Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2016

S. Pallagatti, Ed.
B. Saji
S. Paragiri
Juniper Networks
V. Govindan
M. Mudigonda
Cisco
G. Mirsky
Ericsson

July 6, 2015

BFD for VXLAN
draft-spallagatti-bfd-vxlan-01

Abstract

This document describes use of Bidirectional Forwarding Detection (BFD) protocol for VXLAN . Comments on this draft should be directed to nvo3@ietf.org, rtg-bfd@ietf.org.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 http://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 January 7, 2016.

Table of Contents

1. Introduction 2
2. Use cases 3
3. Deployment 4
4. BFD Packet Encapsulation 5
5. Reception of BFD packet 5
5.1. Demux of the BFD packet 6
6. Echo BFD 6
7. IANA Considerations 6
8. Security Considerations 6
9. Contributors 6
10. Acknowledgements 6
11. Normative References 7
Authors' Addresses 7

1. Introduction

"Virtual eXtensible Local Area Network (VXLAN)" has been defined in [RFC7348] that provides an encapsulation scheme which allows VM's to communicate in data centre network.

VXLAN is typically deployed in data centres on virtualized hosts, which may be spread across multiple racks. The individual racks may be parts 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 a VM in other host only if they are on same VXLAN. VM's are unaware of VXLAN tunnels as VXLAN tunnel terminates on VTEP (hypervisor/TOR). VTEP (hypervisor/TOR) are responsible for encapsulating and decapsulating frames sent from VM's.

Since underlay is a L3 network, connectivity check for these tunnels becomes important. BFD as defined in [RFC5880] can be used to monitor the VXLAN tunnels.

This draft addresses requirements outlined in [I-D.ashwood-nvo3-operational-requirement]. Specifically with reference to the OAM model to Figure 3 of [I-D.ashwood-nvo3-operational-requirement], this draft outlines proposal to implement the OAM mechanism between the NV Edges using BFD.

2. Use cases

Main use case of BFD for VXLAN is for tunnel connectivity check. There are other use cases such as

Layer 2 VM's:

Most deployments will have VM's with only L2 capabilities and may not understand L3. BFD being a L3 protocol can be used for tunnel connectivity check, where BFD will start and terminate at the NV Edge (VTEPs).

It is possible to aggregate the connectivity checks for multiple tenants by running a BFD session between the VTEPs over VxLAN tunnel. In rest of this document terms NV Edge and VTEP are used interchangeably.

Fault localization:

It is also possible that VM's are L3 aware and can possibly host a BFD session. In these cases BFD sessions can be established between VM's for connectivity check. In addition a BFD session can be established between VTEPs for tunnel connectivity check. Having a hierarchical OAM model helps localize faults.

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's. If TOR's/Hypervisor wants to check service node reachability then it would like run BFD session over VXLAN tunnel to service node.

3. Deployment

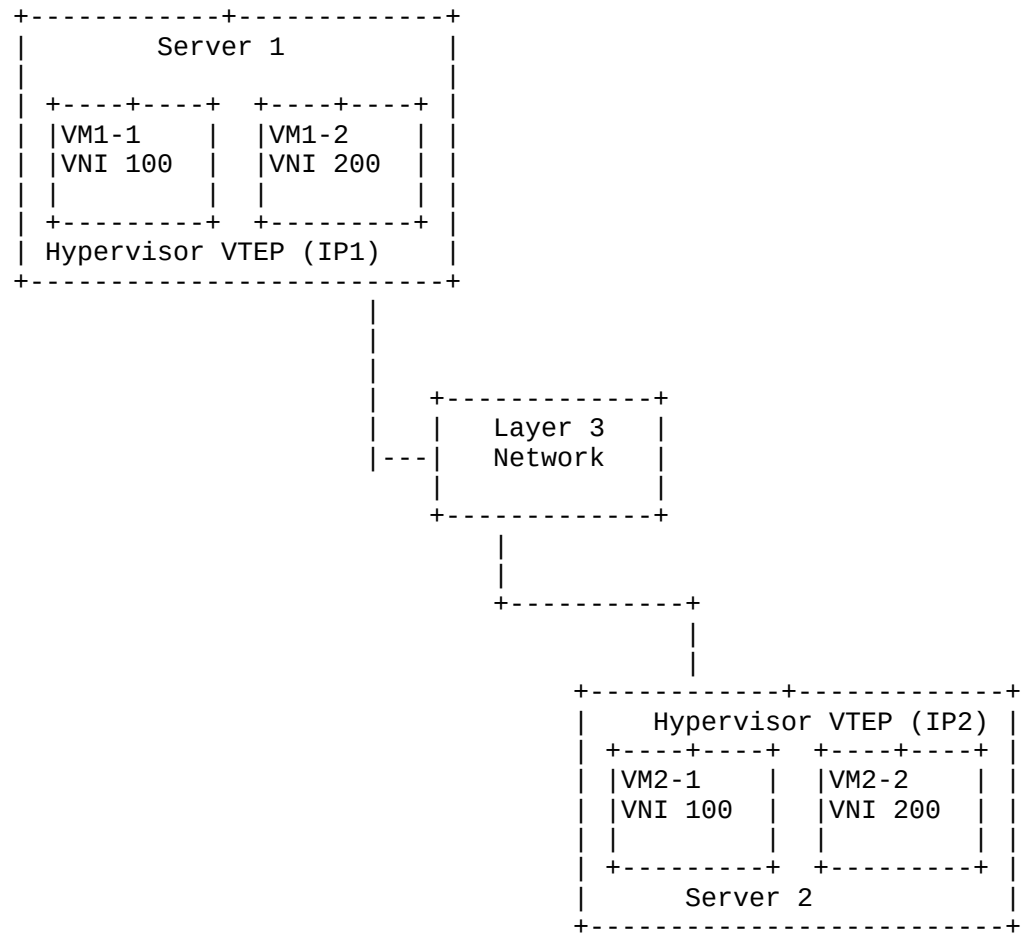

Figure 1

Figure 1 illustrates a scenario where we have two servers, each of them hosting two VMs. These VTEPs terminate two VXLAN tunnels with VNI number 100 and 200 between them. 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 packet intended to Hypervisor VTEP should be forwarded to VM's as VM's may drop this leading to false negative. This method is also applicable VTEP which are either software or physical device.

4. BFD Packet Encapsulation

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

If VTEP is equipped with GPE header capitalises and decides to use GPE instead of VXLAN then GPE header MUST be encoded as per Section 3.3 of [I-D.quinn-vxlan-gpe]. Next Protocol Field in GPE header MUST be set to IPv4 or IPv6.

Details of how VTEP decides to use VXLAN or GPE header is outside the scope of this document.

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 well-known MAC [TBD] OR the MAC address of the destination VTEP. The details of how the destination MAC address is obtained is 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.

Note: Inner source and destination IP needs more discussion in WG.

The fields of the UDP header and the BFD control packet are encoded as specified in RFC 5881.

5. Reception of BFD packet

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 well-known MAC or the MAC address of the VTEP the packet MUST be processed further.

The UDP destination port and the TTL of the inner MAC frame MUST be validated to determine if the received packet can be processed by BFD. BFD packet with inner MAC set to VTEP or well-known MAC address MUST not be forwarded to VM's.

5.1. Demux of the BFD packet

Demux 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 = 0 is different from [RFC5880]. For such packets, the BFD session is identified using the VNID, the source IP and the destination IP of the packet. If BFD packet is received with non-zero your discriminator then BFD session should be demultiplexed only with your discriminator as the key.

6. Echo BFD

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

7. IANA Considerations

The well-known MAC to be used for the Destination MAC address of the inner MAC frame needs to be defined

8. 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 packeted 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.

9. Contributors

Reshad Rahman
rrahman@cisco.com
Cisco

10. Acknowledgements

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

11. Normative References

[I-D.ashwood-nvo3-operational-requirement]
Ashwood-Smith, P., Iyengar, R., Tsou, T., Sajassi, A., Boucadair, M., Jacquenet, C., and M. Daikoku, "NVO3 Operational Requirements", draft-ashwood-nvo3-operational-requirement-03 (work in progress), July 2013.

[I-D.ietf-bfd-seamless-base]
Akiya, N., Pignataro, C., Ward, D., Bhatia, M., and J. Networks, "Seamless Bidirectional Forwarding Detection (S-BFD)", draft-ietf-bfd-seamless-base-05 (work in progress), June 2015.

[I-D.quinn-vxlan-gpe]
Quinn, P., Manur, R., Kreeger, L., Lewis, D., Maino, F., Smith, M., Agarwal, P., Yong, L., Xu, X., Elzur, U., Garg, P., and D. Melman, "Generic Protocol Extension for VXLAN", draft-quinn-vxlan-gpe-04 (work in progress), February 2015.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

[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, August 2014.

Authors' Addresses

Santosh Pallagatti (editor)
Juniper Networks
Embassy Business Park
Bangalore, KA 560093
India

Email: santoshpk@juniper.net

Basil Saji
Juniper Networks
Embassy Business Park
Bangalore, KA 560093
India

Email: sbasil@juniper.net

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
Ericsson

Email: gregory.mirsky@ericsson.com