Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 16, 2018

S. Pallagatti, Ed.
Independent Contributor
S. Paragiri
Juniper Networks
V. Govindan
M. Mudigonda
Cisco
G. Mirsky
ZTE Corp.

October 13, 2017

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

Abstract

This document describes use of Bidirectional Forwarding Detection (BFD) protocol in Virtual eXtensible Local Area Network (VXLAN) 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 April 16, 2018.

Table of Contents

1. Introduction

"Virtual eXtensible Local Area Network (VXLAN)" has been described in [RFC7348]. VXLAN provides an encapsulation scheme that allows virtual machines (VMs) to communicate in a data center network.

VXLAN is typically deployed in data centers interconnecting virtualized hosts, which may be spread across multiple racks. The individual racks may be part of a different Layer 3 network or they could be in a single Layer 2 network. The VXLAN segments/overlay networks are overlaid on top of these Layer 2 or Layer 3 networks.

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

Since underlay is a L3 network, ability to monitor path continuity, i.e. perform proactive continuity check (CC) for these tunnels is important. 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 special service nodes that are designated to properly handle Layer 2 broadcast, unknown unicast, and multicast traffic. Such nodes, often referred "replicators", are usually virtual VTEPs can be monitored by physical VTEPs in order to minimize BUM traffic directed to unavialable replicator.

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

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

Main 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 in 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

Layer 2 VMs:

Most deployments will have VMs with only L2 capabilities that may not support L3. BFD being a L3 protocol can be used as 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. In rest of this document terms NVE and VTEP are used interchangeably.

Fault localization:

It is also possible that VMs are L3 aware and can possibly host a BFD session. In these cases BFD sessions can be established among VMs for CC. In addition, BFD sessions can be established among VTEPs for tunnel CC. Having a hierarchical OAM model helps localize faults though requires additional consideration.

Service node reachability:

Service node is responsible for sending BUM traffic. In case of service node tunnel terminates at VTEP and it 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. These servers host VTEPs that terminate two VXLAN tunnels with VNI number 100 and 200. Separate BFD sessions can be established between the VTEPs (IP1 and IP2) for monitoring each of the VXLAN tunnels (VNI 100 and 200). No BFD packets, intended to Hypervisor VTEP, should be forwarded to a VM as VM may drop BFD packets leading to false negative. This method is applicable whether VTEP is a virtual or physical device.

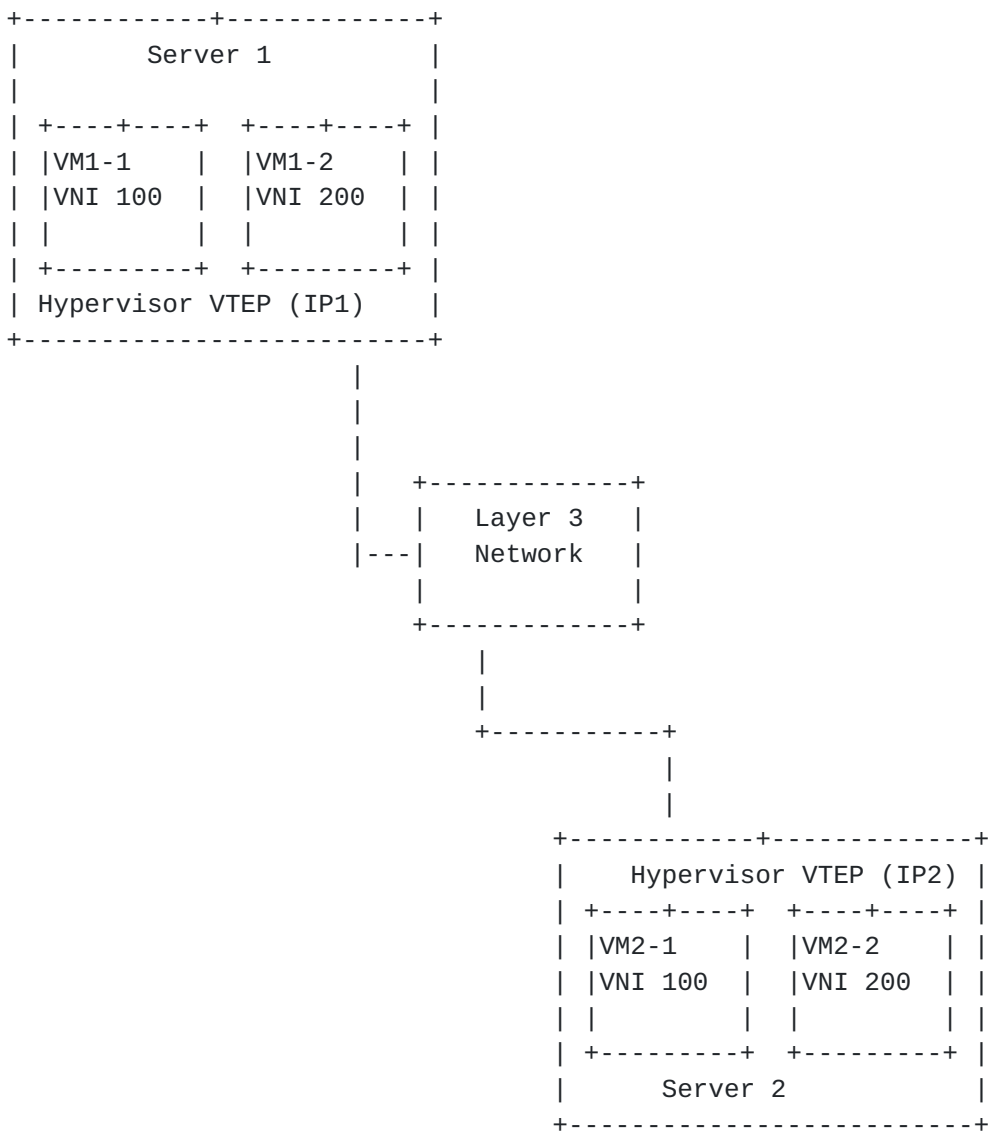

Figure 1: Reference VXLAN domain

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 of VXLAN packets within the sender system.

5.1. BFD Packet Encapsulation in VXLAN

VXLAN packet format has been described in Section 5 of [RFC7348]. The Outer IP/UDP and VXLAN headers MUST be encoded by the sender as per [RFC7348].

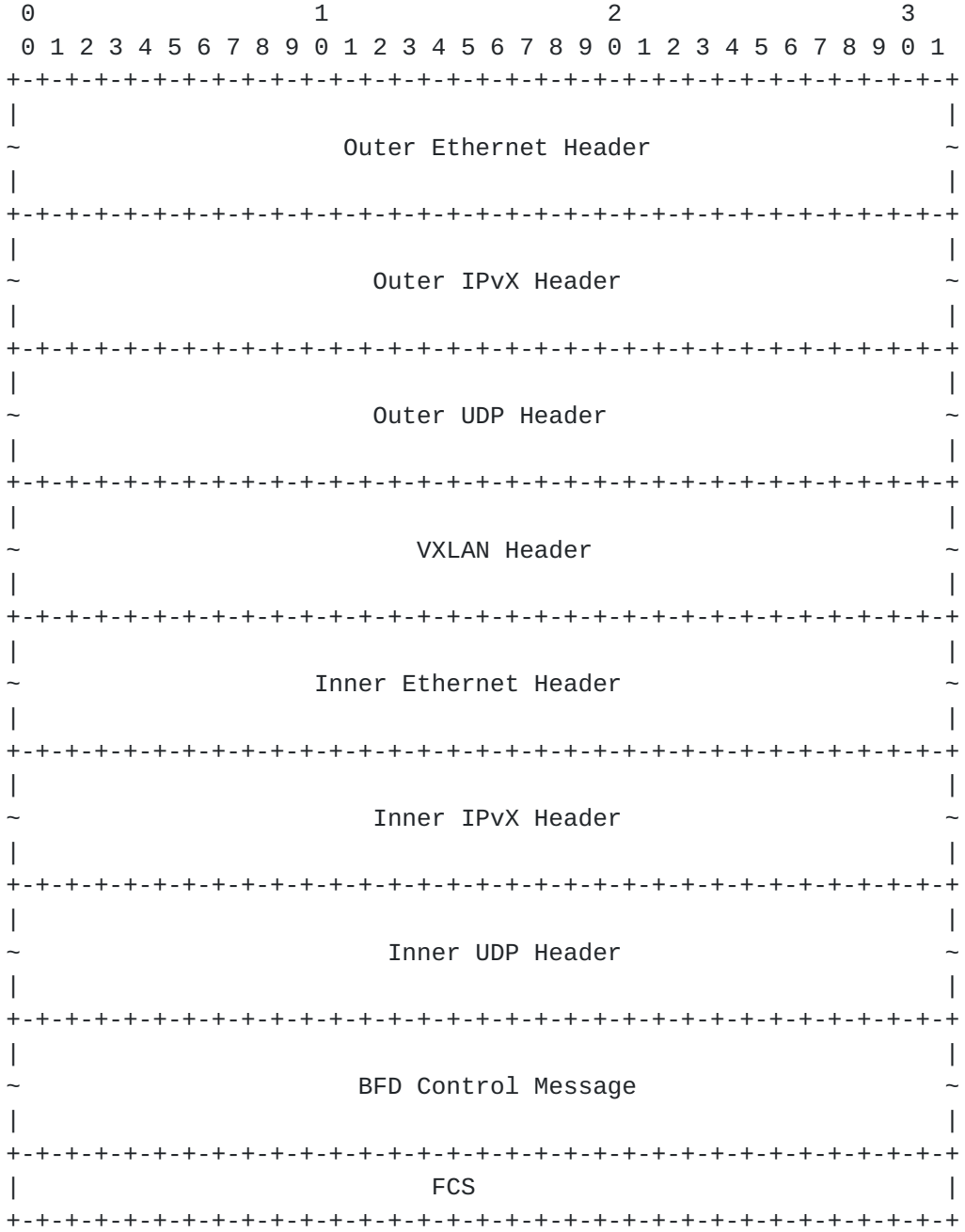

Figure 2: VXLAN Encapsulaion of BFD Control Message

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

[Ed.Note]:Use of inner source and destination IP addresses needs more discussion by the WG.

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 Ethernet frame 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 in order 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 and the destination IP 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

BFD session MAY be established for the reserved VNI 0. One way to aggregate BFD sessions between VTEP's is to establish a BFD session with VNI 0. A VTEP MAY also use VNI 0 to establish a BFD session with a service node.

8. Echo BFD

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

9. IANA Considerations

IANA is requested to assign a dedicated MAC address to be used as the Destination MAC address of the inner Ethernet which carries BFD control packet in IP/UDP encapsulation.

10. Security Considerations

Document recommends setting of inner IP TTL to 1 which could lead to DDoS attack, implementation MUST have throttling in place. Throttling MAY be relaxed for BFD packets based on port number.

Other than inner IP TTL set to 1 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 Hass of Juniper Networks for his reviews and feedback on this material.

Authors would also like to thank Nobo Akiya, Marc Binderberger and Shahram Davari for the extensive review.

13. Normative References

[I-D.ietf-bfd-multipoint] Katz, D., Ward, D., and J. Networks, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint-10 (work in progress), April 2017.

[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>.

Authors' Addresses

Santosh Pallagatti (editor)
Independent Contributor

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