BFD
Internet-Draft
Intended status: Standards Track
Expires: June 1, 2020

S. Pallagatti, Ed.
VMware
S. Paragiri
Individual Contributor
V. Govindan
M. Mudigonda
Cisco
G. Mirsky
ZTE Corp.
November 29, 2019

BFD for VXLAN
draft-ietf-bfd-vxlan-09

Abstract

This document describes the use of the Bidirectional Forwarding Detection (BFD) protocol in point-to-point Virtual eXtensible Local Area Network (VXLAN) tunnels forming up an overlay network.

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 1, 2020.

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 virtual machines (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 by providing a Layer 2 overlay scheme on a Layer 3 network [RFC7348]. 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 VXLAN Tunnel End Points (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 VTEP.

VTEPs are responsible for encapsulating and decapsulating frames exchanged among VMs.

Ability to monitor path continuity, i.e., perform proactive continuity check (CC) for point-to-point (p2p) VXLAN tunnels, is important. The asynchronous mode of BFD, as defined in [RFC5880], is used to monitor a p2p VXLAN tunnel.

In the case where a Multicast Service Node (MSN) (as described in Section 3.3 of [RFC8293]) resides behind a Network Virtualization Endpoint (NVE), the mechanisms described in this document apply and can, therefore, be used to test the connectivity from the source NVE to the MSN.

This document describes the use of Bidirectional Forwarding Detection (BFD) protocol to enable monitoring continuity of the path between VXLAN VTEPs, performing as Network Virtualization Endpoints, and/or availability of a replicator multicast service node.

2. Conventions used in this document

2.1. Terminology

BFD Bidirectional Forwarding Detection

CC Continuity Check

p2p Point-to-point

MSN Multicast Service Node

NVE Network Virtualization Endpoint

VFI Virtual Forwarding Instance

VM Virtual Machine

VNI VXLAN Network Identifier (or VXLAN Segment ID)

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. 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 VXLAN Network Identifier (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). An implementation that supports this specification MUST be able to control the number of BFD sessions that can be created between the same pair of VTEPs. BFD packets intended for a VTEP MUST NOT 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.

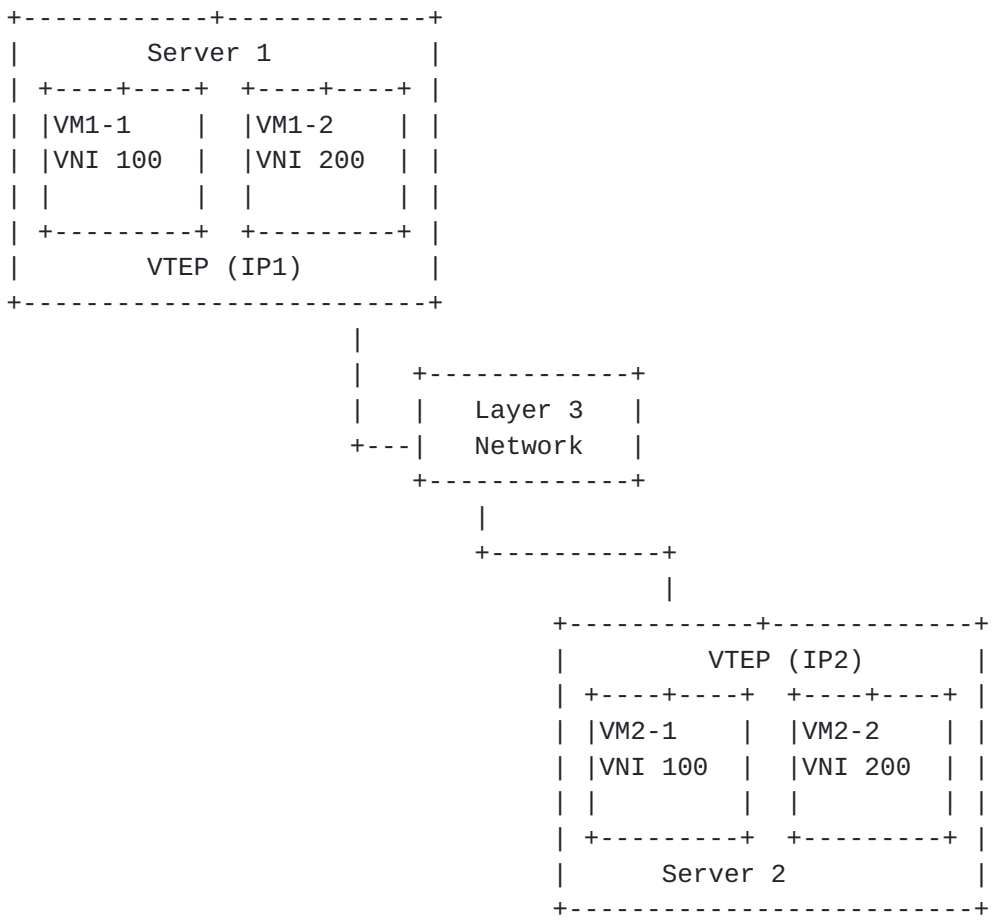

Figure 1: Reference VXLAN Domain

At the same time, a service layer BFD session may be used between the tenants of VTEPs IP1 and IP2 to provide end-to-end fault management. In such case, for VTEPs BFD Control packets of that session are indistinguishable from data packets.

As per Section 4, the inner destination IP address SHOULD be set to one of the loopback addresses (127/8 range for IPv4 and 0:0:0:0:0:FFFF:7F00:0/104 range for IPv6). There could be a firewall configured on VTEP to block loopback addresses if set as the destination IP in the inner IP header. It is RECOMMENDED to allow addresses from the loopback range through a firewall only if it is used as the destination IP address in the inner IP header, and the destination UDP port is set to 3784 [RFC5881].

4. BFD Packet Transmission over VXLAN Tunnel

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

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 Packet

The BFD packet MUST be carried inside the inner Ethernet frame of the VXLAN packet. The choice of Destination MAC and Destination IP addresses for the inner Ethernet frame MUST ensure that the BFD Control packet is not forwarded to a tenant but is processed locally at the remote VTEP. The inner Ethernet frame carrying the BFD Control packet- has the following format:

Ethernet Header:

Destination MAC: This MUST NOT be of one of tenant's MAC addresses. The destination MAC address MAY be the address associated with the destination VTEP. The MAC address MAY be configured, or it MAY be learned via a control plane protocol. The details of how the MAC address is obtained are outside the scope of this document.

Source MAC: MAC address associated with the originating VTEP

IP header:

Destination IP: IP address MUST NOT be of one of tenant's IP addresses. The IP address SHOULD be selected from the range 127/8 for IPv4, for IPv6 - from the range 0:0:0:0:0:FFFF:7F00:0/104. Alternatively, the destination IP address MAY be set to VTEP's IP address.

Source IP: IP address of the originating VTEP.

TTL or Hop Limit: MUST be set to 1 to ensure that the BFD packet is not routed within the Layer 3 underlay network. This addresses the scenario when the inner IP destination address is of VXLAN gateway and there is a router in underlay which removes the VXLAN header, then it is possible to route the packet as VXLAN gateway address is routable address.

The fields of the UDP header and the BFD Control packet are encoded as specified in [RFC5881].

5. Reception of BFD Packet from VXLAN Tunnel

Once a packet is received, VTEP MUST validate the packet. If the Destination MAC of the inner Ethernet frame matches one of the MAC addresses associated with the VTEP the packet MUST be processed further. If the Destination MAC of the inner Ethernet frame doesn't match any of VTEP's MAC addresses, then the processing of the received VXLAN packet MUST follow the procedures described in Section 4.1 [RFC7348]. If the BFD session is using the Management VNI (Section 6), BFD Control packets with unknown MAC address MUST NOT be forwarded to VMs.

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.

5.1. 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. For demultiplexing packets with Your Discriminator equal to 0, a BFD session MUST be identified using the logical link over which the BFD Control packet is received. In the case of VXLAN, the VNI number identifies that logical link. If BFD packet is received with non-zero Your Discriminator, then BFD session MUST be demultiplexed only with Your Discriminator as the key.

6. Use of the Specific 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. When the single BFD session is used to monitor the reachability of the remote VTEP, an implementation SHOULD choose any of the VNIs. An implementation MAY support the use of the Management VNI as control and management channel between VTEPs. The selection of the VNI number of the Management VNI MUST be controlled through management plane. An implementation MAY use VNI number 1 as the default value for the Management VNI. All VXLAN packets received on the Management VNI MUST be processed locally and MUST NOT be forwarded to a tenant.

7. Echo BFD

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

8. IANA Considerations

This specification has no IANA action requested. This section may be deleted before the publication.

9. 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. On the other hand, over-aggressive throttling of BFD Control packets may become the cause of the inability to form and maintain BFD session at scale. Hence, throttling of BFD Control packets SHOULD be adjusted to permit BFD to work according to its procedures.

This document recommends using an address from the Internal host loopback addresses (127/8 range for IPv4 and

0:0:0:0:0:FFFF:7F00:0/104 range for IPv6) as the destination IP address in the inner IP header. Using such address prevents the forwarding of the encapsulated BFD control message by a transient node in case the VXLAN tunnel is broken as according to [RFC1812]:

> A router SHOULD NOT forward, except over a loopback interface, any packet that has a destination address on network 127. A router MAY have a switch that allows the network manager to disable these checks. If such a switch is provided, it MUST default to performing the checks.

If the implementation supports establishing multiple BFD sessions between the same pair of VTEPs, there SHOULD be a mechanism to control the maximum number of such sessions that can be active at the same time.

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.

10. Contributors

Reshad Rahman
rrahman@cisco.com
Cisco

11. 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.

12. References

12.1. Normative References

[RFC1812] Baker, F., Ed., "Requirements for IP Version 4 Routers", RFC 1812, DOI 10.17487/RFC1812, June 1995, <https://www.rfc-editor.org/info/rfc1812>.

[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>.

12.2. 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)
VMware

Email: santosh.pallagatti@gmail.com

Sudarsan Paragiri
Individual Contributor

Email: sudarsan.225@gmail.com

Vengada Prasad Govindan
Cisco

Email: venggovi@cisco.com

Mallik Mudigonda
Cisco

Email: mmudigon@cisco.com

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com