Network Working Group
Internet Draft
Intended status: Informational
Expires: November 4, 2018

V. Ji
Cisco Systems, Inc.

May 4, 2018

E-VPN Ping Mechanism for Virtual eXtensible Local Area Network (VXLAN)
draft-vji-evpn-ping-vxlan-00.txt

Abstract

Ping is a widely deployed Operation, Administration, and Maintenance (OAM) mechanism in networks. This document describes a mechanism for detecting data-plane failures using Ping in RFC7348 VXLAN based EVPN networks.

Table of Contents

1. Introduction

RFC7348 Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks defines means to support data center layer 2 E-VPN over an IP core network.

draft-jain-bess-evpn-lsp-ping defines procedures to detect data-plane failures using LSP Ping in MPLS networks deploying EVPN and PBB-EVPN.

This document outlines how OAM data fields are encapsulated and how connectivity check and fault isolation is performed from edge to edge for VXLAN networks.

2. Conventions used in this document

In examples, "C:" and "S:" indicate lines sent by the client and server respectively.

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying significance described in RFC 2119.

In this document, the characters ">>" preceding an indented line(s) indicates a statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the portions of this RFC covered by these keywords.

3. Acronyms and Definitions

AD	Auto Discovery
CE	Customer Edge Device
ECMP	Equal-Cost Multipath
ESI	Ethernet Segment Identifier
EVPN	Ethernet Virtual Private Network
OAM	Operations, Administration and Maintenance
PE	Provider Edge Device
VLAN	Virtual Local Area Network
VNI	VXLAN Network Identifier (or VXLAN Segment ID)
VTEP	VXLAN Tunnel End Point. An entity that originates and/or terminates VXLAN tunnels
VXLAN	Virtual eXtensible Local Area Network
VXLAN Segment	VXLAN Layer 2 overlay network over which VMs communicate
VXLAN Gateway	an entity that forwards traffic between VXLANs

4. IP ping and trace route extension for VXLAN

In IP network ICMP, UDP or HTTP based ping and traceroute provide ways to perform reachability check and fault isolation, this can be used for OAM purpose for the IP underlay network. E-VPN extension for the existing ping and traceroute operations make it control-plane aware and add additional capability to validate the E-VPN forwarding context, detect data-plane errors and measure PE to PE performance.

5. VXLAN OAM header format

IPv4 underlay OAM information is encoded in the VXLAN header as below.

```
VXLAN Header
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|R|R|R|O|I|R|R|R|            Reserved                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                VXLAN Network Identifier (VNI) |   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

RFC7348 VXLAN header OAM extension

New O bit is selected for OAM purpose, value 1 for OAM packets, 0 for regular VXLAN traffic. This bit is temporarily declared as bit3, subject to be changed

5.1. VXLAN EVPN OAM Header:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Option Class = OAM_ECHO      |    Type       |R|R|R| Length  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Return Code  | Return Subcode|          Must Be Zero         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Sender's Handle                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Sequence Number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    TimeStamp Sent (seconds)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                TimeStamp Sent (seconds fraction)              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  TimeStamp Received (seconds)                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              TimeStamp Received (seconds fraction)            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
.                                                               .
.                             TLVs                              .
.                                                               .
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

VXLAN EVPN OAM header

Type: 0 for echo request; 1 for echo reply.

Return Code and Return sub-code must be zero for the ping or traceroute request. For ping or traceroute reply, the value is defined as:

Return Code #	Value Field
0	Success
1	Context Not Found
2	Context Found but IP address Mis-Match

Return sub-code is reserved for future use.

The Sender's Handle is filled in by the sender and returned unchanged by the receiver in the echo reply (if any). There are no

semantics associated with this handle, although a sender may find this useful for matching up requests with replies.

The Sequence Number is assigned by the sender of the echo request and can be (for example) used to detect missed replies.

The TimeStamp Sent is the time of day (according to the sender's clock) in 64-bit NTP timestamp format [RFC5905] when the echo request is sent. The TimeStamp Received in an echo reply is the time of day (according to the receiver's clock) in 64-bit NTP timestamp format in which the corresponding echo request was received. TimeStamp Received must be zero for the request. Value 0 means the time is not measured or available, shall be ignored.

TLVs (Type-Length-Value tuples) have the following format:

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |             Type              |            Length             |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                             Value                             |
   .                                                               .
   .                                                               .
   .                                                               .
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                                     TLV format
```

TLV type values use the same value of corresponding BGP route type when advertised the route, defined in [RFC7432], [draft-ietf-bess-evpn-prefix-advertisement-04] and [draft-ietf-bess-evpn-igmp-mld-proxy].

Type #	Value Field
1	Ethernet Auto-Discovery (A-D) TLV
2	MAC/IP TLV
3	Inclusive Multicast TLV
4	Ethernet Segment TLV (format to be defined)
5	IP Prefix TLV
6	Selective Multicast Ethernet Tag TLV (format to be defined)

5.2. EVPN MAC/IP TLV

The EVPN MAC/IP TLV is used to identify the MAC for an EVI under test at a peer PE.

The EVPN MAC TLV fields are derived from the MAC/IP advertisement route defined in [RFC7432] Section 7.2 and has the format as shown in Figure 4. This TLV is included in the Echo Request sent to the Peer PE by the PE that is the originator of the request.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                     Route Distinguisher                       |
   |                         (8 octets)                            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                      Ethernet Tag ID                          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                Ethernet Segment Identifier                    |
   |                      (10 octets)                              |
   +                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                               | must be zero  |  MAC Addr Len |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 MAC Address                                   |
   +                  (6 Octets)   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                               | Must be zero  |  IP Addr Len  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 IP Address (0, 4 or 16 Octets)                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    L2VNI (3 Octets)           |   Reserved    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    L3VNI (3 Octets)           |   Reserved    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

EVPN MAC TLV format

The ping echo request is sent using the EVPN VNI(s) associated with the MAC route announced by a remote PE to reach the remote PE.

5.3. EVPN Inclusive Multicast TLV

The EVPN Inclusive Multicast sub-TLV fields are based on the EVPN Inclusive Multicast route defined in [RFC7432] Section 7.3. The EVPN Inclusive Multicast TLV has the format as shown in Figure 5. This TLV is included in the echo request sent to the EVPN peer PE by the originator of request to verify the multicast connectivity state on the peer PE(s) in EVPN and PBB-EVPN.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Route Distinguisher                        |
   |                        (8 octets)                             |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                     Ethernet Tag ID                           |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | IP Addr Len |                                                 |
   +-+-+-+-+-+-+-+                                                 |
   ~               Originating Router's IP Addr                    ~
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                     VNI (3 Octets)            |   Reserved    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                        EVPN Inclusive Multicast TLV format
```

Broadcast, multicast and unknown unicast traffic can be sent using ingress replication or P2MP P-tree in EVPN network.

5.4. EVPN Auto-Discovery TLV

The EVPN Auto-Discovery (AD) TLV fields are based on the Ethernet AD route advertisement defined in [RFC7432] Section 7.1. EVPN AD TLV applies to only EVPN. The EVPN AD sub-TLV has the format shown in Figure 6.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Route Distinguisher                        |
   |                        (8 octets)                             |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                     Ethernet Tag ID                           |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               Ethernet Segment Identifier                     |
   |                     (10 octets)                               |
   +                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                               |         must be zero          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                   VNI (3 Octets)              |   Reserved    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                           EVPN Auto-Discovery TLV format
```

5.5. EVPN IP Prefix TLV

The EVPN IP Prefix TLV is used to identify the IP Prefix for an EVI under test at a peer PE. The EVPN IP Prefix sub-TLV fields are derived from the IP Prefix Route (RT-5) advertisement defined in [I-D.ietf-bess-evpn-prefix-advertisement] and has the format as shown in Figure 7. This TLV is included in the Echo Request sent to the Peer PE by the PE that is the originator of the request.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Route Distinguisher                       |
|                         (8 octets)                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Ethernet Tag ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Ethernet Segment Identifier                    |
|                      (10 octets)                              |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               | must be zero  | IP Prefix Len |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                  IP Prefix  (4 or 16 Octets)                  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                 GW IP Address (4 or 16 Octets)                ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   L3VNI (3 Octets)            |   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

EVPN IP Prefix TLV format

6. E-VPN Context Validation procedure

The TLVs in the EVPN OAM header is collect from the control-plane of the ping or traceroute initiator PE, and to be validated by control-plane of the peer PE, mid-node transmit routers may ignore it. For traceroute, when the packet is punted to OAM for each TTL expiry event, transmitter router may update the TimeStamp field in the header to provide performance measurement.

This procedure do not have preference of protocol selection of ping or trace route. Typically, ICMP echo request and ICMP echo reply is used for ping; while ICMP echo request, UDP, HTTP or other protocols may be used for traceroute. There is no change to these upper level protocols.

7. Security Considerations

The proposal introduced in this document does not introduce any new security considerations beyond that already apply to [RFC7432], [RFC7348], [RFC7623] and [RFC6425] and draft-jain-bess-evpn-lsp-ping.

8. IANA Considerations

8.1. Sub-TLV Type

This document defines 6 new TLV types, which is intend to use the same value as RT types defined in .

IANA is requested to assign a sub-TLV type value to the following

8.2. Proposed new Return Codes

[RFC8029] defines values for the Return Code field of Echo Reply. This document proposes two new Return Codes, which SHOULD be included in the Echo Reply message by a PE in response to LSP Ping Echo Request message:

1. The FEC exists on the PE and the behavior is to drop the packet because of not DF.

2. The FEC exists on the PE and the behavior is to drop the packet because of Split Horizon Filtering.

9. References

9.1. Normative References

[RFC7348] M. Mahalingam, Storvisor, D. Dutt, K. Duda, P. Agarwal, L. Kreeger, T. Sridhar, M. Bursell, C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[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>.

[I-D.ietf-bess-evpn-prefix-advertisement] Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-09 (work in progress), November 2017.

[draft-ietf-bess-evpn-igmp-mld-proxy] Ali Sajassi, Samir Thoria, Keyur Patel, Derek Yeung, John Drake and Wen Lin "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-00 (work in progress), March 2017.

[RFC6425] Saxena, S., Ed., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", RFC 6425, DOI 10.17487/RFC6425, November 2011, <https://www.rfc-editor.org/info/rfc6425>.

[RFC6426] Gray, E., Bahadur, N., Boutros, S., and R. Aggarwal, "MPLS On-Demand Connectivity Verification and Route Tracing", RFC 6426, DOI 10.17487/RFC6426, November 2011, <https://www.rfc-editor.org/info/rfc6426>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[I-D.ietf-evpn-lsp-ping] P. Jain, Ed., S. Salam, A. Sajassi, S. Boutros and G. Mirsky, "LSP-Ping Mechanisms for EVPN and PBB-EVPN", draft-jain-bess-evpn-lsp-ping-06 (work in progress), January, 2018

[RFC5905] D. Mills, J. Martin, Ed., J. Burbank, W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", June 2010, <https://www.rfc-editor.org/info/rfc5905>.

9.2. Informative References

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <https://www.rfc-editor.org/info/rfc7365>.

[RFC792] J. Postel, "INTERNET CONTROL MESSAGE PROTOCOL", RFC792, September 1981, <https://www.rfc-editor.org/info/rfc792>.

10. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Victor Ji
Cisco Systems, Inc.
Email: vji@cisco.com