BESS Z. Zhang
Internet-Draft Juniper Networks
Intended status: Standards Track J. Rabadan, Ed.
Expires: September 12, 2019 Nokia
A. Sajassi
Cisco Systems
March 11, 2019

MVPN/EVPN Composite Tunnel
draft-zzhang-bess-mvpn-evpn-composite-tunnel-00

Abstract

EVPN E-Tree defines a composite tunnel to be used for a Root PE to simultaneously indicate a non-Ingress-Replication tunnel (e.g., P2MP tunnel) in the transmit direction and an Ingress Replication tunnel in the receive direction for BUM traffic. This document extends it to more generic use in both MVPN and general EVPN.

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 RFC2119.

Status of This Memo

Copyright Notice

Table of Contents

1. Terminologies

Familiarity with BIER/MVPN/EVPN protocols and procedures is assumed. Some terminologies are listed below for convenience.

[To be added].

2. Introduction

The composite tunnel defined in [RFC8317] is specifically designed for the particular use case of EVPN E-Tree in that the Root PE only needs to receive on the Ingress Replication (IR) tunnel and transmit on the non-IR tunnel encoded in the PMSI Tunnel Attribute (PTA) that specifies a Composite Tunnel, hence the following language quoted from [RFC8317]:

> Composite tunnel type is advertised by the Root PE to simultaneously indicate a non-Ingress-Replication tunnel (e.g., P2MP tunnel) in the transmit direction and an Ingress Replication tunnel in the receive direction for the BUM traffic.

However, the underlying principal of MVPN PMSI A-D route, EVPN IMET route and the PTA that the routes carry allows the composite tunnel to be used in more generic use cases for both MVPN and EVPN, as explained in Section Section 2.1 and Section 2.2.

The EVPN IMET route is the equivalent of MVPN I-PMSI A-D route. In the rest of the document, unless explicitly stated, I-PMSI A-D route refers to MVPN Intra-AS I-PMSI A-D route and/or EVPN IMET route.

2.1. P2MP Tunnels

As specified in [RFC6514] [RFC7432], an I-PMSI A-D route advertises a PE's membership in a VPN or Broadcast Domain (BD). The route carries a PTA, whose Tunnel Identifier field specifies the tunnel that the advertising PE uses to send traffic unless the tunnel type is either "No tunnel information present" or "Ingress Replication". A PE that imports the route into a VRF/BD/EVI will join the specified tunnel if it needs to receive traffic from the advertising PE.

As specified in [RFC6514] and clarified in [RFC7988], if the tunnel type is Ingress Replication, and the Leaf Information Required (LIR) bit in the PTA's Flags field is set to 0, the advertise PE is actually not indicating that it uses IR to send traffic, but that it will receive traffic using the label that is part of the Tunnel Identifier field of the PTA. A PTA with tunnel type set to IR and LIR bit set to 1 does indicate that the advertising router will use IR to send traffic. In that case, the label field in the Tunnel Identifier is set to 0 and receiving PEs will need to send a Leaf A-D route to "join" the IR tunnel. The label value in the Tunnel Identifier of the Leaf A-D route's PTA is used when sending traffic to the advertiser of the Leaf A-D route.

While [RFC7988] is MVPN specific, the above IR procedures and clarifications are also applicable to EVPN, as the EVPN IMET route is the equivalent of an I-PMSI A-D route with the LIR flag set to 0.

In summary, w/o considering composite tunnel, when IR is specified in I-PMSI A-D routes w/ the LIR bit NOT set, the tunnel is used to receive traffic (even from PEs not advertising IR in its PMSI A-D routes). The composite tunnel introduced in [RFC8317] combines a transmitting non-IR tunnel and a receiving IR tunnel, but a PE advertising a composite tunnel should be still be able to send to certain PEs using IR.

2.2. MP2MP Tunnels

When the PTA specifies one of the MP2MP tunnels (BIDIR-PIM, mLDP MP2MP, BIER), it means the advertising PE will use the MP2MP tunnel for both sending and receiving. While [RFC8317] specifies composite tunnel only as transmitting non-IR + receiving IR, an MP2MP tunnel can also be part of composite tunnel to receive traffic. The rest of the document focuses on BIER, but it equally applies to mLDP MP2MP or BIDIR-PIM.

3. Specifications

While not previously done so, this document makes it explicit that, an MVPN/EVPN PE1 advertising a non-IR tunnel for sending traffic can also send to another PE2 using IR if that PE2 advertises to receive traffic with IR (whether PE advertises IR standalone or as part of a composite tunnel), as long as it is known that PE2 does not also join the non-IR tunnel on which PE1 is also sending the same data.

3.1. General MVPN/EVPN Use of Composite Tunnels

This document extends the use of composite tunnel to appropriate general MVPN/EVPN scenarios where a PE advertises a composite tunnel in its I-PMSI A-D route to receive traffic on IR tunnel and send traffic on non-IR tunnel.

This document also allows an MP2MP tunnel to be part of a composite tunnel so that the advertising PE can use both the MP2MP tunnel and IR to receive traffic.

For a regular, non-composite tunnel in the PMSI Tunnel Attribute (PTA) of a PMSI/Leaf A-D route, the PTA includes an "MPLS Label" field between the "Tunnel Type" field and the "Tunnel Identifier" field. The label is for "tunnel aggregation" purpose - traffic on the same tunnel could carry different labels for multiplexing purpose (e.g. for different VPNs/BDs). For an IR tunnel, the label is downstream- assigned; for non-IR tunnels, the label is either 0 (no aggregation) or upstream-assigned, or from a Domain-wide Common Block (DCB) [I-D.ietf-bess-mvpn-evpn-aggregation-label].

```
                +---------------------------------+
                |  Flags (1 octet)                |
                +---------------------------------+
                |  Tunnel Type (1 octets)         |
                +---------------------------------+
                |  MPLS Label (3 octets)          |
                +---------------------------------+
                |  Tunnel Identifier (variable)   |
                +---------------------------------+
```

PTA Fields [RF6514]

[RFC8317] specifies that the "Tunnel Identifier" field includes a three-octet label before the actual identifier of the non-IR tunnel, though the text/diagram about the roles of the labels is unclear/ confusing. For easier reference this document moves the added label out, so that the "Tunnel Identifier" is the actual identifier of the non-IR tunnel:

```
                +-------------------------------------------+
                | Flags (1 octet)                           |
                +-------------------------------------------+
                | Tunnel Type (1 octet)                     |
                +-------------------------------------------+
                | Non-IR Tunnel Aggregation Label (3 octets)|
                +-------------------------------------------+
                | Ingress Replication MPLS Label (3 octets) |
                +-------------------------------------------+
                | Non-IR Tunnel Identifier (variable)       |
                +-------------------------------------------+
```

PTA Fields for Composite Tunnel [This Document]

An example of composite tunnel is BIER-IR tunnel, where the tunnel type is set to 0x8B, and BIER Tunnel Aggregation Label and BIER Tunnel Identifier are as specified in [I-D.ietf-bier-mvpn].

Section Section 4.1 gives an example application of using BIER-IR tunnel for BIER capable EVPN PEs to send/receive BUM traffic via BIER, and receive BUM traffic from BIER incapable PEs via IR; BIER incapable PEs send BUM traffic using IR; BIER traffic from BIER capable PEs will have the BIER header popped off by a Penultimate Hop before reaching BIER incapable PEs [I-D.ietf-bier-php]. The same can be used for MVPN as well.

3.2. EVPN-IP Assisted Replication with BIER-IR Composite Tunnel

For the example in Section Section 4.1, instead of having BIER incapable PEs send BUM traffic using IR to every PE, Assisted Replication (AR) [I-D.ietf-bess-evpn-optimized-ir] can be used for a BIER incapable PE to send BUM traffic to a BIER capable Assisted Replication Replicator (AR-R) via IR, who will then relay to other PEs via BIER.

The same concept applies to MVPN as well, though AR for MVPN is via Virtual Hub and Spoke (VHS) [RFC7024], similar to AR for EVPN-MPLS [I-D.keyupate-bess-evpn-virtual-hub] ([I-D.ietf-bess-evpn-optimized-ir] is for EVPN-IP). The procedures for those two cases will be specified in separate documents.

EVPN-IP AR with BIER-IR composite tunnels follows similar procedures as in [I-D.ietf-bess-evpn-optimized-ir], with the following differences:

- The IMET route from a BIER capable AR-Replicator that has the IR-IP address in the Originating PE field encodes BIER tunnel in the PTA, as specified in [EVPN-BIER].

- An AR-Leaf originates an IMET route with BIER-IR tunnel with AR-Leaf flag. If it is BIER capable, it both sends and receives BM traffic via BIER. If it is not BIER capable, it sends BM traffic via IR to the AR-Replicator, who will then relay to other PEs using BIER.

- The AR-R does NOT relay traffic that arrive with BIER encapsulation.

- Only non-selective mode is supported.

The above rules are illustrated in further details in Section Section 4.2. Notice that, composite-tunnel is used because [I-D.ietf-bess-evpn-optimized-ir] requires falling back to IR when the AR-Replicator is not available.

4. Use-cases

This section describes some Composite Tunnel use-cases. We refer to BIER-IR as the PTA's Tunnel Type with the high-order bit set and BIER type, I.e., Tunnel Type = 0x8B, as per [RFC8317]. In this section, a BIER non-capable PE is assumed to be a PE that does not support BIER tunnel data plane transmission or termination. However, these BIER non-capable PEs support the required control plane extensions to advertise BIER tunnel information in the IMET PTAs.

4.1. BIER-IR Composite Tunnels in EVPN Networks

BIER-IR composite tunnels may be used in a group of PEs attached to the same EVPN tenant network. This would allow some of those PEs to use BIER P-tunnels where other PEs in the same group may use Ingress Replication (IR). While these BIER-IR composite tunnels can be used in a similar use case as described in [RFC8317], they can also be used along with PHP as a way to introduce BIER in EVPN networks where some of the PEs do not support BIER data plane.

Figure 1 illustrates an example of an EVPN BD where the PE1 and PE2 support BIER data plane, but PE3/PE4/PE5 do not. The network could still benefit of BIER if the BFRs connected to the receiver PEs (BFR1 and BFR2), either directly or through a tunnel, pop the BIER header and send the EVPN payload natively to PE3/PE4/PE5, as described in [I-D.ietf-bier-php].

```
                                           IMET(BIER-IR)
                                       <---------
                 PE2        IMET(BIER-IR)             PE3
            +-----------+       ---------->      +-----------+
            | +-------+ |                        | +-------+ |--> R1
            | |   BD  | |<---+       +----------->| |   BD  | |
            | +-------+ |    |  BFR1 |           | +-------+ |--> R2
            +-----------+    |     +---+         +-----------+
                             +---->|   |              PE4
                 PE1         |     +---+         +-----------+
            +-----------+    |       |           | +-------+ |
            | +-------+ |    |       +----------->| |   BD  | |--> R3
      S1--> | |   BD  | |----+                   | +-------+ |
            | +-------+ |    |                   +-----------+
            +-----------+    |  BFR2                  PE5
                             |     +---+         +-----------+
                             +---->|   |--------->| +-------+ |
                                   +---+         | |   BD  | |--> R4
                                                 | +-------+ |
                  |                  |           +-----------+
                  |<---------------->|
                  |        BIER      |
                BFIR                 BFER
                                     (PHP)
                  |                                       |
                  |<------------------------------------->|
                  |              EVPN                     |
```

Figure 1 - BIER Composite Tunnels in EVPN

In this example:

- All the PEs advertise an IMET route containing a BIER-IR composite tunnel in the PTA (PMSI Tunnel Attribute):

    * The Tunnel Type has a value of 0x8B (BIER-IR composite tunnel).

    * The BIER Tunnel Identifier (composed Sub-Domain ID, BFR-ID and BFR-Prefix) and Flags are populated as in [EVPN-BIER]. The IR Label is a downstream allocated Label that allows remote PEs to send BUM traffic to the advertising PE using Ingress Replication, as in [RFC8317].

- When PE1/PE2 need to transmit BUM packets, they follow the procedures in [EVPN-BIER]. BUM packets received on PE1/PE2 from other BIER capable PEs will be received with a BIER encapsulation and procedures in [EVPN-BIER] will be followed. PHP nodes pop the BIER header before delivering the EVPN packets to PE3/PE4/PE5.

- When PE3/PE4/PE5 need to send BUM packets to each other or to PE1/PE2, they use Ingress Replication and the IR label that is received from the other PEs as part of the composite tunnel Tunnel Identifier.

4.2. Assisted Replication and BIER Composite Tunnels

The use case in section Section 4.1 allows the introduction of BIER in EVPN tenant networks where BIER capable and BIER non-capable PEs are attached to the same EVPN tenant network. However, BIER non-capable PEs still send multiple copies of the same BUM packet to reach the other PEs.

In overlay networks, the use case can be optimized so that the BIER non-capable PEs send a single copy per packet by using Assisted Replication along with BIER-IR composite tunnels. Figure 2 illustrates this use case with an example, where AR-L1/AR-L2/AR-L3 and AR-L4 are Assisted Replication Leaf routers [AR] that do not support BIER data plane. AR-R1/AR-R2 are Non-Selective Assisted Replication Replicators [AR] that do support BIER data plane and are connected to other BFRs, such as BFR1 and BFR2.

```
                                                    AR-L2
                                                   +-------+ --> R1
                                              +--->|   BD  |
                IMET                    BFR1  |    +-------+ --> R2
                (BIER-IR,AR-Leaf)           +---+
               --->                     +-->|   |
                                        |   +---+   AR-L3
             AR-L1            AR-R1     |     |    +-------+
            +-------+        +-------+  |     +--->|   BD  | --> R3
       S1-->|   BD  |------->|   BD  |--+          +-------+
            +-------+        +-------+  |
                                        |
                  <----       AR-R2     | BFR2      AR-L4
                    IMET     +-------+  |   +---+  +-------+
                   (AR,AR-R) |   BD  |  +-->|   |->|   BD  | --> R4
                             +-------+      +---+  +-------+

                               |              |
                               |<------------>|
                               |     BIER     |
                             BFIR            BFER
                                             (PHP)
               |                                       |
               |<------------------------------------->|
               |              EVPN                     |
```

Figure 2 - BIER-IR Composite Tunnels and AR

In this example:

- o The AR-R PEs issue two IMET routes each:

  - * An IMET route that includes the AR-IP in the Originating PE, Tunnel Type AR, IR label, Flags Type = 01 (AR-Replicator) and L = 0 (no Leaf Information Required). The IR Label is a downstream allocated Label that will be used by the AR-L PEs that transmit BUM traffic to the receiving AR-R for replication to remote AR-L PEs. No change with respect to [AR].

  - * And an IMET route that includes the IR-IP in the Originating PE field, Tunnel Type and Tunnel Identifier with BIER information, as in [EVPN-BIER].

- o Each AR-L PE issues an IMET route with:

  - * The Flags field populated as in [AR] with AR Type set to AR-LEAF.

- The Tunnel Type and Tunnel Identifier have composite BIER-IR information, as in Section Section 4.1. The IR Label is a downstream allocated Label that will be used by the remote AR-L PEs when IR is used for unknown unicast traffic. The MPLS Label field in the PTA MAY be zero.

When an AR-R receives a BM packet encapsulated in an overlay tunnel, it will do a tunnel destination IP lookup and if the destination IP is the AR-R IR-IP Address, the AR-R will proceed as in [AR]. If the destination IP is the AR-R AR-IP Address, the AR-R MUST forward the packet to the BIER network and any local AC (if any). When creating the BIER header, the AR-R will behave as a BFIR and will include all the remote AR-L and AR-R in the BIER header, excluding the AR-L from which the BM packet was received.

If an AR-R receives a BM packet encapsulated in BIER, it will follow the procedures in [EVPN-BIER] as any other BIER PE. It MUST NOT send the BM packets to any overlay tunnels, only to local ACs.

In the example of Figure 2, when AR-R1 receives a BM packet from AR-L1 in an overlay tunnel with its AR-IP as tunnel destination address, it will forward the packet encapsulated with a BIER header that includes AR-L2, AR-L3 and AR-L4 as BFERs, but not AR-L1.

As in [AR], if the AR-L does not discover any AR-R in the service, it MUST use IR to send BM traffic to the remote AR-L PEs and AR-R PEs with local ACs. If there is one or more AR-Rs (discovered by tracking the received AR-R routes) the AR-L selects a AR-R to send the BM traffic to. The selection rules are described in [AR]. The AR-L encapsulates the BM packets into an overlay tunnel that uses the AR-IP and AR Label advertised by the selected AR-R. In the example of Figure 2, AR-L1 selects AR-R1 as the AR-R. If AR-R1 becomes unavailable, AR-R2 is selected. If no AR-R is available, AR-L1 would use IR to send the BM packets to the remote AR-L PEs.

AR-L PEs receive the BUM packets without a BIER header (since it is popped by the PHP node) and with the MPLS Label / VNI imposed by the AR-R (or the source AR-L if there is no AR for the packet).

5. Security Considerations

This specification does not introduce additional security concerns.

6. IANA Considerations

7. Acknowledgements

8. References

8.1. 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, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

[RFC8317] Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J., Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/RFC8317, January 2018, <https://www.rfc-editor.org/info/rfc8317>.

8.2. Informative References

[I-D.ietf-bess-evpn-optimized-ir]
Rabadan, J., Sathappan, S., Lin, W., Katiyar, M., and A. Sajassi, "Optimized Ingress Replication solution for EVPN", draft-ietf-bess-evpn-optimized-ir-06 (work in progress), October 2018.

[I-D.ietf-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-ietf-bess-mvpn-evpn-aggregation-label-02 (work in progress), December 2018.

[I-D.ietf-bier-evpn]
Zhang, Z., Przygienda, T., Sajassi, A., and J. Rabadan, "EVPN BUM Using BIER", draft-ietf-bier-evpn-01 (work in progress), April 2018.

[I-D.ietf-bier-mvpn]
Rosen, E., Sivakumar, M., Aldrin, S., Dolganow, A., and T. Przygienda, "Multicast VPN Using BIER", draft-ietf-bier-mvpn-11 (work in progress), March 2018.

[I-D.ietf-bier-php]
Zhang, Z., "BIER Penultimate Hop Popping", draft-ietf-bier-php-01 (work in progress), November 2018.

[I-D.keyupate-bess-evpn-virtual-hub]
Patel, K., Sajassi, A., Drake, J., Zhang, Z., and W. Henderickx, "Virtual Hub-and-Spoke in BGP EVPNs", draft-keyupate-bess-evpn-virtual-hub-01 (work in progress), October 2018.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7024] Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", RFC 7024, DOI 10.17487/RFC7024, October 2013, <https://www.rfc-editor.org/info/rfc7024>.

[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>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Jorge Rabadan (editor)
Nokia

EMail: jorge.rabadan@nokia.com

Ali Sajassi
Cisco Systems

EMail: sajassi@cisco.com