Abstract

This document describes use of Bidirectional Forwarding Detection (BFD) protocol for VXLAN. Comments on this draft should be directed to nvo3@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 November 5, 2015.

Copyright Notice

Copyright (c) 2015 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 (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents.
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). VETP (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.
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 aware and may not understand L3. BFD being a L3 protocol can be used for tunnel connectivity check, where BFD will start and terminate at VTEP on different host.

Fault localization:

It is also possible that VM’s are L3 aware and can possible host a BFD session. In these cases BFD session can be used to run between VM’s for VM’s connectivity check, also BFD session between VTEP’s for tunnel connectivity check. With both VM’s BFD session and tunnel BFD can easily localize the fault to either VM or tunnel.

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
Consider the above diagram, where we have two servers with IP1 and IP2 and each of them are hosting two VM’s. There are two VXLAN tunnels with VNI number 100 and 200. For connectivity check of these two VXLAN tunnels, BFD sessions needs to be established per tunnel. In the diagram above two BFD will be established between server 1’s Hypervisor VTEP (IP1) and server2 Hypervisor VTEP(IP2). BFD session will originate from Hypervisor VTEP (IP1) and terminate at Hypervisor VTEP (IP2) and visa versa. Each BFD session on Hypervisor VTEP will be identified by its VNI in VXLAN header. 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. Packet Format

Packet format has been defined in Section 5 of [RFC7348]. Outer IP/UDP and VXLAN header will remain same and they should be filled by sending VTEP as per [RFC7348]. Inner packet format has been defined as below for BFD packet which terminates at VETP. BFD packet MUST have a inner IP/UDP header followed by BFD payload.

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

Inner IPv4 Header:

+-----------------------------------------------+
|Version|IHL|Type of Service|Total Length|
+-----------------------------------------------+
|Identification|Flags|Fragment Offset|
+-----------------------------------------------+
|TTL = 1|Protocol=17(UDP)|Header Checksum|
+-----------------------------------------------+
|Inner Source IPv4 Address|
+-----------------------------------------------+
|Inner Destination IPv4 Address = 127/8 address|
+-----------------------------------------------+

Inner IPv6 Header:

+-----------------------------------------------+
|Version|Traffic Class|Flow Label|
+-----------------------------------------------+
|Payload Length|NxtHdr=17(UDP)|Hop Limit = 1|
+-----------------------------------------------+
|Inner Source IPv6 Address|
+-----------------------------------------------+
|Inner Destination IPv6 Address = 0:0:0:0:FFFF:7F00/104|
+-----------------------------------------------+
Inner UDP Header:

```
+-----------------------------+
|           Source Port       |       Dest Port = 3784 |
| UDP Length                  |        UDP Checksum    |
+-----------------------------+
```

BFD packet:

```
+-----------------------------+
| Vers |  Diag   |Sta|P|F|C|A|D|M|  Detect Mult  |    Length     |
| My Discriminator            |
| Your Discriminator          |
| Desired Min TX Interval     |
| Required Min RX Interval    |
| Required Min Echo RX Interval|
+-----------------------------+
```

5. Transmission of BFD Packet:

This section describes BFD packet encapsulation while transmitting BFD packet from VTEP

Outer IP/UDP and VXLAN header:

VTEP which is transmitting the packet MUST encapsulate the BFD packet in outer IP/UDP and VXLAN header as described in Section 5 of [RFC7348]. IP TOS value MUST be set to "Internetwork Control".

Inner IP/UDP header:

Source IP address:

MUST be set to outgoing interface of sending VTEP interface.

Destination IP address:
MUST be set to non-routable 127/8 or 0:0:0:0:0:FFFF:7F00/104 range address.

IP TTL/HOP LIMIT:

MUST be set to 1.

UDP header:

UDP header is set as defined in Section 4 of [RFC5881]

BFD Packet:

BFD packet SHOULD be constructed as defined in Section 4 of [RFC5880].

6. Reception of BFD Packet:

Once a packet is received VTEP MUST validate packet as described in Section 4.1 of [RFC7348]. Since inner IP TTL is set to 1 packet SHOULD be consumed by VTEP and should not be forwarded further to VM. It is recommended that BFD packets should not be throttled with TTL 1. Implementation MAY have a check to relax throttling if the inner IP address is 127/8 range for IPv4 and 0:0:0:0:0:FFFF:7F00/104 for IPv6 then UDP destination port is 3784.

6.1. Demux of BFD Packet:

Demux of IP BFD packet has been defined in Section 3 of [RFC5881]. BFD demultiplexing for VXLAN is going to be different as destination IP is same for all sessions and underlay is layer 3 and may have ECMP. Source address and VNI should identify a BFD session on VTEP, initially when BFD packets are sent with with your discriminator set to 0 BFD packets MUST be demultiplexed with source address and VNI as the key. If BFD packet is received with non-zero your discriminator then BFD session should be demultiplexed only with your discriminator as the key.

7. Echo BFD:

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

8. S-BFD:

S-BFD can also be used for connectivity check as defined in [I-D.ietf-bfd-seamless-base]
8.1. Transmission of S-BFD:

VTEP MUST encapsulate S-BFD packet as defined in Section 5. For S-BFD however your Discriminator will be set to VNI from the VXLAN header.

8.2. Reception of S-BFD:

VTEP MUST decapsulate S-BFD packet as defined in above section "reception of BFD packet". Reflector MUST validate if your Discriminator belongs any one of the VNI on that VTEP.

9. IANA Considerations

This document has no actions for IANA.

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 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.

11. Acknowledgements

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

12. Normative References

[I-D.ietf-bfd-seamless-base]


Authors’ Addresses

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

Email: santoshpk@juniper.net

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

Email: sparagiri@juniper.net

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

Email: sbasil@juniper.net