bess
Internet-Draft
Intended status: Standards Track
Expires: 22 August 2025

Z. Zhang
W. Lin
Juniper Networks
J. Rabadan
Nokia
A. Sajassi
Cisco
C. Lin
New H3C Technologies

18 February 2025

Dynamic Overlay Load Balancing
draft-zzhang-bess-dynamic-overlay-lb-00

Abstract

This document specifies a mechanism for an overlay service ingress PE to dynamically load-balance traffic to Multi-Homing PEs based on near real-time access link information advertised by those PEs.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 22 August 2025.

Copyright Notice

Table of Contents

1. Introduction

[I-D.ietf-bess-evpn-unequal-lb] specifies a mechanism to do weighted load-balancing on an Ethernet Virtual Private Network (EVPN) [RFC7432] ingress PE to egress PEs of Multi-Homed Ethernet Segments (MHES) based on the capacity of the MHES advertised by the egress PEs. The capacity advertisement is not real-time, and load-balancing based on dynamic information is outside the scope of that document and left for further study.

[I-D.wang-idr-next-next-hop-nodes] describes a scenario where global load-balancing can be achieved in a CLOS network by considering the real-time load information on the next hop router in addition to considering the real-time local load information of the path to that next hop router.

[I-D.zzhang-rtgwg-router-info] specifies a UDP-based mechanism to advertise router information including real-time load information for links, or for paths to some neighbors.

This document specifies how dynamic load-balancing can be achieved on ingress PEs based on near real-time access link information advertised by the multi-homing egress PEs for both L2 and L3 services. The difference from [I-D.wang-idr-next-next-hop-nodes] is that [I-D.wang-idr-next-next-hop-nodes] is for load-balancing across the underlay, while this document is for overlay services.

In the case of EVPN L2 services (via EVPN Type 2 routes) and L3 services (via EVPN Type 5 routes with a non-zero ESI) [RFC9136], in addition to calculating load-balancing weights according to

[I-D.ietf-bess-evpn-unequal-lb], the forwarding component of an EVPN PE can further fine-tune the weights based on real-time link information advertised from the MHPEs according to [I-D.zzhang-rtgwg-router-info]. The EVPN signaling is extended to signal a 32-bit local link ID for each MHES, and the link ID is used in the link load signaling per [I-D.zzhang-rtgwg-router-info].

In the case of EVPN L3 services via EVPN routes with a zero-ESI but a MAC/IP overlay index, if the overlay index itself is resolved via an MHES, the dynamic load-balancing of the L3 services recursively follows the dynamic load-balancing for the overlay index (see previous paragraph).

Otherwise, or in the case of IP-VPN [RFC4364], Labeled Unicast [RFC8277] among border routers (BDRs), or plain IP over tunnels to egress BDRs/ASBRs, the IP/VPN routes can carry a next-next-hop, which is used in the link/path load signaling via UDP, just as in [I-D.wang-idr-next-next-hop-nodes] with the difference that the signaling is to (remote) ingress PEs (or tunnel ingress nodes) instead of link-local flooding, and that the dynamic load-balancing is done by the ingress routers.

In addition to the dynamic load-balancing, the up/down status of an access link in the UDP advertisement per [I-D.zzhang-rtgwg-router-info] can also be used by the ingress PEs to quickly stop using an egress PE when its access link goes down - assuming that the UDP advertisement can arrive at the ingress PEs faster than the withdrawal of the EVPN Ethernet A-D per ES route (in the case of EVPN) or L3 IP/VPN routes, and that the UDP advertisement can be handled in the fast forwarding path in a fast reroute fashion similar to a local link down case.

When there are multiple paths to a PE in the underlay, the dynamic load-balancing to that PE via multiple paths in the underlay can be done in addition to the load-balancing to multiple PEs in the overlay.

2. Specification

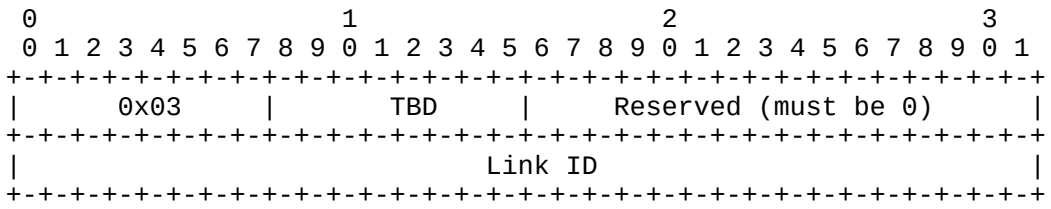
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The signaling and procedures in this document are optional. They are only used when dynamic load-balancing to egress PEs is desired.

2.1. EVPN

Each EVPN PE on an MHES assigns a local 32-bit link ID for its link to the MHES. In the Ethernet Auto-Discover (Type 1) per ES route originated by a PE for the MHES, a new Link ID Extended Community is attached to signal the local link ID.

The Link ID Extended Community is a transitive opaque Extended Community with a sub-type TBD:



The load info of the link is signaled using the Link Information TLV per [I-D.zzhang-rtgwg-router-info]. Each link to an MHES has a link record in the TLV, and each record's Link ID is the link's local link ID that is also signaled in the Link ID Extended Community.

The UDP messages are delivered in the underlay via one of the following methods:

* Individually addressed and delivered to each PE via unicast.

* Addressed to an IPv4 "All Routers on this Subnet" multicast address or an IPv6 Node-local All Routers Address (multicast) [RFC4291] and delivered over a P2MP tunnel to all other PEs.

* Addressed to an operator-specified multicast address and delivered over the multicast tree for that address in the underlay network. All PEs MUST join that multicast tree.

The Link ID in the UDP message MAY be set to a value that identifies the EVPN domain. It MAY be zero if the link info signaling is not used for other purposes (as the remote ingress PEs don't care from which link the egress PE sent the UDP message).

With the Link ID in the Link ID Extended Community attached to Ethernet Auto-Discovery per ES routes from egress PEs on an MHES, and the link info signaled from the MHES PEs, an ingress PE learns the up/down status and dynamic link load of each MHES link and adjusts the load-balancing weight dynamically, for both MAC-based L2 forwarding and IP-based L3 forwarding [I-D.ietf-bess-evpn-ip-aliasing] [I-D.mackenzie-bess-evpn-l3mh-proto].

An ingress PE may receive the dynamic link load information from some but not all of the PEs for an MHES, or the link load information from some may time out. The Link ID Extended Community may also be present in some but not all the Ethernet Auto-Discovery per ES routes. In that case, the ingress PE cannot determine the dynamic load information for some links of the MHES and SHOULD act as if such a link had a load of X% of its static bandwidth as advertised per [I-D.ietf-bess-evpn-unequal-lb]. If the link does not even have the static bandwidth information, then it MAY be considered to have a static bandwidth of the least of all received static bandwidths for all other links on the MHES. If the [I-D.ietf-bess-evpn-unequal-lb] signaling is not used at all, then all the links are considered to have an equal reference static bandwidth Y and the dynamic link load (either signaled per [I-D.zzhang-rtgwg-router-info] or assumed as X% per above) is based on Y. The actual value of Y does not matter and choice of X is a local operational consideration, but it SHOULD be consistent across all PEs. This document suggests a default value of 50 for X, and an operator can adjust X’s value depending on how much it wants to utilize a link that lacks the dynamic load info. Alternatively, an implementation MAY allow disabling the dynamic load-balancing for such a MHES.

The exact implementation for the load-balancing details is outside the scope of this document.

2.2. IP-VPN, Labeled Unicast, and Tunneled IP

The BGP procedures and signaling are the same as described in [I-D.wang-idr-next-next-hop-nodes] for prefixes multi-homed to multiple nodes (which could be egress PEs, BDRs, or ASBRs). In summary, the BGP routes for the multi-homed prefixes are advertised with a Next-next Hop Nodes (NNHN) TLV, which includes (among other things) one or more Next-next-hop BGP ID. These routes include EVPN Type 5 routes with a zero-ESI, for which the overlay index can not resolve to an MHES.

Each of the Next-next-hop BGP ID corresponds to a link/path on an egress node to the prefix. The node signals the link/path's dynamic load information using the Neighbor Path Information as specified in [I-D.zzhang-rtgwg-router-info], so that ingress nodes can dynamically load-balancing corresponding traffic via different egress nodes.

The UDP messages are delivered in the underlay to the ingress nodes as described in the EVPN case.

3. Security Considerations

To be added.

4. IANA Considerations

This document requests IANA to assign a sub-type value TBD from the Transitive Opaque Extended Community Sub-Types registry:

Sub-Type Value	Name
TBD	Link ID Extended Community

5. Acknowledgements

The authors thank Kevin Wang for his review, comments, and suggestions to make this document and solution more complete.

6. References

6.1. Normative References

[I-D.wang-idr-next-next-hop-nodes]
Wang, K., Haas, J., Lin, C., and J. Tantsura, "BGP Next-next Hop Nodes", Work in Progress, Internet-Draft, draft-wang-idr-next-next-hop-nodes-02, 2 December 2024, <https://datatracker.ietf.org/doc/html/draft-wang-idr-next-next-hop-nodes-02>.

[I-D.zzhang-rtgwg-router-info]
Zhang, Z. J., Wang, K., Lin, C., and N. Vaidya, "Advertising Router Information", Work in Progress, Internet-Draft, draft-zzhang-rtgwg-router-info-01, 18 September 2024, <https://datatracker.ietf.org/doc/html/draft-zzhang-rtgwg-router-info-01>.

[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>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

6.2. Informative References

[I-D.ietf-bess-evpn-ip-aliasing]
Sajassi, A., Rabadan, J., Pasupula, S., Krattiger, L., and J. Drake, "EVPN Support for L3 Fast Convergence and Aliasing/Backup Path", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-ip-aliasing-02, 4 November 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-ip-aliasing-02>.

[I-D.ietf-bess-evpn-unequal-lb]
Malhotra, N., Sajassi, A., Rabadan, J., Drake, J., Lingala, A. R., and S. Thoria, "Weighted Multi-Path Procedures for EVPN Multi-Homing", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-unequal-lb-24, 12 November 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-unequal-lb-24>.

[I-D.mackenzie-bess-evpn-l3mh-proto]
MacKenzie, M., Brissette, P., Matsushima, S., Lin, W., and J. Rabadan, "EVPN multi-homing support for L3 services", Work in Progress, Internet-Draft, draft-mackenzie-bess-evpn-l3mh-proto-05, 9 September 2024, <https://datatracker.ietf.org/doc/html/draft-mackenzie-bess-evpn-l3mh-proto-05>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[RFC9136] Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021, <https://www.rfc-editor.org/info/rfc9136>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks
Email: zzhang@juniper.net

Wen Lin
Juniper Networks
Email: wlin@juniper.net

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

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Changwang Lin
New H3C Technologies
Email: linchangwang.04414@h3c.com