bess
Internet-Draft
Intended status: Standards Track
Expires: 3 September 2026

Z. Zhang
W. Lin
HPE
C. Lin
New H3C Technologies

2 March 2026

EVPN Fast Notification and Handling for Multihoming
draft-zzhang-bess-evpn-mh-fast-notification-00

Abstract

EVPN supports powerful multihoming features, but it depends on the timely notification and handling of link going down or coming up on multihomed Ethernet Segments, which may not be guaranteed by the nature of control plane BGP signaling. When the handling is delayed, traffic duplication or loss may happen. This document specifies a UDP-based notification that can be originated and processed (near) instantly, greatly eliminate the potential traffic duplication or loss.

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 3 September 2026.

Copyright Notice

Copyright (c) 2026 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights

and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction 2
2. Specification 4
3. Security Considerations 5
4. IANA Considerations 5
5. Acknowledgments 5
6. References 5
6.1. Normative References 5
6.2. Informative References 5
Authors' Addresses 6

1. Introduction

EVPN [RFC7432] supports powerful multihoming features, but it depends on the timely notification and handling of link going down or coming up on multihomed Ethernet Segments, which may not be guaranteed by the nature of control plane BGP signaling. When the handling is delayed, traffic duplication or loss may happen.

PEs attached to a Multi-Homed Ethernet Segment (MHES) elect a Designated Forwarder (DF) and optionally elect a Backup Designated Forwarder (BDF). All the PEs in an EVPN instance learn the DF/BDF roles through the Primary/Backup-bit (P/B-bit) [RFC8214] in the EVPN Layer 2 Attributes Extended Community advertised with the EVPN Ethernet AutoDiscovery (type-1) route.

In the EVPN multihoming case, all the PEs on the MHES will receive a copy of the BUM traffic. In single-active multihoming case, only the DF sends the traffic to its locally attached MHES. In all-active multihomng case, the forwarding to the MHES depends on the following:

* With the MPLS data plane, if the traffic originated from the same MHES on another PE, it is not forwarded to the MHES at all by any receiving PE. Otherwise, only the DF forwards BUM traffic to the MHES.

* With the IP data plane (e.g., VXLAN), locally originated BUM traffic is forwarded to all MHESes regardless of the DF status. Remote BUM traffic is dropped on an MHES if the source PE is also attached to the MHES. Otherwise, only the DF forwards to the MHES. This behavior is called Local Bias.

If the MHES goes down on a DF, a new DF (typically the pre-elected BDF) needs to quickly take over and forward traffic on the local ES.

Remote ingress PEs need to quickly know the changes on a remote MHES, too. In the case of Single-Active, if the DF's link goes down on an MHES, the remote ingress PEs need to quickly switch to sending unicast to the new DF. In the case of All-Active, the remote ingress PEs may load balance to all the peers on the remote MHES, so any down event needs to be quickly notified to all remote ingress PEs.

In the case of Local Bias, if the MHES goes down on a PE, other PEs on the same MHES need to quickly update their Local Bias state, so that BUM traffic from that PE will be forwarded by the DF (which could be the current one or a new one - the original pre-elected BDF - if that PE was the DF) on the MHES.

Conversely, if the MHES comes up on a PE, other PEs on the same MHES need to quickly update their Local Bias state so that BUM traffic from that PE is no longer forwarded to the MHES even by the DF.

If the handling is delayed, traffic duplication/loss will happen. However, the delay is easy to happen due to the following reasons:

* The "slow" control plane (e.g., a routing engine) needs to learn the MHES state change and then originate or withdraw a corresponding EVPN Ethernet Segment (type-4) route and EVPN Ethernet Auto Discovery (type-1) per ES route.

* The route is subject to various BGP route propagation control/ effects and even TCP flow control.

* The receiving PE's "slow" control plane needs to get the route, process it, and then update the forwarding states in the "fast" forwarding plane (e.g., the line cards).

As a result, significant duplication/loss could happen, especially in scaled scenarios.

While BFD [RFC5880] can be used to quickly detect the link faliures, one BFD sessoin would be needed for each MHES, with the BFD packets sent all the time. This document specifies an alternative light-weight solution.

2. Specification

To speed up the process, the notification SHOULD be sent in the forwarding plane via UDP per [I-D.zzhang-rtgwg-router-info] and [I-D.zzhang-bess-dynamic-overlay-lb], and handled in the forwarding plane on the receiving PEs.

The MHES down notification on a DF in the case of Single-Active or of any role in the case of All-Active MAY be sent via multicast to all PEs, so that for unicast traffic destined to the MHES they can quickly switch to sending to the pre-elected BDF in case of Single-Active, or load balancing to the rest of the PEs on the MHES in the case of All-Active. This significantly speeds up the global repair on the ingress PEs, reducing the importance of egress protection.

If multicast is not available, the above mentioned-notification to all PEs MAY be sent via unicast, or may be skipped to avoid the burden of individual notifications.

Additionally, except in the case of down notifications already being sent to all PEs as mentioned in the previous paragraph:

* If Local Bias is used, any MHES up/down event SHOULD be notified via unicast to known peers on the MHES (so that they can update their Local Bias state, and if the down event is on the DF the BDF can take over).

* If Local Bias is not used, any MHES down event on a DF SHOULD be notified via unicast to the BDF on the MHES (so that the BDF can quickly take over).

To compensate for potential loss of the unreliable UDP notifications, several copies SHOULD be sent quickly in a row.

[I-D.zzhang-bess-dynamic-overlay-lb] specifies that MHES PEs can dynamically signal link loads on the MHES so that remote ingress PEs may load-balance unicast traffic to different MHES PEs based on the dynamic load information.

Even if the dynamic load-balancing based on dynamic load information is not used, the same UDP notification specified in {{!I-D.zzhang-bess-dynamic-overlay-lb" is used for the fast notification in this document, with the following differences:

* The notification is only sent when an MHES comes up or goes down. The load information is set to 0 and ignored by the receiving PEs.

* In all cases, the notification carries an indication if the originating PE was a DF on the MHES. The exact encoding is TBD.

3. Security Considerations

To be added.

4. IANA Considerations

To be added.

5. Acknowledgments

6. References

6.1. Normative References

[I-D.zzhang-bess-dynamic-overlay-lb]
Zhang, Z. J., Lin, W., Rabadan, J., and C. Lin, "Dynamic Overlay Load Balancing", Work in Progress, Internet-Draft, draft-zzhang-bess-dynamic-overlay-lb-01, 20 October 2025, <https://datatracker.ietf.org/doc/html/draft-zzhang-bess-dynamic-overlay-lb-01>.

[I-D.zzhang-rtgwg-router-info]
Zhang, Z. J., Wang, K., Lin, C., Vaidya, N., Tantsura, J., and Y. Liu, "Advertising Router Information", Work in Progress, Internet-Draft, draft-zzhang-rtgwg-router-info-06, 23 February 2026, <https://datatracker.ietf.org/doc/html/draft-zzhang-rtgwg-router-info-06>.

[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>.

6.2. Informative References

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

Authors' Addresses

Zhaohui Zhang
HPE
Email: zhaohui.zhang@hpe.com

Wen Lin
HPE
Email: wen.lin@hpe.com

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