                 Weighted Multi-Path Procedures for EVPN Multi-Homing
                    draft-ietf-bess-evpn-unequal-lb-11

Abstract

   EVPN enables all-active multi-homing for a CE device connected to two
   or more PEs via a LAG, such that bridged and routed traffic from
   remote PEs to hosts attached to the Ethernet Segment can be equally
   load balanced (it uses Equal Cost Multi Path) across the multi-homing
   PEs. EVPN also enables multi-homing for IP subnets advertised in IP
   Prefix routes, so that routed traffic from remote PEs to those IP
   subnets can be load balanced. This document defines extensions to
   EVPN procedures to optimally handle unequal access bandwidth
   distribution across a set of multi-homing PEs in order to:

   o provide greater flexibility, with respect to adding or removing
      individual multi-homed PE-CE links.

   o handle multi-homed PE-CE link failures that can result in unequal
      PE-CE access bandwidth across a set of multi-homing PEs.

Status of This Memo

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 November 8, 2021.

Copyright Notice

Table of Contents

1. Requirements Language and 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 [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, whose overlay reachability to the ESI
is via the Local PE.

o BW: BandWidth

o LAG: Link Aggregation Group

o ES: Ethernet Segment

o vES: Virtual Ethernet Segment

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

o IMET: Inclusive Multicast Route

o DF: Designated Forwarder

o BDF: Backup Designated Forwarder

o DCI: Data Center Interconnect Router

2. 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:

o ECMP Load-balancing for bridged unicast traffic is enabled via

o aliasing and mass-withdraw procedures detailed in RFC 7432.

o ECMP Load-balancing for routed unicast traffic is enabled via
   existing L3 ECMP mechanisms.

o 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.

2.1. PE-CE Link Provisioning

```
                    +-----------------------+
                    | Underlay Network Fabric|
                    +-----------------------+


                  +-----+   +-----+
                  | PE1 |   | PE2 |
                  +-----+   +-----+
                    \          /
                     \ ESI-1 /
                      \      /
                      +\---/+
                      | \ / |
                      +--+--+
                         |
                        CE1


                        Figure 1
```
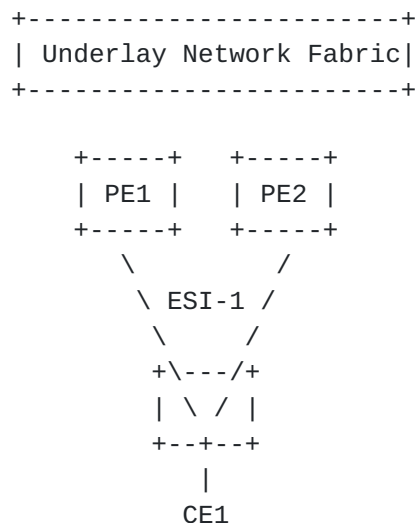
Consider CE1 that is dual-homed to PE1 and PE2 via EVPN all-active
multi-homing with single member links of equal bandwidth to each PE
(aka, equal access bandwidth distribution across PE1 and PE2). If
the provider wants to increase link bandwidth to CE1, it must add a

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 be treated as equal cost paths at remote
PEs, and as a result may attract approximately equal amount of CE1
destined traffic, 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.

2.2. 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.

```
               +------------------------+
               | Underlay Network Fabric|
               +------------------------+

                 +-----+   +-----+
                 | PE1 |   | PE2 |
                 +-----+   +-----+
                   \\           //
                    \\ ESI-1 //
                     \\      /X
                     +\\---//+
                     | \\ // |
                     +---+---+
                         |
                        CE1
```
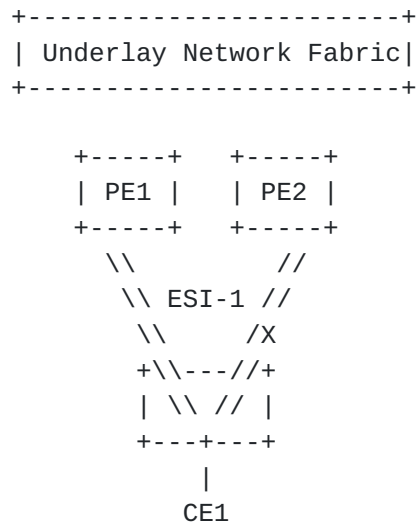
                        Figure 2

   Consider a CE1 that is multi-homed to PE1 and PE2 via a LAG with two
   member links to each PE. On a PE2-CE1 physical link failure, LAG
   represented by an Ethernet Segment ESI-1 on PE2 stays up, however,
   its bandwidth is cut in half. With existing ECMP procedures, both
   PE1 and PE2 may 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. As an alternative, min-link on LAGs is
   sometimes used to bring down the LAG interface on member link
   failures. This however results in loss of available bandwidth in the
   network, and is not ideal.

2.3. Design Requirement

```
                +-----------------------+
                |Underlay Network Fabric|
                +-----------------------+

        +-----+    +-----+           +-----+    +-----+
        | PE1 |    | PE2 |   .....   | PEx |    | PEn |
        +-----+    +-----+           +-----+    +-----+
           \          \                //         //
            \ L1       \ L2           // Lx      // Ln
             \          \            //         //
           +-\-------\-----------//--------//-+
           |  \          \  ESI-1  //         //  |
           +--------------------------------+
                            |
                           CE
```

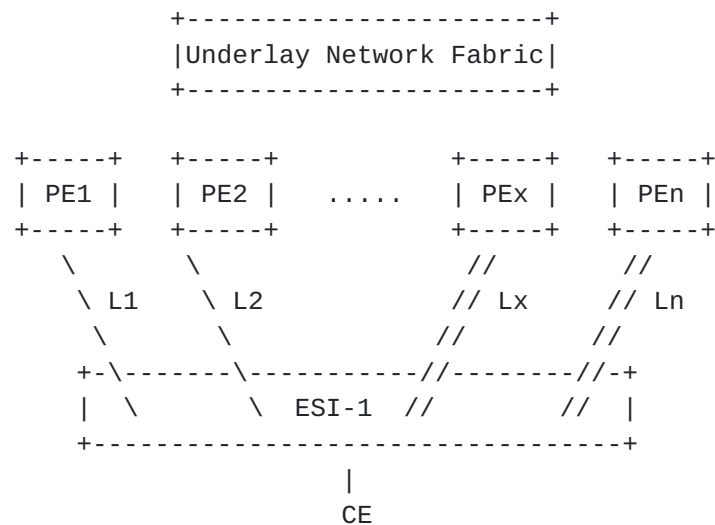   Figure 3

   To generalize, if total link bandwidth to a CE is distributed across
   "n" multi-homing PEs, with Lx being the total bandwidth to PEx across
   all links, 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]

   Figure 3 illustrates a scenario where PE1..PEn are attached to a
   multi-homed Ethernet Segment, however this document generalizes this
   requirement so that the unequal load balancing can be applied to PEs
   attached to a vES or to a multi-homed subnet advertised by EVPN IP
   Prefix routes.

   The solution proposed below includes extensions to EVPN procedures to
   achieve the above.

3. Solution Overview

   In order to achieve weighted load balancing to an ES or vES for
   overlay unicast traffic, Ethernet A-D per ES route (EVPN Route Type
   1) is leveraged to signal the Ethernet Segment weight to remote PEs.
   Using Ethernet A-D per-ES route to signal the Ethernet Segment weight
   provides a mechanism that reacts to changes in access bandwidth or
   number of access links 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 or number of
   links received from each PE that the ES is multi-homed to.

In order to achieve weighted load balancing of overlay BUM traffic,
EVPN ES route (Route Type 4) is leveraged to signal the ESI weight 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 weight.

Unequal load balancing to multi-homed subnets is achieved by
signaling the weight along with the IP Prefix routes advertised for
the subnet.

Procedures to accomplish this are described in greater detail next.

4. EVPN Link Bandwidth Extended Community

A new EVPN Link Bandwidth extended community is defined for the
solution specified in this document:

o This extended-community is defined of type 0x06 (EVPN).

o IANA is requested to assign a sub-type value of 0x10 for the EVPN
   Link bandwidth extended community, of type 0x06 (EVPN).

o EVPN Link Bandwidth extended community is defined as transitive.

4.1. Encoding and Usage of EVPN Link Bandwidth Extended Community

EVPN Link Bandwidth Extended Community value field is to be treated
as a 6 octet unsigned integer representing total bandwidth of PE's
all physical links in an ethernet segment, expressed in Mbits/sec
(MegabitsPerSecond). Note however that the load balancing algorithm
defined in this document uses ratio of Link Bandwidths. Hence, the
operator may choose a different unit or use the community as a
generalized weight that may be set to link count, locally configured
weight, or a value computed based on an attribute other than link
bandwidth. In such case, the operator MUST ensure consistent usage
of the unit across all PEs in an ethernet segment. This may involve
multiple routing domains/Autonomous Systems.

In order to avoid interop issues because of provisioning error, one
octet in the extended community's six octet 'value' field is used to
explicitly signal if the weight encoded in the remaining five octets
is link bandwidth expressed in Mbits/sec or a generalized weight
value. This results in the following encoding for EVPN link
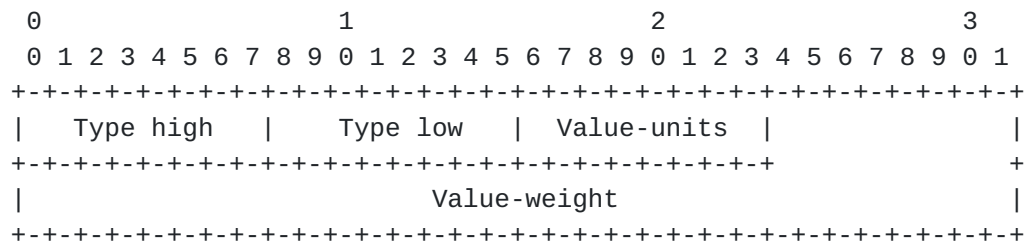bandwidth extended community:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type high   |   Type low    |  Value-units  |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               +
|                         Value-weight                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                                Figure 4

   Value-units is encoded as:

   o 0x00: weight expressed using default units of Mbits/sec

   o 0x01: generalized weight expressed in something other than Mbits/
      sec

   Generalized weight units are intentionally left arbritrary to allow
   for flexibility in its usage for different applications without
   having to define new encoding for each non-default application.
   Implementations SHOULD support the default units of Mbits/sec, while
   support of non-default generalized weight is considered optional.

4.2. Note on 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, link-bandwidth needs to be
   signaled to eBGP neighbors. When signaled across AS boundary, this
   attribute can be used to achieve optimal load-balancing towards
   source PEs from a different AS. This is applicable both when next-
   hop is changed or unchanged across AS boundaries.

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

5.1. Local PE Behavior

   A PE that is part of an Ethernet Segment's redundancy group would
   advertise an additional "link bandwidth" extended community 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. New
   EVPN link bandwidth extended community defined in this document is
   used for this purpose.

5.2. 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,
and then use this relative weight to compute a weighted path-list to
be used for load balancing, as opposed to using an ECMP path-list for
load balancing across the PE paths. PE Weight and resulting weighted
path-list computation at remote PEs is a local matter. An example
computation algorithm is shown below to illustrate the idea:

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 may 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 LAG represented by
ESI-10:

$L(1, 10) = 2000$ Mbits/sec

$L(2, 10) = 1000$ Mbits/sec

$L(3, 10) = 1000$ Mbits/sec

$H(10) = 1000$

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 may 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.

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 an
unlikely event that link bandwidth attribute is not received from one
or more subset of PEs, forwarding path-list should be computed using
regular ECMP semantics. Note that a default weight cannot be assumed
for a PE that does not advertise its link bandwidth as the weight
attribute t be used in path-list computation is relative.

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

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.

6.1. 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 IANA for 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:

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

o Type 1: HRW algorithm, as in [RFC8584]

o Type 2: Preference algorithm, as in [EVPN-DF-PREF]

o 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.

6.2. 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 as
defined in [RFC7432] is modified as follows:

o Each PE advertises a "Link Bandwidth" extended community attribute
   along with the ES route to signal the PE-CE link bandwidth (LBW)
   for the ES.

o A receiving PE MUST use the ES link bandwidth attribute received
   from each PE to compute a relative weight for each remote PE.

o 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. As a
   result, a single PE may have multiple ordinals in the DF candidate
   PE list and 'N' used in (V mode N) operation as defined in
   [RFC7432] is modified to be total number of ordinals instead of
   being total number of PEs.

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.

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 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.

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] 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)

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] 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:

   o 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.

   o 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.

   o The LBW exchanged value has no impact on the Non-Revertive option
      described in [EVPN-DF-PREF].

7. 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 DF elections are re-adjusted on link failures or
   bandwidth changes. 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.

8. 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.

9. 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 advertising PE's relative weight
for the prefix included in this RT-5. Receiving PE will then compute
the forwarding path-list for the prefix route using weighted paths
received from each remote PE.

10. 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 is an area of further consideration.

11. Operational Considerations

None

12. Security Considerations

This document raises no new security issues for EVPN.

13. IANA Considerations

[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 for a bit in
the DF Election extended community Bitmap:

Bit 28: BW (Bandwidth Weighted DF Election)

A new EVPN Link Bandwidth extended community is defined to signal
local ES link bandwidth to remote PEs. This extended-community is
defined of type 0x06 (EVPN). IANA is requested to assign a sub-type
value of 0x10 for the EVPN Link bandwidth extended community, of type
0x06 (EVPN). EVPN Link Bandwidth extended community is defined as
transitive.

14. 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. Authors would also like to thank Bruno
Decraene and Sergey Fomin for valuable review and comments.

15. Contributors

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

16. References

16.1. 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", draft-ietf-bess-evpn-pref-df-06
          (work in progress), June 2020.

[EVPN-PER-MCAST-FLOW-DF]
          Sajassi, A., mishra, m., Thoria, S., Rabadan, J., and J.
          Drake, "Per multicast flow Designated Forwarder Election
          for EVPN", draft-ietf-bess-evpn-per-mcast-flow-df-
          election-04 (work in progress), August 2020.

[EVPN-VIRTUAL-ES]
          Sajassi, A., Brissette, P., Schell, R., Drake, J.,
          Rabadan, J., and , "EVPN Virtual Ethernet Segment", draft-
          ietf-bess-evpn-virtual-eth-segment-06 (work in progress),
          March 2020.

[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>.

   [RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., and B. Fee,
              "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension
              Solution", RFC 7814, DOI 10.17487/RFC7814, March 2016,
              <https://tools.ietf.org/html/rfc7814>.

   [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>.

16.2. Informative References

   [BGP-LINK-BW]
              Mohapatra, P. and R. Fernando, "BGP Link Bandwidth
              Extended Community", draft-ietf-idr-link-bandwidth-07
              (work in progress), March 2019.

Authors' Addresses

   Neeraj Malhotra (editor)
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: nmalhotr@cisco.com

   Ali Sajassi
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: sajassi@cisco.com

   Jorge Rabadan
   Nokia
   777 E. Middlefield Road
   Mountain View, CA 94043
   USA

   Email: jorge.rabadan@nokia.com

   John Drake
   Juniper

   Email: jdrake@juniper.net

   Avinash Lingala
   ATT
   200 S. Laurel Avenue
   Middletown, CA 07748
   USA

   Email: ar977m@att.com

   Samir Thoria
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: sthoria@cisco.com