BFD
Internet-Draft
Intended status: Standards Track
Expires: June 29, 2019

S. Pallagatti, Ed.
Rtbrick
S. Paragiri
Juniper Networks
V. Govindan
M. Mudigonda
Cisco
G. Mirsky
ZTE Corp.
December 26, 2018

BFD for VXLAN
draft-ietf-bfd-vxlan-06

Abstract

This document describes the use of the Bidirectional Forwarding Detection (BFD) protocol in Virtual eXtensible Local Area Network (VXLAN) overlay networks.

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 June 29, 2019.

Copyright Notice

Copyright (c) 2018 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 Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

"Virtual eXtensible Local Area Network" (VXLAN) [RFC7348]. provides an encapsulation scheme that allows building an overlay network by decoupling the address space of the attached virtual hosts from that of the network.

One use of VXLAN is in data centers interconnecting VMs of a tenant. VXLAN addresses requirements of the Layer 2 and Layer 3 data center network infrastructure in the presence of VMs in a multi-tenant environment, discussed in section 3 [RFC7348], by providing Layer 2 overlay scheme on a Layer 3 network. Another use is as an encapsulation for Ethernet VPN [RFC8365].

This document is written assuming the use of VXLAN for virtualized hosts and refers to VMs and VTEPs in hypervisors. However, the concepts are equally applicable to non-virtualized hosts attached to VTEPs in switches.

In the absence of a router in the overlay, a VM can communicate with another VM only if they are on the same VXLAN segment. VMs are unaware of VXLAN tunnels as a VXLAN tunnel is terminated on a VXLAN Tunnel End Point (VTEP) (hypervisor/TOR). VTEPs (hypervisor/TOR) are responsible for encapsulating and decapsulating frames exchanged among VMs.

Ability to monitor path continuity, i.e., perform proactive continuity check (CC) for these tunnels, is important. The asynchronous mode of BFD, as defined in [RFC5880], can be used to monitor a VXLAN tunnel. Use of [I-D.ietf-bfd-multipoint] is for future study.

Also, BFD in VXLAN can be used to monitor the particular service nodes that are designated to handle Layer 2 broadcast properly, unknown unicast, and multicast traffic. Such nodes, discussed in details in [RFC8293], are often referred to as "replicators", are usually virtual VTEPs and can be monitored by physical VTEPs to minimize BUM traffic directed to the unavailable replicator.

This document describes the use of Bidirectional Forwarding Detection (BFD) protocol VXLAN to enable monitoring continuity of the path between Network Virtualization Edges (NVEs) and/or availability of a replicator service node using BFD.

In this document, the terms NVE and VTEP are used interchangeably.

2. Conventions used in this document

2.1. Terminology

BFD - Bidirectional Forwarding Detection

CC - Continuity Check

NVE - Network Virtualization Edge

TOR - Top of Rack

VM - Virtual Machine

VTEP - VXLAN Tunnel End Point

VXLAN - Virtual eXtensible Local Area Network

2.2. Requirements Language

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.

3. Use cases

The primary use case of BFD for VXLAN is for continuity check of a tunnel. By exchanging BFD control packets between VTEPs, an operator exercises the VXLAN path in both the underlay and overlay thus ensuring the VXLAN path availability and VTEPs reachability. BFD failure detection can be used for maintenance. There are other use cases such as the following:

Layer 2 VMs:

Deployments might have VMs with only L2 capabilities and not have an IP address assigned or, in other cases, VMs are assigned IP address but are restricted to communicate only within their subnet. BFD being an L3 protocol can be used as a tunnel CC mechanism, where BFD will start and terminate at the NVEs, e.g., VTEPs.

It is possible to aggregate the CC sessions for multiple tenants by running a BFD session between the VTEPs over VxLAN tunnel.

Fault localization:

It is also possible that VMs are L3 aware and can host a BFD session. In these cases, BFD sessions can be established among VMs for CC. Also, BFD sessions can be created among VTEPs for tunnel CC. Having a hierarchical OAM model helps localize faults though it requires additional consideration of, for example, coordination of BFD intervals across the OAM layers

Service node reachability:

The service node is responsible for sending BUM traffic. In case a service node tunnel terminates at a VTEP, and that VTEP might not even host VM. BFD session between TOR/hypervisor and service node can be used to monitor service node reachability.

4. Deployment

Figure 1 illustrates the scenario with two servers, each of them hosting two VMs. The servers host VTEPs that terminate two VXLAN tunnels with VNI number 100 and 200 respectively. Separate BFD sessions can be established between the VTEPs (IP1 and IP2) for monitoring each of the VXLAN tunnels (VNI 100 and 200). The implementation SHOULD have a reasonable upper bound on the number of BFD sessions that can be created between the same pair of VTEPs. No BFD packets intended for a Hypervisor VTEP should be forwarded to a VM as a VM may drop BFD packets leading to a false negative. This method is applicable whether the VTEP is a virtual or physical device.

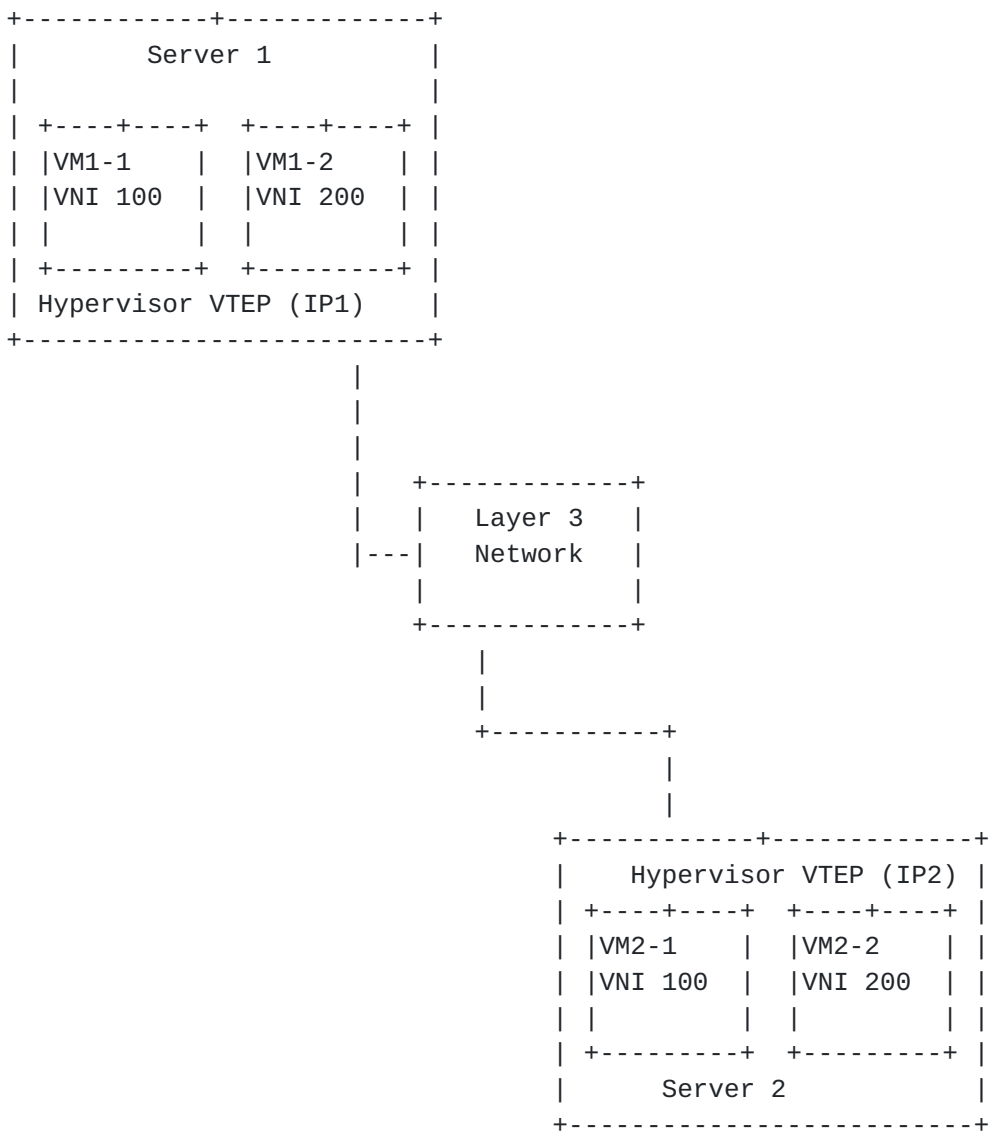

Figure 1: Reference VXLAN domain

5. BFD Packet Transmission over VXLAN Tunnel

BFD packet MUST be encapsulated and sent to a remote VTEP as explained in Section 5.1. Implementations SHOULD ensure that the BFD packets follow the same lookup path as VXLAN data packets within the sender system.

5.1. BFD Packet Encapsulation in VXLAN

BFD packets are encapsulated in VXLAN as described below. The VXLAN packet format is defined in Section 5 of [RFC7348]. The Outer IP/UDP and VXLAN headers MUST be encoded by the sender as defined in [RFC7348].

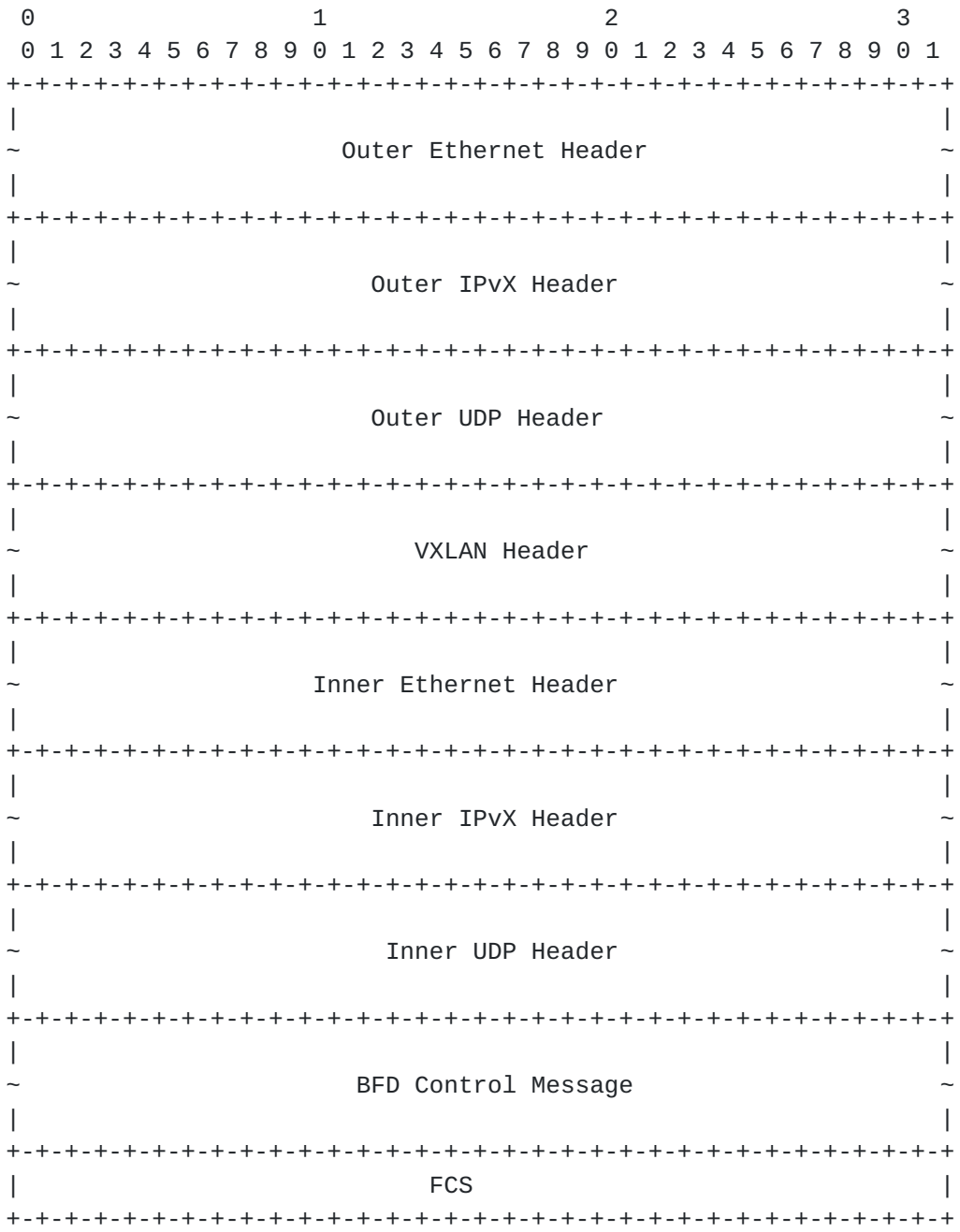

Figure 2: VXLAN Encapsulation of BFD Control Message

The BFD packet MUST be carried inside the inner MAC frame of the VXLAN packet. The inner MAC frame carrying the BFD payload has the following format:

Ethernet Header:

Destination MAC: This MUST be the dedicated MAC TBA (Section 9) or the MAC address of the destination VTEP. The details of how the MAC address of the destination VTEP is obtained are outside the scope of this document.

Source MAC: MAC address of the originating VTEP

IP header:

Source IP: IP address of the originating VTEP.

Destination IP: IP address of the terminating VTEP.

TTL: MUST be set to 1 to ensure that the BFD packet is not routed within the L3 underlay network.

The fields of the UDP header and the BFD control packet are encoded as specified in [RFC5881] for p2p VXLAN tunnels.

6. Reception of BFD packet from VXLAN Tunnel

Once a packet is received, VTEP MUST validate the packet as described in Section 4.1 of [RFC7348]. If the Destination MAC of the inner MAC frame matches the dedicated MAC or the MAC address of the VTEP the packet MUST be processed further.

The UDP destination port and the TTL of the inner IP packet MUST be validated to determine if the received packet can be processed by BFD. BFD packet with inner MAC set to VTEP or dedicated MAC address MUST NOT be forwarded to VMs.

To ensure BFD detects the proper configuration of VXLAN Network Identifier (VNI) in a remote VTEP, a lookup SHOULD be performed with the MAC-DA and VNI as key in the Virtual Forwarding Instance (VFI) table of the originating/terminating VTEP to exercise the VFI associated with the VNI.

6.1. Demultiplexing of the BFD packet

Demultiplexing of IP BFD packet has been defined in Section 3 of [RFC5881]. Since multiple BFD sessions may be running between two VTEPs, there needs to be a mechanism for demultiplexing received BFD

packets to the proper session. The procedure for demultiplexing packets with Your Discriminator equal to 0 is different from [RFC5880]. For such packets, the BFD session MUST be identified using the inner headers, i.e., the source IP, the destination IP, and the source UDP port number present in the IP header carried by the payload of the VXLAN encapsulated packet. The VNI of the packet SHOULD be used to derive interface-related information for demultiplexing the packet. If BFD packet is received with non-zero Your Discriminator, then BFD session MUST be demultiplexed only with Your Discriminator as the key.

7. Use of reserved VNI

In most cases, a single BFD session is sufficient for the given VTEP to monitor the reachability of a remote VTEP, regardless of the number of VNIs in common. When the single BFD session is used to monitor reachability of the remote VTEP, an implementation SHOULD use a VNI of 0.

8. Echo BFD

Support for echo BFD is outside the scope of this document.

9. IANA Considerations

IANA has assigned TBA as a dedicated MAC address from the IANA 48-bit unicast MAC address registry to be used as the Destination MAC address of the inner Ethernet of VXLAN when carrying BFD control packets.

10. Security Considerations

The document requires setting the inner IP TTL to 1 which could be used as a DDoS attack vector. Thus the implementation MUST have throttling in place to control the rate of BFD control packets sent to the control plane. Throttling MAY be relaxed for BFD packets based on port number.

The implementation SHOULD have a reasonable upper bound on the number of BFD sessions that can be created between the same pair of VTEPs.

Other than inner IP TTL set to 1 and limit the number of BFD sessions between the same pair of VTEPs, this specification does not raise any additional security issues beyond those of the specifications referred to in the list of normative references.

11. Contributors

Reshad Rahman
rrahman@cisco.com
Cisco

12. Acknowledgments

Authors would like to thank Jeff Haas of Juniper Networks for his reviews and feedback on this material.

Authors would also like to thank Nobo Akiya, Marc Binderberger, Shahram Davari, Donald E. Eastlake 3rd, and Anoop Ghanwani for the extensive reviews and the most detailed and helpful comments.

13. References

13.1. Normative References

[I-D.ietf-bfd-multipoint]
Katz, D., Ward, D., Networks, J., and G. Mirsky, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint-19 (work in progress), December 2018.

[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>.

[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>.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, DOI 10.17487/RFC5881, June 2010, <https://www.rfc-editor.org/info/rfc5881>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[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>.

13.2. Informational References

[RFC8293] Ghanwani, A., Dunbar, L., McBride, M., Bannai, V., and R. Krishnan, "A Framework for Multicast in Network Virtualization over Layer 3", RFC 8293, DOI 10.17487/RFC8293, January 2018, <https://www.rfc-editor.org/info/rfc8293>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

Authors' Addresses

Santosh Pallagatti (editor)
Rtbrick

Email: santosh.pallagatti@gmail.com

Sudarsan Paragiri
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, California 94089-1206
USA

Email: sparagiri@juniper.net

Vengada Prasad Govindan
Cisco

Email: venggovi@cisco.com

Mallik Mudigonda
Cisco

Email: mmudigon@cisco.com

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com