BIER WG
Internet-Draft
Intended status: Informational
Expires: November 25, 2017
May 24, 2017

Z. Zhang
B. Wu
ZTE Corporation

# Analysis of BIER in NVO3 network
draft-zhangwu-bier-nvo3-analysis-00

## Abstract

This document analyses BIER deployment in NVO3 network. The intent is to evaluate whether BIER could achieve some simplicity and efficiency.

Table of Contents



## 1. Introduction

As mentioned in [I-D.ietf-nvo3-mcast-framework], there is multicast requirement in NVO3 network, such as BUM packets for infrastructure multicast and other application-specific multicast. Besides PIM, BIER is mentioned to be used as one of IP multicast underlay technology in section 3, which introduces several multicast mechanisms for NVO3 network.

Bit Index Explicit Replication (BIER) [I-D.ietf-bier-architecture] is a new architecture for the forwarding of multicast data packets. It provides optimal forwarding of multicast packets through a "multicast domain". It does not require a protocol for explicitly building multicast distribution trees, nor does it require intermediate devices to maintain any per-flow state. When a multicast data packet enters the BIER domain, the BIER ingress router determines the set of BIER egress routers to which the packet needs to be sent. The BIER ingress router then encapsulates the packet in a BIER header. The BIER header contains a bitstring in which each bit represents exactly one BIER egress router in the domain; to forward the packet to a given set of egress routers, the bits corresponding to those routers are set in the BIER header. In this way, elimination of the per-flow state and the explicit tree-building protocols results in a considerable simplification.

This document will give some basic analysis on how to deploy BIER in the NVO3 network to mitigate multicast states.

## 2. BIER in NVO3 network

```
       +--------+                                             +--------+
       | Tenant +--+                                     +----| Tenant |
       | System |  |                                    (')   | System |
       +--------+  |          ................         (   )  +--------+
                   |  +-+--+  .              .  +--+-+  (_)
                   |  | NVE|--.              .--| NVE|   |
                   +--|    |  .              .  |    |---+
                      +-+--+  .              .   +--+-+
                      /       .              .
                     /        .  L3 Overlay  .   +--+-++--------+
       +--------+   /         .    Network   .   | NVE|| Tenant |
       | Tenant +--+          .              .- -|    || System |
       | System |             .              .   +--+-++--------+
       +--------+             ................
                                     |
                                   +----+
                                   | NVE|
                                   |    |
                                   +----+
                                     |
                                     |
                           =====================
                             |               |
                         +--------+      +--------+
                         | Tenant |      | Tenant |
                         | System |      | System |
                         +--------+      +--------+
```

Figure 1: NVO3 Generic Reference Model

As described in [RFC8014], figure 1 is a generic reference model of NVO3.  BIER can be used in this model to provide underlay multicast function.

```
                  ...........................................
                  .                                 ------- .   -------
                  .            ................     | Leaf| .---| NVE |
                  .-------     .              .-----|     | .   |     |
        -------   .| Leaf|     .              .     ------- .   -------
        | NVE |---.|     |-----.              .             .
        |     |   .-------     .    Spine     .             .
        -------   .            .              .             .
                  .            .              .             .
                  .            .              .     ------- .
                  .            .              .-----| Leaf| .   -------
                  .            ................     |     | .---| NVE |
                  .                   |             ------- .   |     |
                  .                   |                     .   -------
                  .                 -------                 .
                  .                 | Leaf|                 .
                  .                 |     |                 .
                  .                 -------                 .
                  ...........................................
                                |              |
                                |              |
                             -------        -------
                             | NVE |        | NVE |
                             |     |        |     |
                             -------        -------
```

Figure 2: Example topology of NVO3 underlay network

The architecture of L3 underlay Network can be various. As depicted in figure 2, the Clos topology mentioned in [RFC7938] is taken as an example to offer underlay L3 connectivity. And it is assumed that the leaf and spine devices support BIER functionality including exchanging BFR-id and other information and building BIER forwarding table.

In this document, two possible ways to handle multicast packets are identified. One of the choices is placing BIER boundaries at leaf switches which are a general choice of implementing BIER. The other way is putting the BIER boundary at NVE to achieve better efficiency.

## 2.1. Leaf devices as the edge BIER nodes

Leaf devices are taken as BIER edge node. Each leaf device is allocated one unique BFR-id. And leaf devices are the ingress BIER nodes and egress BIER nodes. Both the leaf and spine devices exchange the BIER information by IGP/BGP extensions. The according extensions are defined in [I-D.ietf-bier-ospf-bier-extensions], [I-D.ietf-bier-isis-extensions], and [I-D.ietf-bier-idr-extensions].

IGMP/MLD protocol is used between NVE and leaf device like the description in [I-D.ietf-nvo3-mcast-framework]. The BUM flows are encapsulated corresponding multicast group address. BMLD [I-D.pfister-bier-mld] protocol is used among all the leaf devices to exchange multicast group information. After BMLD protocol process is completed, each leaf devices knows the other leaf devices associated with specific multicast group address. When one packet with multicast group address reaches leaf device, leaf device encapsulates the packet with BIER header which indicates all the destination leaf devices that belong to the same multicast group.

After the BIER packet reaches the destination leaf devices through the spine network forwarding, the destination leaf device removes the BIER header of packet and forwards to corresponding NVEs.

So the multicast state is eliminated because of removing of multicast tree in L3 underlay network. NVE only needs to support IGMP/MLD protocol. But one or several multicast group addresses for a tenant is still needed.

```
                                                           IGMP/MLD
        IGMP/MLD  ..........................................   |
            \     .    ..............................     .    |
             \    .    .  BIER forwarding           .     .    |
              \   .    .                          ------- .    |-------
               \  .    .       ................   |Leaf2| .---| NVE2|
                \ .-------     .              .-----|     | .    |     |
        -------   .|Leaf1|     .              .   ------- .    -------
        | NVE1|---.|     |-----.              .      .    .
        |     |   .-------     .    Spine     .      .    .
        -------   .    .       .              .      .    .
                  .    .       .              .      .    .
                  .    .       .              .   ------- .
                  .    .       .              .-----|Leaf3| .    -------
                  .    .       ................   |     | .---| NVE3|
                  .    .              |           ------- . |  |     |
                  .    .              |              .    . |  -------
                  .    . BMLD      -------           .    . |
                  .    ............ |Leaf4|............     . IGMP/MLD
                  .                 |     |                 .
                  .                 -------                 .
                  ..........................................
                               |   IGMP/MLD   |
                               |              |
                            -------        -------
                            | NVE4|        | NVE5|
                            |     |        |     |
                            -------        -------
```

Figure 3: BIER in VNO3

For example, as illustrated in Figure 3, NVE1 needs to send BUM flows to NVE2, NVE3, NVE4 and NVE5. The NVE supports the basic IGMP/MLD snooping function. In most of condition, there would have to be a separate group for each tenant, plus a separate group for each multicast address (used for multicast applications) within a tenant. NVE1 sends the packet to Leaf1. According the BMLD exchange Leaf1 encapsulate BIER header with the destination of Leaf2, Leaf3 and Leaf4. After BIER forwarding the packet reaches Leaf2, Leaf3 and Leaf4, the leaf devices remove the BIER header and send it to NVE2, NVE3, NVE4 and NVE5.

In some use cases, there could be a big amount of leaf switch, and it is impossible to encapsulate destination BFR-ids in one same BIER header because of the limitation of BitStringLength, the packet should be sent more than once to reach destination. In case the BFR-id of Leaf2 is 28, the BFR-id of Leaf3 is 78; the BitStringLength used in BIER encapsulation is 64, it is impossible to encapsulate the two BFR-ids in one BIER header. In this situation, one solution is

Leaf1 send two copies of packet to Leaf2 and Leaf3. SI in BIER header defined in [I-D.ietf-bier-architecture] is be used to distinguish these two copies to deliver to the final destination. The other solution is increase the forwarding BitStringLength to 128. From the above analysis, BIER is quite applicable in the scenario with limited size of leaf switches. But in a large scale of NVO3 underlay network, there is some limitation due to the BitStringLength.

## 2.2. NVE as edge BIER nodes

In last section, IGMP/MLD is still needed to run between NVE and leaf devices. And the multicast groups for Tenants and specific multicast applications are needed.

If NVE is the edge BIER node, IGMP/MLD protocol does not need to run between NVE and leaf device. NVE encapsulates the BIER header with the BFR-ids of destination NVEs straightly and send it to leaf devices. Leaf and spine devices forward the BIER packet to destination NVEs. Destination NVEs remove the BIER header and forwarding according to the inner encapsulation.

In this situation, leaf device does not need to be allocated BFR-id. Every NVE should be allocated with one unique BFR-id. The BFR-id information should be exchanged within the L3 underlay network. If NVE supports the IGP/BGP BIER extension, NVE takes part in the BIER information exchange. The forwarding plane will be established easily. But in some situations NVE does not support IGP or BGP; NVE can not take part in the BIER information exchange. So the BFR-id of NVE should be advertised by some other method.

Besides multicast state elimination in L3 underlay network, IGMP/MLD does not need to run between NVE and leaf devices. And BMLD does not need to run among all the leaf devices. This function makes the BIER deployment more simply.

```
                  ...........................................
                  .    ................................     .
                  .    .  BIER forwarding             .     .
                  .    .                            ------- .   -------
                  .    .       ................     |Leaf2| .---| NVE2|
                  .-------     .              .-----|     | .   |     |
        -------   .|Leaf1|     .              .     ------- .   -------
        | NVE1|---.|     |-----.              .       .     .
        |     |   .-------     .    Spine     .       .     .
        -------   .    .       .              .       .     .
                  .    .       .              .       .     .
                  .    .       .              .     ------- .
                  .    .       .              .-----|Leaf3| .   -------
                  .    .       ................     |     | .---| NVE3|
                  .    .              |             ------- .   |     |
                  .    .              |               .     .   -------
                  .    .            -------           .     .
                  .    ............ |Leaf4|............     .
                  .                 |     |                 .
                  .                 -------                 .
                  ...........................................
                                |              |
                                |              |
                             -------        -------
                             | NVE4|        | NVE5|
                             |     |        |     |
                             -------        -------
```

Figure 4: BIER in VNO3

The same example for figure 3, NVE1 should send BUM packet to NVE2, NVE3, NVE4 and NVE5. NVE1 encapsulates BIER header with NVE2, NVE3, NVE4 and NVE5 as destination and send it to Leaf1. Leaf1 and spine devices forward the packet to NVE2, NVE3, NVE4 and NVE5 according to the BIER header. The NVEs remove the BIER header and forward it.

The BitStringLength limitation also remains in this solution. And the situation may be more seriously because of the larger number of NVEs than leaf devices. If the BFR-id of NVE is not allocated reasonably, in the worst situation, the forwarding efficiency is the same with the source replication described in [I-D.ietf-nvo3-mcast-framework] section 3.2.

## 3. Other Consideration

As illustrated in [RFC7938], there could be hundred thousand servers connected by underlay network in some NVO3 network. So there are more than thousands of NVEs and leaf devices in the network. Using BIER as multicast underlay protocol make significant advantage because of the elimination of multicast state stored in the underlay network. But the BitStringLength limitation is one problem.

In order to achieve the optimization of BIER, the BFR-ids allocation should be more reasonable. The BFR-id of NVE/leaf device that is belong to one same VN could be allocated adjacent as much as possible. So the encapsulation of BIER header can be more efficient.

Along with the number of BFR-id increasing for NVE/leaf devices, there are thousands BIER forwarding items in the L3 underlay network. The forwarding efficiency in the L3 underlay network should also be considered.

## 4. Normative References

[I-D.ietf-bier-architecture]
Wijnands, I., Rosen, E., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast using Bit Index Explicit Replication", draft-ietf-bier-architecture-06 (work in progress), April 2017.

[I-D.ietf-bier-idr-extensions]
Xu, X., Chen, M., Patel, K., Wijnands, I., and T. Przygienda, "BGP Extensions for BIER", draft-ietf-bier-idr-extensions-02 (work in progress), January 2017.

[I-D.ietf-bier-isis-extensions]
Ginsberg, L., Przygienda, T., Aldrin, S., and Z. Zhang, "BIER support via ISIS", draft-ietf-bier-isis-extensions-04 (work in progress), March 2017.

[I-D.ietf-bier-ospf-bier-extensions]
Psenak, P., Kumar, N., Wijnands, I., Dolganow, A., Przygienda, T., Zhang, Z., and S. Aldrin, "OSPF Extensions for BIER", draft-ietf-bier-ospf-bier-extensions-05 (work in progress), March 2017.

[I-D.ietf-nvo3-mcast-framework]
Ghanwani, A., Dunbar, L., McBride, M., Bannai, V., and R. Krishnan, "A Framework for Multicast in Network Virtualization Overlays", draft-ietf-nvo3-mcast-framework-08 (work in progress), May 2017.

[I-D.pfister-bier-mld]
Pfister, P., Wijnands, I., Venaas, S., Wang, C., Zhang, Z., and M. Stenberg, "BIER Ingress Multicast Flow Overlay using Multicast Listener Discovery Protocols", draft-pfister-bier-mld-03 (work in progress), March 2017.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <http://www.rfc-editor.org/info/rfc7938>.

[RFC8014] Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)", RFC 8014, DOI 10.17487/RFC8014, December 2016, <http://www.rfc-editor.org/info/rfc8014>.

## Authors' Addresses

Zheng(Sandy) Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zhang.zheng@zte.com.cn

Bo Wu
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: wu.bo@zte.com.cn