BFD
Internet-Draft
Intended status: Standards Track
Expires: November 26, 2020

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

May 25, 2020

BFD for VXLAN
draft-ietf-bfd-vxlan-12

Abstract

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

Status of This Memo

Copyright Notice

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.

The 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]) participates in VXLAN, 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 MSN using a Management VNI (Section 4). All other uses of the specification to test toward other VXLAN endpoints are out of the scope.

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). Using a BFD session to monitor a set of VXLAN VNIs between the same pair of VTEPs might help to detect and localize problems caused by misconfiguration. 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.

```
      +------------+-------------+
      |        Server 1          |
      | +----+----+  +----+----+ |
      | |VM1-1    |  |VM1-2    | |
      | |VNI 100  |  |VNI 200  | |
      | |         |  |         | |
      | +---------+  +---------+ |
      |        VTEP (IP1)        |
      +--------------------------+
                            |
                            |   +-------------+
                            |   |   Layer 3   |
                            +---|   Network   |
                                +-------------+
                                    |
                                    +-----------+
                                                |
                                         +------------+-------------+
                                         |         VTEP (IP2)       |
                                         | +----+----+  +----+----+ |
                                         | |VM2-1    |  |VM2-2    | |
                                         | |VNI 100  |  |VNI 200  | |
                                         | |         |  |         | |
                                         | +---------+  +---------+ |
                                         |      Server 2            |
                                         +--------------------------+
```

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 (this use case is outside the scope of this document). In such a case, for VTEPs BFD Control packets of that session are indistinguishable from data packets.

For BFD Control packets encapsulated in VXLAN (Figure 2), the inner destination IP address SHOULD be set to one of the loopback addresses from 127/8 range for IPv4 or to one of IPv4-mapped IPv4 loopback addresses from ::ffff:127.0.0.0/104 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 they are used as the destination IP addresses in the inner IP header and the destination UDP port is set to 3784 [RFC5881].

4. Use of the Management 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 that supports this specification MUST 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 a 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.

5. BFD Packet Transmission over VXLAN Tunnel

BFD packets MUST be encapsulated and sent to a remote VTEP as explained in this section. Implementations SHOULD ensure that the BFD packets follow the same forwarding 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 value in the VNI field of the VXLAN header MUST be set to the value selected as the Management VNI. The Outer IP/UDP and VXLAN headers MUST be encoded by the sender as defined in [RFC7348].

```
     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
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                      Outer Ethernet Header                    ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                        Outer IPvX Header                      ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                        Outer UDP Header                       ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                           VXLAN Header                        ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                    Inner Ethernet Header                      ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                        Inner IPvX Header                      ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                         Inner UDP Header                      ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                       BFD Control Packet                     ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                        Outer  FCS                             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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: since a Management VNI is the VNI that does not have any tenants, the value of this field is not analyzed by the receiving VTEP.

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 ::ffff:127.0.0.0/104. 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 255 in accordance with the Generalized TTL Security Mechanism [RFC5082].

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

6. Reception of BFD Packet from VXLAN Tunnel

Once a packet is received, the VTEP MUST validate the packet. If the packet is received on the management VNI and is identified as BFD control packet addressed to the VTEP, and then the packet can be processed further. Processing of BFD control packets received on non-management VNI is outside the scope of this specification.

Validation of TTL / Hop Limit of the inner IP packet, as long as the related considerations for BFD control packet demultiplexing and authentication, is performed as described in Section 5 [RFC5881].

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

Security issues discussed in [RFC5880], [RFC5881], and [RFC7348] apply to this document.

This document recommends using an address from the Internal host loopback addresses 127/8 range for IPv4 or an IP4-mapped IPv4 loopback address from ::ffff:127.0.0.0/104 range for IPv6 as the destination IP address in the inner IP header. Using such an 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.

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, Anoop Ghanwani, Dinesh Dutt, Joel Halpern, and Carlos Pignataro for the extensive reviews and the most detailed and constructive 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>.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <https://www.rfc-editor.org/info/rfc5082>.

[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