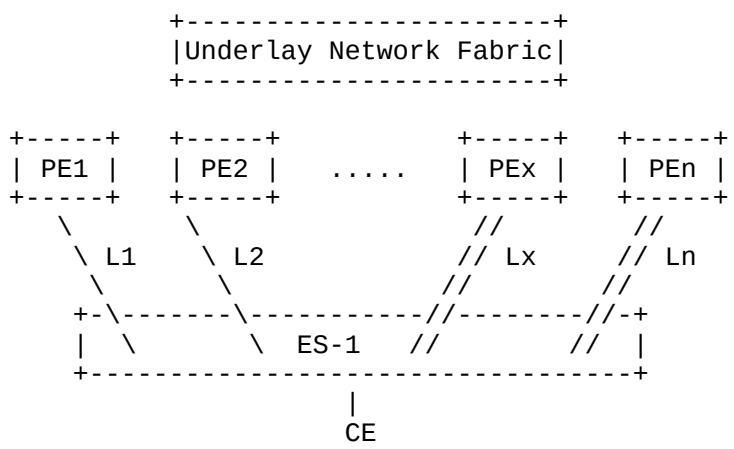

Figure 3

To generalize, consider a CE for which the total access bandwidth is distributed across "n" egress PEs, where "Lx" represents the aggregate bandwidth between the CE and egress PEx. Traffic from ingress PEs destined for the CE needs to be load balanced across the set of egress PEs [PE1, PE2, 宴ヲ, PEn] in proportion to their respective access bandwidths. Specifically, the fraction of unicast and Broadcast, Unknown Unicast, and Multicast (BUM) traffic serviced by egress PEx SHOULD be:

Lx / (L1 + L2 + ... + Ln)

Figure 3 illustrates a scenario in which egress PE1 through PEn are connected to a multi-homed Ethernet Segment. However, the requirement described in this section is not limited to physical

Ethernet Segments. It equally applies to virtual Ethernet Segments (vES) and to multi-homed subnets advertised using EVPN IP Prefix routes.

The solution specified in this document defines extensions to existing EVPN procedures to satisfy the above requirement. The following assumptions apply to the procedures described herein:

* For procedures related to bridged unicast and BUM traffic, EVPN all-active multi-homing is assumed.

* Procedures related to bridged unicast and BUM traffic apply to both aliasing and non-aliasing modes, as defined in [RFC7432].

2. Requirements Language and 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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

* BW: BandWidth

* LAG: Link Aggregation Group

* ES: Ethernet Segment

* ESI: Ethernet Segment ID

* vES: Virtual Ethernet Segment

* EVI: Ethernet virtual Instance, this is a mac-vrf

* RT-1: EVPN Route Type 1 as defined in [RFC7432]

* RT-2: EVPN Route Type 2 as defined in [RFC7432]

* RT-5: EVPN Route Type 5 as defined in [RFC7432]

* PE: Provider Edge Router

* Path-List: A forwarding object used to load-balance routed or bridged traffic across multiple forwarding paths.

* Access Bandwidth: Bandwidth of PE-CE links in an Ethernet Segment

* Egress PE: In the context of an Ethernet Segment or a route, this is the PE that advertises a locally attached Ethernet Segment RT-1, or a locally attached host or prefix route (RT-2, RT-5)

* Ingress PE: In the context of an Ethernet Segment or a route, this is the receiving PE that learns remote Ethernet Segment RT-1 and/or host and prefix routes (RT-2, RT-5) from the Egress PE

* IMET: Inclusive Multicast Route

* DF: Designated Forwarder

* BDF: Backup Designated Forwarder

* DCI: Data Center Interconnect Router

3. Solution Overview

To enable weighted load balancing of overlay unicast traffic toward an ES or vES, this document leverages the Ethernet A-D per ES route (EVPN Route Type 1) to signal an ES weight to ingress PEs. Signaling the ES weight via the Ethernet A-D per ES route provides a service- and host-independent mechanism that dynamically reflects changes in access bandwidth or the number of PE寰鼎E links. Ingress PEs that compute MAC path lists based on global and aliasing Ethernet A-D routes can therefore construct weighted load balancing path lists proportional to the relative access bandwidth advertised by each egress PE.

Weighted load balancing of overlay BUM traffic is achieved by using the EVPN ES route (EVPN Route Type 4) to signal the ES weight to egress PEs within the ES redundancy group. This information is used to influence per-service DF election, allowing egress PEs to perform service carving in proportion to their relative ES weights.

Unequal load balancing toward multi-homed subnets is supported by advertising the corresponding weight together with the EVPN IP Prefix routes associated with the subnet.

The detailed procedures for these mechanisms are specified in the following sections.

4. EVPN Link Bandwidth Extended Community

This document defines a new EVPN Link Bandwidth Extended Community to support the solution described herein.

* The extended community is of Type 0x06 (EVPN Extended Community Sub-Types).

* IANA has assigned Sub-Type value 0x10 for the EVPN Link Bandwidth Extended Community.

* The EVPN Link Bandwidth Extended Community is defined as transitive.

4.1. Encoding and Usage

The EVPN Link Bandwidth Extended Community is used to advertise the aggregate access bandwidth of all physical links between an egress PE and an Ethernet Segment. The value is expressed as an unsigned integer representing megabits per second (Mbps). Since the load-balancing procedures defined in this document operate on relative weights, the value MAY alternatively be used as a generalized weight (e.g., link count, locally configured weight, or a value derived from other operational considerations).

When a generalized weight is used, the operator MUST ensure consistent interpretation of the advertised value across all egress PEs associated with the Ethernet Segment. This requirement applies even when the egress PEs span multiple routing domains or Autonomous Systems.

To enable unambiguous interpretation and to reduce the risk of provisioning errors, one octet of the six-octet extended community value field is used to explicitly indicate whether the remaining five octets encode link bandwidth in Mbps or a generalized weight. The EVPN Link Bandwidth Extended Community is encoded as follows:

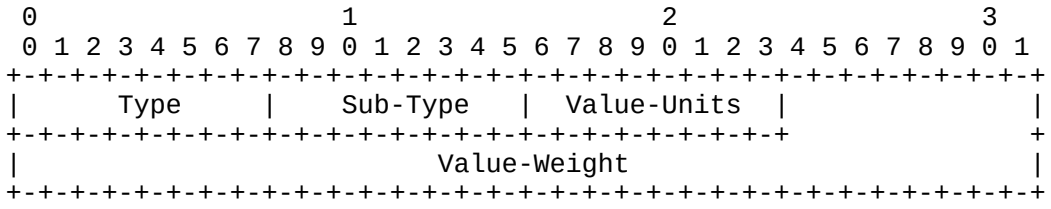

Figure 4

The Value-Weight field is encoded as a 5-octet unsigned integer, in network byte order, with a valid range of 0 to 2^40亓.

The Value-Units field is encoded as follows:

* 0x00: Weight expressed in Mbps (default)

* 0x01: Generalized weight expressed in units other than Mbps

Generalized weight units are intentionally left unspecified to allow flexibility across different applications without requiring additional encodings. Implementations MUST support the default unit of Mbps. Support for generalized weight values is OPTIONAL. Value-Units code points other than those defined in this document are out of scope.

4.1.1. BGP Error Handling

The following error-handling procedures apply to the processing of the EVPN Link Bandwidth Extended Community:

* Value-Units Consistency: When an EVPN Link Bandwidth Extended Community is received with a route, a PE MUST verify that the Value-Units field is consistent across all paths associated with that route. For Route Types 1 (RT-1) and 4 (RT-4), if the receiving PE is also a member of the corresponding Ethernet Segment, it MUST additionally verify that the received Value-Units value is consistent with its locally configured Value-Units for that Ethernet Segment. If any inconsistency is detected, the extended community MUST be ignored for all paths associated with the route.

* Multiplicity: A PE MUST ensure that at most one instance of the EVPN Link Bandwidth Extended Community is received per path. If more than one instance is present, the extended community MUST be ignored for all paths associated with the route.

* Non-Zero Value-Weight: Only non-zero Value-Weight values are considered valid. If a Value-Weight value of zero is received on any path, the extended community MUST be treated as invalid and ignored for all paths associated with the route.

* Unsupported Value-Units: If a PE receives a Value-Units value that it does not support, the EVPN Link Bandwidth Extended Community MUST be ignored and MUST NOT be used for load-balancing purposes.

* Invalid or Missing Extended Community: If the EVPN Link Bandwidth Extended Community is not received on any path, or is determined to be invalid on any path for any of the reasons listed above, it MUST be ignored for all paths associated with the route. In such cases:

  - For RT-1, regular EVPN ECMP procedures apply to all routes associated with the Ethernet Segment.

- For RT-4, regular DF election procedures apply to the Ethernet Segment.

- For RT-5, regular ECMP procedures apply to the IP Prefix route. A syslog message SHOULD be generated indicating the reason why the extended community was ignored.

* Unexpected Route Types: This document specifies the use of the EVPN Link Bandwidth Extended Community only with per-ES RT-1, RT-4, and RT-5 routes. If the extended community is received with any other EVPN route type, including per-[ES, EVI] RT-1 or RT-2 routes, it MUST be ignored, and a syslog message SHOULD be generated indicating the reason.

5. Weighted Unicast Traffic Load-balancing to an Ethernet Segment

5.1. Egress PE Behavior

A PE that is part of an Ethernet Segment’s redundancy group MUST advertise an additional "EVPN link bandwidth" extended community with Ethernet A-D per ES route (EVPN Route Type 1), that carries total bandwidth of PE’s physical links in an Ethernet Segment or a generalized weight. New EVPN link bandwidth extended community defined in this document is used for this purpose.

EVPN link bandwidth extended community MUST NOT be attached to per-EVI RT-1 or to EVPN RT-2 as it is a physical ESI property and hence advertised per-ESI.

5.2. Ingress PE Behavior

An ingress PE MUST ensure that the EVPN Link Bandwidth Extended Community is received from all egress PEs associated with a given ES, and MUST verify that the received Value-Units are consistent across all such egress PEs. If the EVPN Link Bandwidth Extended Community is missing from one or more egress PEs, or if inconsistent Value-Units are detected, the ingress PE MUST ignore the EVPN Link Bandwidth Extended Community for that ES and MUST revert to regular ECMP forwarding toward that ES. When the EVPN Link Bandwidth Extended Community is ignored, the ingress PE SHOULD generate a syslog notification.

For all paths that are accepted by BGP and installed in the RIB as forwarding paths, once Value-Units consistency has been successfully validated, the ingress PE SHOULD use the received Value-Weight from each egress PE to derive a relative (normalized) weight per egress PE for the ES. These normalized weights SHOULD then be used to construct a weighted forwarding path-list for load balancing, instead

of using an ECMP-based path-list. The computation of egress PE weights and the resulting weighted path-list at the ingress PE is a local implementation matter.

An example computation algorithm is shown below for illustrative purposes.

Let:

L(x,y) denote the link bandwidth advertised by egress PE-x for ES-y

W(x,y) denote the normalized weight assigned to egress PE-x for ES-y

H(y) denote the highest common factor (HCF) of [L(1,y), L(2,y), ⋯ , L(n,y)]

The normalized weight assigned to egress PE-x for ES-y may be computed as:

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

For a MAC+IP Advertisement Route (EVPN Route Type 2) received for ES-y, the ingress PE MAY compute a MAC and IP forwarding path-list weighted according to the normalized weights defined above.

As an example, consider a CE device multi-homed to PE-1, PE-2, and PE-3 via physical links of 2 Gbps, 1 Gbps, and 1 Gbps respectively, as part of a LAG represented by ES-10:

L(1,10) = 2000 Mbps

L(2,10) = 1000 Mbps

L(3,10) = 1000 Mbps

H(10) = 1000

The resulting normalized weights are:

W(1,10) = 2

W(2,10) = 1

W(3,10) = 1

For a remote MAC+IP host route associated with ES-10, the resulting forwarding path-list MAY therefore be computed as:

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

instead of:

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

This results in traffic destined to ES-10 being load-balanced across the egress PEs in proportion to the bandwidth advertised for ES-10 by each egress PE.

The above computation algorithm is provided for illustration only. Weighted path-list computation based on the EVPN Link Bandwidth Extended Community is a local implementation choice. If the received bandwidth values do not yield a suitable HCF that allows programming reasonable integer weights in hardware, an implementation MAY apply alternative approximation or rounding methods to derive implementable weight values.

Weighted path-list computation MUST be performed for an ES only if the EVPN Link Bandwidth Extended Community is received from all egress PEs advertising reachability to that ES via Ethernet A-D per-ES Route Type 1. If the EVPN Link Bandwidth Extended Community is not received from one or more such egress PEs, the ingress PE MUST compute the forwarding path-list using regular ECMP semantics. A default weight MUST NOT be assumed for an egress PE that does not advertise link bandwidth, as the computed weights are strictly relative.

If a per-ES Route Type 1 is not advertised, or is withdrawn, by an egress PE as specified in [RFC7432], that egress PE MUST be removed from the forwarding path-list for the corresponding [EVI, ES], and the weighted path-list MUST be recomputed accordingly.

If a per-[ES, EVI] Route Type 1 is not advertised by an egress PE as specified in [RFC7432], that egress PE MUST NOT be included in the forwarding path-list for the corresponding [EVI, ES]. In this case, the weighted path-list MUST be computed using only the weights received from egress PEs that advertised the per-[ES, EVI] Route Type 1.

6. Weighted BUM Traffic Load-Sharing across an Ethernet Segment

Optionally, load-sharing of per-service DF roles, weighted by the relative link bandwidth contribution of each egress PE within a multi-homed ES, may be supported.

To enable this behavior, this document defines a new DF Election Capability, as specified in [RFC8584], called BW (Bandwidth Weighted DF Election). The BW Capability MAY be used in conjunction with selected DF Election Types, as described in the following sections.

6.1. The BW Capability in the DF Election Extended Community

This document requests IANA to allocate a new bit in the DF Election Capabilities registry defined by [RFC8584]:

Bit 4: BW (Bandwidth Weighted DF Election)

When an ES route is advertised with the BW bit set, the advertising egress PE indicates its desire for link-bandwidth information to be considered in the DF Election algorithm specified by the associated DF Type.

As specified in [RFC8584], all egress PEs associated with an ES MUST advertise identical DF Election Capabilities and the same DF Type. If this condition is not met, all PEs MUST revert to the default DF Election procedure defined in [RFC7432].

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

* Type 0: Default DF Election algorithm, as specified in [RFC7432]

* Type 1: Highest Random Weight (HRW) algorithm, as specified in [RFC8584]

* Type 2: Preference-based DF Election algorithm, as specified in [EVPN-DF-PREF]

* Type 4: HRW per-multicast-flow DF Election algorithm, as specified in [EVPN-PER-MCAST-FLOW-DF]

The following sections describe the modifications to the DF Election procedures for these DF Types when the BW Capability is in use.

6.2. BW Capability and Default DF Election Algorithm

When all PEs in an Ethernet Segment agree to use the BW Capability with DF Type 0, the Default DF Election procedure defined in [RFC7432] is modified as follows:

* Each egress PE MUST advertise an EVPN Link Bandwidth Extended Community along with the ES route to signal the PECE link bandwidth associated with the ES.

* A receiving PE MUST use the EVPN Link Bandwidth Extended Community received from all egress PEs to compute a relative (normalized) weight for each egress PE within the ES.

* The DF Election procedure MUST use the resulting weighted list of candidate egress PEs when computing the per-VLAN Designated Forwarder, such that DF roles are distributed in proportion to the normalized weights.

As a result, a given PE MAY appear multiple times in the DF candidate list. Consequently, the value N used in the (V mod N) operation defined in [RFC7432] MUST be interpreted as the total number of ordinals in the weighted candidate list, rather than the total number of distinct egress PEs in the ES.

Using the example from Section 5.2, the DF candidate list becomes:

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

For a given VLAN-a on ES-10, the DF is computed as (VLAN-a mod 4). This results in DF role assignment across PE-1, PE-2, and PE-3 in proportion to their normalized weights for ES-10.

6.3. 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.

6.3.1. BW Increment

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

In the context of an ES,

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

Lmin = 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) / Lmin

As an example,

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

bandwidth increment for each PE would be computed as:

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

with L(1) = 10, L(2) = 10, L(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 Lmin. If it computes to a non-integer value (including as a result of link failure), it MUST be rounded down to an integer.

6.3.2. HRW Hash Computations with BW Increment

HRW algorithm as described in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] computes a random hash value 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(i).

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

We refer to this hash value as "affinity" in this document. Hash or 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] MUST 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] MUST 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)

6.4. 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] is updated to now consider the LBW value as below:

f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP (Don't Preempt me) 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].

6.5. Cost-Benefit Tradeoff on Link Failures

Incorporating link bandwidth into the DF election process enables more optimal distribution of BUM traffic across the links of an ES. However, doing so also causes DF election outcomes to change in response to link failures or link bandwidth variations.

Operators that do not wish to incur this level of DF election churn SHOULD NOT advertise the BW Capability. In the absence of the BW Capability, BUM traffic distribution across the ES links may be suboptimal; however, this approach preserves DF election stability while still allowing ingress PEs to apply weighted ECMP for unicast traffic toward the set of egress PEs.

7. Additional Considerations

7.1. Real-time Available Bandwidth

PE-CE link bandwidth availability may sometimes vary in real-time disproportionately across PE-CE links within a multi-homed ES 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.

7.2. Weighted Load-balancing to Multi-homed Subnets

EVPN Link bandwidth extended community may also be used to achieve unequal load-balancing of prefix routed traffic by including this extended community in EVPN Route Type 5. When included in EVPN RT-5, its value is to be interpreted as egress PE's relative weight for the prefix included in this RT-5. Ingress PE will then compute the forwarding path-list for the prefix route using weighted paths received from each egress PE. EVPN Link bandwidth extended community

MUST be encoded with "Value-Units = 0x01" to signal a generalized weight associated with the advertising PE.

7.3. Weighted Load-balancing without EVPN aliasing

[RFC7432] defines per-[ES, EVI] RT-1 based EVPN aliasing procedure as an optional propcedure. In an unlikely scenario where an EVPN implementation does not support EVPN aliasing procedures, MAC forwarding path-list at the ingress PE is computed based on per-ES RT-1 and RT-2 routes received from egress PEs instead of per-ES RT-1 and per-[ES, EVI] RT-1 from egress PEs. In such a case, only the weights received via per-ES RT-1 from the egress PEs included in the MAC path-list are to be considered for weighted path-list computation.

7.4. 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. Applicability of weighted ECMP procedures specified in this document to these set of use cases is an area of further consideration beyond the scope of this document.

7.5. EVPN Link Bandwidth Extended Community in Non-EVPN Networks

While this document does not preclude future applicability to non-EVPN networks, it considers usage and handling of EVPN Link Bandwidth Extended Community specified in this document with non-EVPN routes out of scope.

7.6. Preference for EVPN Link Bandwidth in EVPN Networks

It is possible that a non-EVPN Link Bandwidth extended community such as [BGP-LINK-BW] is leaked from an IP or IPVPN route into an EVPN RT-5 towards an EVPN network. If an EVPN PE receives an EVPN route with both the EVPN Link Bandwidth extended community specified in this document and a non-EVPN Link Bandwidth extended community such as the one specified in [BGP-LINK-BW], it MUST as default behavior, prefer the EVPN Link Bandwidth extended community and handle it as per procedures specified in this document. In other words, any non-EVPN Link Bandwidth extended community is to be ignored if an EVPN route is received with the EVPN Link Bandwidth extended community specified in this document.

It is recommended that an implementation SHOULD provide a way to modify the above default local preference order via optional provisioning.

7.7. Interworking with Non-EVPN networks

In EVPN routing interworking use cases with IPVPN and IPv4/IPv6 routing, it is not beneficial to preserve the the EVPN Link Bandwidth extended community from EVPN routes to non-EVPN routes as the next-hop is rewritten when a prefix learnt via EVPN RT-5 is advertised into IPVPN or IP routing networks. Interworking procedures, including preservation, cummulation or translation of EVPN Link Bandwidth extended community to address current or future use cases are however considered beyond the scope of this document. Readers are encouraged to refer to [EVPN-IPVPN] for interworking specification.

8. Operational Considerations

* In order for the solution specified in this document to function correctly, implementation SHOULD ensure that EVPN Link Bandwidth Extended Comminiuty is being advertised with same "Value-Units" across all PEs.

* Further, when a generalized weight option is used with "Value-Units = 0x1", implementation SHOULD ensure that the weights are assigned to each PE in a consistent manner.

* Implementation SHOULD alert the users via syslog when an inconsistency in "Value-Units" is detected across the PE set for a given ESI or prefix.

* Implementation SHOULD also alert users via syslog if an unreasonable discrepancy is detected across advertised BW or weights from different PEs, such that the implementation is unable to compute a weighted pathlist that can be programmed in hardware. This could likely result from inconsistent units of weight used by different PEs.

* Operators MAY monitor the traffic flow distribution and DF election distribution across the egress PE set to ensure that the implementation is working as expected.

9. Security Considerations

Security considerations discussed in [RFC7432] and [RFC8584] apply to this document. Methods described in this document further extend signaling of multi-homed devices using ESI LAG. They are hence subject to same considerations if the control plane or data plane was to be compromised. As an example, if control plane is compromised, signaling of heavily skewed Link Bandwidth Attributes could result in all traffic to be directed towards one PE resulting in its host facing links to be overloaded. Exposure to such an attack is limited by suggested syslogs discussed in Operational Consideration section. Considerations for protecting control and data plane described in [RFC7432] are equally applicable to signaling of Link Bandwidth Attribute defined in this document.

10. IANA Considerations

10.1. Bandwidth Weighted DF Election Capability

[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 IANA to allocate a bit in the "DF Election capabilities" registry setup by [RFC8584] with the following suggested bit number:

Bit 4: BW (Bandwidth Weighted DF Election)

10.2. EVPN Link Bandwidth Extended Community

This document defines a new EVPN Link Bandwidth extended community to signal local ES link bandwidth to ingress PEs. This extended community is defined of type 0x06 (EVPN Extended Community Sub-Types). IANA has assigned a sub-type value of 0x10 for the EVPN Link bandwidth extended community, of type 0x06 (EVPN Extended Community Sub-Types). EVPN Link Bandwidth extended community is defined as transitive.

10.3. Value-Units Registry

IANA is requested to set up a registry called "Value-Units" for the 1-octet field in the EVPN Link Bandwidth Extended Community. New registrations will be made through the "RFC Required" procedure defined in [RFC8126]. The following are suggested initial values in that registry exist:

Value	Name	Reference
0	Weight in units of Mbps	This document
1	Generalized Weight	This document
2-255	Unassigned	

11. Acknowledgements

Authors would like to thank Gunter Van de Velde, Jeffrey Haas, Stephane Litkowski, and Dhruv Dhody for multiple reviews and contributing valuable improvements to document. Authors would also like to thank Satya Mohanty for valuable review and inputs with respect to HRW and weighted HRW algorithm refinements specified in this document. Authors would also like to thank Bruno Decraene and Sergey Fomin for valuable review and comments.

12. Contributors

Satya Ranjan Mohanty
Cisco Systems
US
Email: satyamoh@cisco.com

13. References

13.1. Normative References

[EVPN-DF-PREF]
Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., Mohanty, S., and , "Preference-based EVPN DF Election", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-pref-df-13, 19 June 2020, <https://tools.ietf.org/html/draft-ietf-bess-evpn-pref-df-13.txt>.

[EVPN-PER-MCAST-FLOW-DF]
Sajassi, A., mishra, m., Thoria, S., Rabadan, J., and J. Drake, "Per multicast flow Designated Forwarder Election for EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-per-mcast-flow-df-election-04, 31 August 2020, <http://www.ietf.org/internet-drafts/draft-ietf-bess-evpn-per-mcast-flow-df-election-04.txt>.

[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>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 8126, June 2017, <https://www.rfc-editor.org/rfc/rfc8126>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", RFC 8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

[RFC8584] Rabadan, J., Ed., Mohanty, R., Sajassi, N., Drake, A., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <https://www.rfc-editor.org/info/rfc8584>.

13.2. Informative References

[BGP-LINK-BW]
Mohapatra, P., Mohanty, S., and S. Krier, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-19, May 2025, <https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-19.txt>.

[EVPN-IPVPN]
Rabadan, J., Sajassi, A., Drake, J., Lin, W., and J. Uttaro, "EVPN Interworking with IPVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-ipvpn-interworking, June 2025, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-ipvpn-interworking-14.txt>.

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

Appendix A. BGP-Link-Bandwidth-Extended-Community

Link bandwidth extended community described in [BGP-LINK-BW] for layer 3 VPNs was considered for re-use here. This Link bandwidth extended community is however defined in [BGP-LINK-BW] as optional non-transitive. Since it is not possible to change deployed behavior of extended community defined in [BGP-LINK-BW], it was decided to define a new one. In inter-AS scenarios within an EVPN network, EVPN

link-bandwidth needs to be signaled to eBGP neighbors. When signaled across AS boundary, this extended community can be used to achieve optimal load-balancing towards egress PEs in a different AS. This is applicable both when next-hop is changed or unchanged across AS boundaries.

Authors' Addresses

Neeraj Malhotra (editor)
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: nmalhotr@cisco.com

Ali Sajassi
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: sajassi@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
United States of America
Email: jorge.rabadan@nokia.com

John Drake
Independent
Email: je_drake@yahoo.com

Avinash Lingala
ATT
200 S. Laurel Avenue
Middletown, CA 07748
United States of America
Email: ar977m@att.com

Samir Thoria
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: sthoria@cisco.com