EVPN VXLAN Bypass VTEP
<draft-eastlake-bess-evpn-vxlan-bypass-vtep-01.txt>

Abstract

   A principal feature of EVPN is the ability to support multihoming
   from a customer equipment (CE) to multiple provider edge equipment
   (PE) with all-active links. This draft specifies a mechanism to
   simplify PEs used with VXLAN tunnels and enhance VXLAN Active-Active
   reliability.

Status of This Memo

Table of Contents

1. Introduction

A principal feature of EVPN is the ability to support multihoming
from a customer equipment (CE) to multiple provider edge equipment
(PE) with links used in the all-active redundancy mode. That mode is
where a device is multihomed to a group of two or more PEs and where
all PEs in such a redundancy group can forward traffic to/from the
multihomed device or network for a given VLAN [RFC7209]. This draft
specifies a VXLAN gateway mechanism to simplify PE processing in the
multi-homed case and enhance VXLAN Active-Active reliability.

1.1 Terminology and Acronyms

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.

This document uses the following acronyms and terms:

All-Active Redundancy Mode - When a device is multihomed to a group
    of two or more PEs and when all PEs in such redundancy group can
    forward traffic to/from the multihomed device or network for a
    given VLAN.

BUM - Broadcast, Unknown unicast, and Multicast.

CE - Customer Edge equipment.

DCI - Data Center Interconnect.

ESI - Ethernet Segment Identifier - A unique non-zero identifier that
    identifies an Ethernet segment.

NVE - Network Virtualization Edge.

PE - Provider Edge equipment.

Single-Active Redundancy Mode - When a device or a network is
    multihomed to a group of two or more PEs and when only a single PE
    in such a redundancy group can forward traffic to/from the
    multihomed device or network for a given VLAN.

VXLAN - Virtual eXtensible Local Area Network [RFC7348].

VXTEP - VXLAN Tunnel End Point.

2. **VXLAN Gateway High Reliability**

   One example of the current situation would be a DCI (data center
   interconnect) using VXLAN tunnels that is multihomed for reliability
   as show in Figure 1. Each PE as a VXLAN Tunnel End Point (VTEP) uses
   a different IP adress. Thus each PE must process EVPN updates based
   on the ESIs [RFC7432].

```
                              .........
                              .  DCI  .
              +----------+    .       .      +----------+
              | PE       +--------------------+ PE       |
              |VTEP IP-1 +---   . VXLAN .   ---+VTEP IP-3 |
              +----------+   \  .Tunnels.  /   +----------=
          /      |            -----   -----       |     \
        +--+     |            .  \ /  .            |      +--+
        |CE|     |            .   X   .            |      |CE|
        +--+     |            .  / \  .            |      +--+
          \      |            -----   -----        |     /
              +----------+   /  . VXLAN .  \   +----------+
              | PE       +---   .Tunnels.   ---+ PE       |
              |VTEP IP-2 +--------------------+VTEP IP-4 |
              +----------+    .       .      +----------+
                              .........
```
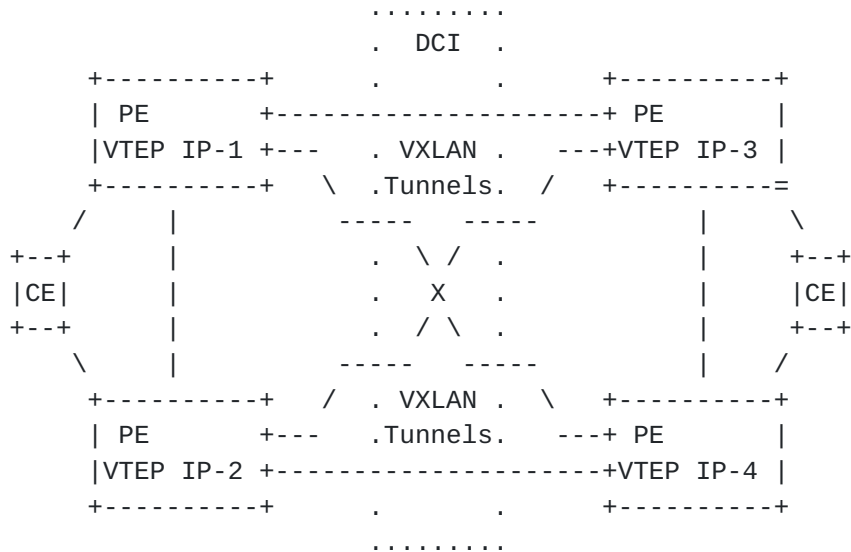
                    Figure 1. Current Situtation

   The situation is greatly simplified if the set of VTEPs connected to
   a particular Ethernet segment all use the same anycast IP address.
   PEs no longer need to conern themselves with whether a remote CE is
   single or multi-homed. The situation is as shown in Figure 2. The IP
   address within each VTEP group is synchronized by messages within
   that group.

```
                             . . . . . . . . .
                             .   DCI   .
             +----------+        .       .        +----------+
             | Anycast  |        .       .        | Anycast  |
             |VTEP IP-1 +---     .       .     ---+VTEP IP-2 |
             +----------+   \  .         .  /     +----------=
            /      ^         \ .         . /           ^      \
      +--+         |          \.         ./            |       +--+
      |CE|     Sy|nc           >-------<           Sy|nc       |CE|
      +--+         |          /. VXLAN .\            |         +--+
         \         v         / . Tunnel. \           v        /
          +----------+    /  .         .  \    +----------+
          | Anycast  +---     .         .     ---+ Anycast  |
          |VTEP IP-1 |        .         .        |VTEP IP-2 |
          +----------+        .         .        +----------+
                             . . . . . . . .
```

                  Figure 2. Situtation Using Anycast

[3]. **Detailed Problem and Solution Requirement**

   In the scenario illustrated in Figure 3, where an enterprise site and
   a data center are interconnected, the VPN gateways (PE1 and PE2) and
   the enterprise site (CPE) are connected through a VXLAN tunnel to
   provide L2/L3 services between the enterprise site (CPE) and data
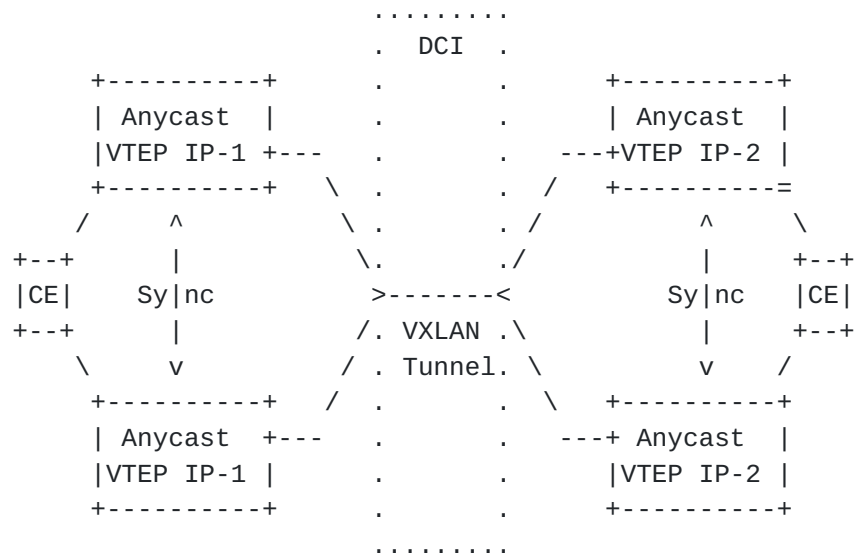   center. The data center gateway (CE1) is dual-homed to PE1 and PE2
   to access the VXLAN network, which enhances network access
   reliability. When one PE fails, services can be rapidly switched to
   the other PE, minimizing the impact on services.

   As shown in Figure 3, PE1 and PE2 use a virtual address as a Network
   Virtualization Edge (NVE) interface address at the network side,
   namely, the Anycast VTEP address. In this way, the CPE is aware of
   only one remote NVE interface and establishes a VXLAN tunnel with the
   virtual address. The packets from the CPE can reach CE1 through
   either PE1 or PE2. However, single-homed CEs may exist, such as CE2
   and CE3. As a result, after reaching a PE, the packets from the CPE
   may need to be forwarded by the other PE to a single-homed CE.
   Therefore, a bypass VXLAN tunnel needs to be established between PE1
   and PE2. An EVPN peer relationship is established between PE1 and
   PE2. Different addresses, namely, bypass VTEP addresses, are
   configured for PE1 and PE2 so that they can establish a bypass VXLAN
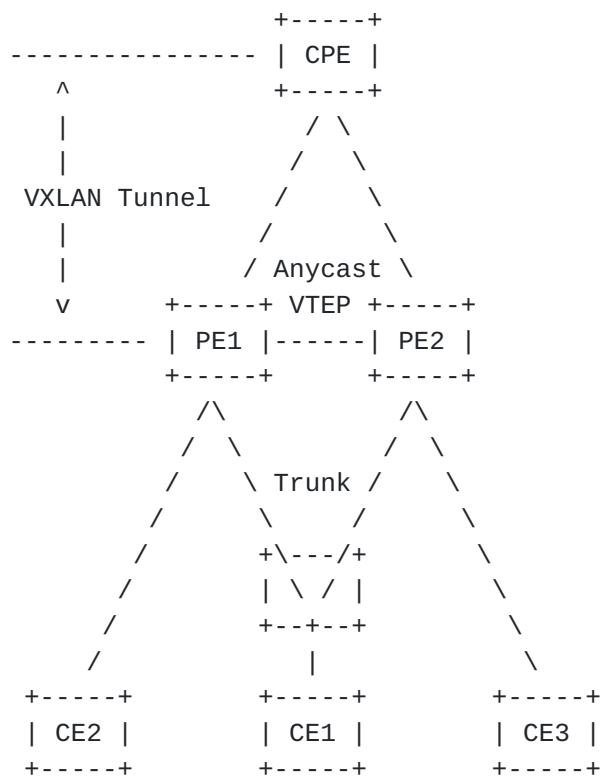   tunnel.

```
                              +-----+
                 --------------- | CPE |
                 ^               +-----+
                 |                  / \
                 |                 /   \
            VXLAN Tunnel          /     \
                 |               /       \
                 |              / Anycast \
                 v      +-----+ VTEP +-----+
            --------- | PE1 |------| PE2 |
                      +-----+      +-----+
                        /\           /\
                       /  \         /  \
                      /    \ Trunk /    \
                     /      \     /      \
                    /       +\---/+       \
                   /        | \ / |        \
                  /         +--+--+         \
                 /            |              \
            +-----+        +-----+        +-----+
            | CE2 |        | CE1 |        | CE3 |
            +-----+        +-----+        +-----+
```

Figure 3. Basic networking of the VXLAN active-active scenario

4. The Bypass VXLAN Extended Community Attribute

This sections describes the extensions specified to meeting the
requirements given in Section 3 and enhance VXLAN active-active
reliability.

This document specifies two new BGP extended communities, called the
Bypass VXLAN Extended Community. The extended communities have a
Type indicating they are transitive and are IPv4-address-specific or
IPv6-address-specific, depending on whether the VTEP address to be
accommodated is IPv4 or IPv6. In the new extended communities, the
4-byte or 16-byte global administrator field encodes the IPv4 or IPv6
address that is the VTEP address and the 2-byte local administrator
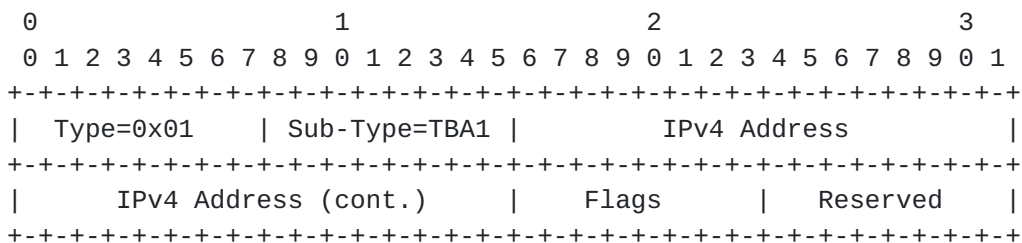field is formatted as shown in Figures 4 and 5.

```
 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=0x01    | Sub-Type=TBA1 |          IPv4 Address          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      IPv4 Address (cont.)     |    Flags     |    Reserved     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

  Figure 4. IPv4-address-specific Bypass VXLAN Extended Community

```
 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=0x00/0x40| Sub-Type=TBA2 |   Target IPv6 Address         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Target IPv6 Address (cont.)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Target IPv6 Address (cont.)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Target IPv6 Address (cont.)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Target IPv6 Address (cont.)  |    Flags     |    Reserved     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

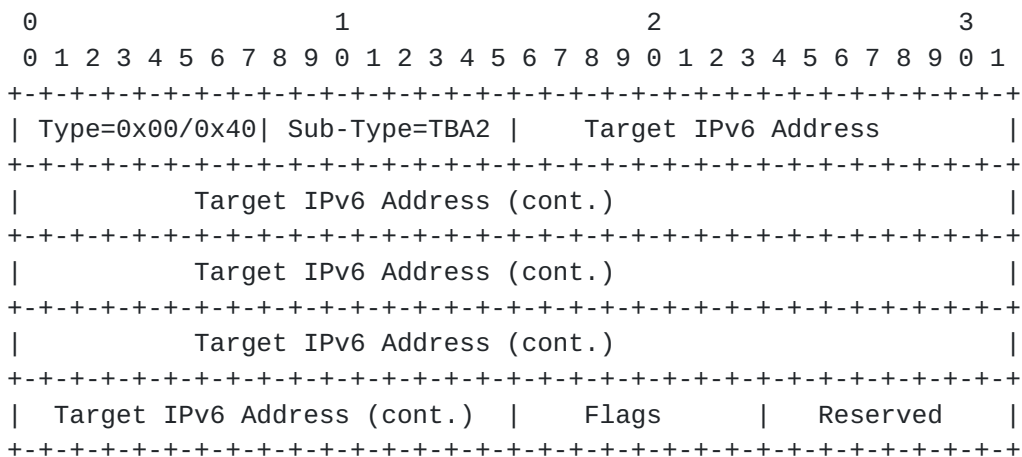 Figure 5. IPv6-address-specific Bypass VXLAN Extended Community

   Where

      Type:
         0x01 = type for IPv4 specific use.
         0x00 = type for transitive IPv6 specific use.
         0x40 = type for non-transitive IPv6 specific use.

      Sub-Type:

TBA1 = subtype for IPv4 specific use.

TBA2 = subtype for IPv6 specific use.

IPv4/IPv6: An address of that type.

Flags: MUST be sent as zero and ignored on receipt.

Reserved: MUST be sent as zero and ignored on receipt.

5. **Control Plane Processing**

   Using the topology in Figure 3:

   1) PE2 sends a multicast route to PE1. The source address of the
   route is the Anycast VTEP address shared by PE1 and PE2. The route
   carries the bypass VXLAN extended community attribute, including the
   bypass VTEP address of PE1.

   2) After receiving the multicast route from PE2, PE1 considers that
   an Anycast relationship be established with PE2. This is because the
   source address (Anycast VTEP address) of the route is the same as the
   local virtual address of PE1 and the route carries the bypass VTEP
   extended community attribute. Based on the bypass VXLAN extended
   attribute of the route, PE1 establishes a bypass VXLAN tunnel to PE2.

   3) PE1 learns the MAC address of the CEs through upstream packets
   from the CEs and advertises them as routes to PE2 through BGP EVPN.
   The routes carry the ESI of the links accessed by the CEs, and
   information about the VLANs that the CE access, and the bypass VXLAN
   extended community attribute.

   4) PE1 learns the MAC address of the CPE through downstream packets
   at the network side, specifies that the next-hop address of the MAC
   route can be iterated to a static VXLAN tunnel, and advertises the
   route to PE2. The next-hop address of the MAC route cannot be
   changed.

6. Data Packet Processing

   This section describes how Layer 2 unicast and BUM (Broadcast,
   Unknown unicast, and Multicast) packets are forwarded. A description
   of how Layer 3 packets transmitted on the same subnet and Layer 3
   packets transmitted across subnets cases are forwarded will be
   provided in a furture version of this document.

6.1 Layer 2 Unicast Packet Forwarding

   The following two subsections discuss Layer 2 unicast forwarding in
   the topology shown in Figure 3.

6.1.1 Uplink

   After receiving Layer 2 unicast packets destined for the CPE from
   CE1, CE2, and CE3, PE1 and PE2 search for their local MAC address
   table to obtain outbound interfaces, perform VXLAN encapsulation on
   the packets, and forward them to the CPE.

6.1.2 Downlink

   After receiving a Layer 2 unicast packet sent by the CPE to CE1, PE1
   performs VXLAN decapsulation on the packet, searches the local MAC
   address table for the destination MAC address, obtains the outbound
   interface, and forwards the packet to CE1.

   After receiving a Layer 2 unicast packet sent by the CPE to CE2, PE1
   performs VXLAN decapsulation on the packet, searches the local MAC
   address table for the destination MAC address, obtains the outbound
   interface, and forwards the packet to CE2.

   After receiving a Layer 2 unicast packet sent by the CPE to CE3, PE1
   performs VXLAN decapsulation on the packet, searches the local MAC
   address table for the destination MAC address, and forwards it to PE2
   over the bypass VXLAN tunnel. After the packet reaches PE2, PE2
   searches the destination MAC address, obtains the outbound interface,
   and forwards the packet to CE3.

   The process for PE2 to forward packets from the CPE is the same as
   that for PE1 to forward packets from the CPE.

6.2 BUM Packet Forwarding

   Using the topology in Figure 3, if the destination address of a BUM
   packet from the CPE is the Anycast VTEP address of PE1 and PE2, the
   BUM packet may be forwarded to either PE1 or PE2. If the BUM packet
   reaches PE2 first, PE2 sends a copy of the packet to CE3 and CE1. In
   addition, PE2 sends a copy of the packet to PE1 through the bypass
   VXLAN tunnel between PE1 and PE2. After the copy of the packet
   reaches PE1, PE1 sends it to CE2, not to the CPE or CE1. In this
   way, CE1 receives only one copy of the packet.

   Using the topology in Figure 3, after a BUM packet from CE2 reaches
   PE1, PE1 sends a copy of the packet to CE1 and the CPE. In addition,
   PE1 sends a copy of the packet to PE2 through the bypass VXLAN tunnel
   between PE1 and PE2. After the copy of the packet reaches PE2, PE2
   sends it to CE3, not to the CPE or CE1.

   Using the topology in Figure 3, after a BUM packet from CE1 reaches
   PE1, PE1 sends a copy of the packet to CE2 and the CPE. In addition,
   PE1 sends a copy of the packet to PE2 through the bypass VXLAN tunnel
   between PE1 and PE2. After the copy of the packet reaches PE2, PE2
   sends it to CE3, not to the CPE or CE1.

7. IANA Considerations

   IANA is requested to assign two new Extended Community attribute
   SubTypes as follows:

7.1 IPv4 Specific

```
   Sub-Type Value   Name                            Reference
   --------------   ------------------------------  ----------
       TBA1         Bypass VXLAN Extended Community  [this doc]
```

7.2 IPv6 Specific

```
   Sub-Type Value   Name                            Reference
   --------------   ------------------------------  ----------
       TBA2         Bypass VXLAN Extended Community  [this doc]
```

8. Security Considerations

   TBD

   For general EVPN Security Considerations, see [RFC7432].

Acknowledgements

   The authors would like to thank the following for their comments and
   review of this document:

      TBD

Contributors

   The following individuals made significant contributions to this
   document:

      Haibo Wang
      Huawei Technologies
      Huawei Bldg., No. 156 Beiqing Road
      Beijing 100095
      China

      Email: rainsword.wang@huawei.com

Normative References

   [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate
         Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119,
         March 1997, <http://www.rfc-editor.org/info/rfc2119>.

   [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>.

   [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>.

Informative References

   [RFC7209] - Sajassi, A., Aggarwal, R., Uttaro, J., Bitar, N.,
         Henderickx, W., and A. Isaac, "Requirements for Ethernet VPN
         (EVPN)", RFC 7209, DOI 10.17487/RFC7209, May 2014,
         <https://www.rfc-editor.org/info/rfc7209>.

   [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, <http://www.rfc-
         editor.org/info/rfc7348>.

Authors' Addresses

      Donald E. Eastlake, 3rd
      Huawei Technologies
      1424 Pro Shop Court
      Davenport, FL 33896 USA

      Phone: +1-508-333-2270
      Email: d3e3e3@gmail.com

      Zhenbin Li
      Huawei Technologies
      Huawei Bld., No.156 Beiqing Rd.
      Beijing 100095
      China

      Email: lizhenbin@huawei.com

      Shunwan Zhuang
      Huawei Technologies
      Huawei Bld., No.156 Beiqing Rd.
      Beijing 100095
      China

      Email: zhuangshunwan@huawei.com