

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

L. Geng
L. Wang
China Mobile
J. Xie
M. McBride
G. Yan
Huawei Technologies
July 2, 2018

MVPN using Segment Routing and BIER for High Reachability Multicast
Deployment
draft-geng-bier-sr-multicast-deployment-00

Abstract

Bit Index Explicit Replication (BIER) introduces a stateless multicast approach for a specific IGP area. Segment Routing introduces an approach for end-to-end stateless deployment for both inter-area and inter-as scenarios. This document proposes a MVPN using Segment Routing and BIER for a high reachability multicast deployment.

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

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 <https://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 January 3, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem Statement and Considerations	3
3.1. Problem Statement and Considerations	3
4. MVPN Using SR-MPLS and BIER-MPLS Encapsulation	4
4.1. Anchor information Advertisement and Usage	4
4.2. MVPN Forwarding State and Forwarding Procedure	6
5. MVPN Using SRv6 and BIER-IPv6 Encapsulation	7
5.1. Anchor information Advertisement and Usage	7
5.2. MVPN Forwarding State and Forwarding Procedure	7
6. Security Considerations	8
7. IANA Considerations	8
8. Acknowledgements	8
9. References	8
9.1. Normative References	8
9.2. Informative References	9
Authors' Addresses	9

1. Introduction

Bit Index Explicit Replication (BIER) [RFC8279] introduces a stateless multicast approach for a specific IGP area. Segment Routing [I-D.ietf-spring-segment-routing] introduces an approach for end-to-end stateless deployment for both inter-area and inter-as scenario. An end-to-end VPN deployment may benefit from the combination of this two technology in which the stateless nature can be maintained. This document proposes an MVPN deployment with high reachability in such scenario using both Segment Routing and BIER.

2. Terminology

Readers of this document are assumed to be familiar with the terminology and concepts of the documents listed as Normative References.

3. Problem Statement and Considerations

3.1. Problem Statement and Considerations

In a BIER deployment in multi-area or multi-AS network, a segmented MVPN has to be used. As a result, multicast states are created at the segment boundary. The per-flow multicast states are maintained on the routers which are considered beyond of the "MVPN service" sites. This significant disadvantage for multicast service deployment is due to the poor reachability of BIER and is hard to solve solely by BIER itself.

Segment Routing, however, has high reachability for both multi-area and multi-as deployment. VPN services can use pre-defined Segments (SIDs) on the area boundary routers (ABR) or AS boundary routers (ASBR) for end-to-end deployment, without requiring such boundary routers to include per-VPN or per-flow states, or per-VPN or per-flow signaling to establish the end-to-end connection.

BIER and Segment Routing can be used for different partition of an end-to-end MVPN service deployment. A packet with BIER encapsulation is carried by Segment Routing to a boundary router. When reaching the boundary router, it is replicated according to the BitString in the BIER encapsulation to destination routers. Hence, the whole multicast deployment can be stateless end-to-end.

A typical scenario for this type of deployment is in a service-provider network for business L3VPN service with multicast as defined in [I-D.ietf-bier-use-cases]. Service provider network tends to be very heterogeneous with full-mesh backbone network, ring-shaped metro networks for sparse area coverage, and sometime a fabric for dense area coverage. A source router can send multicast packets to each of the boundary routers of each metro network, with a loose path selection in the full-mesh core network to avoid overloading by using Segment Routing. The boundary router or boundary routers replicate the packets to its own metro network according to the BIER encapsulation.

To achieve the end-to-end statelessness, the boundary router will not proxy any per-VPN or per-flow state. Instead, each of the edge routers, in a specific metro network, directly tell the interest of some multicast flow to the ingress edge router. This is the same as

the L3VPN deployed end-to-end on Option-C style or SR style. For MVPN service, this can be done by the current BGP MVPN signaling. While for MVPN using Segment Routing and BIER, it is required to include the information of boundary router(s) of the area the egress edge router belongs to. The boundary router(s) can be thought as anchor(s) of the area for BIER replication.

Below is an example of end-to-end MVPN deployment on a simple network containing one ABR in each of the edge network area.

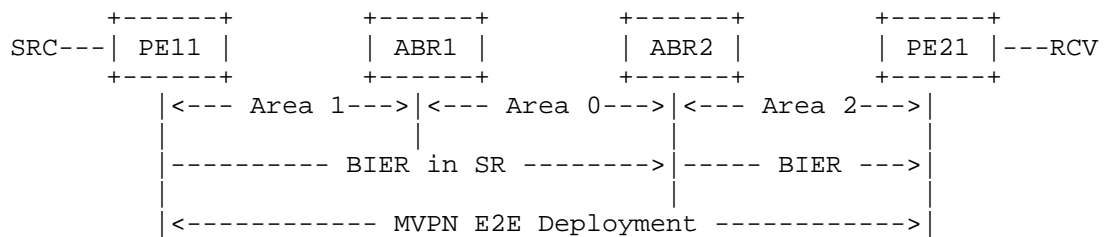


Figure 1: MVPN using BIER and SR for E2E deployment

A more realistic network may contain two ABRs in each metro network area for realibility.

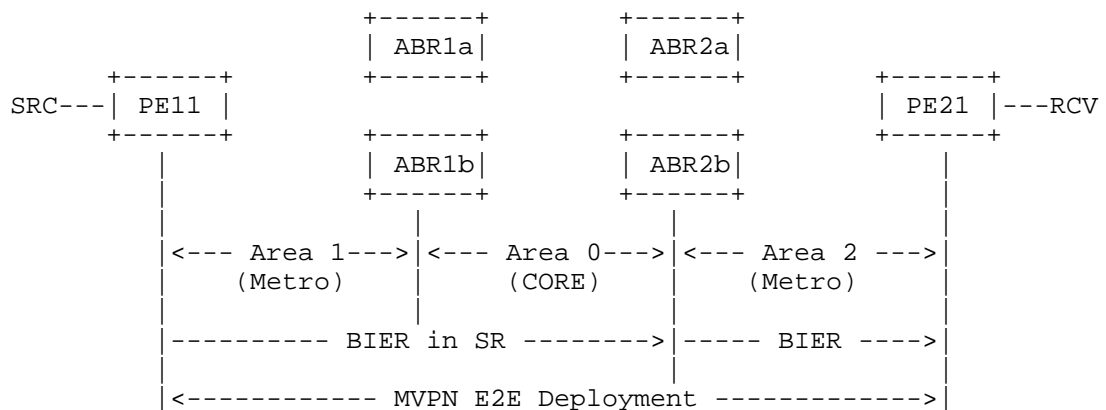


Figure 2: MVPN using BIER and SR for E2E deployment and protection

4. MVPN Using SR-MPLS and BIER-MPLS Encapsulation

4.1. Anchor information Advertisement and Usage

In an area of the receiver side, the anchor router or routers advertise the BIER Label, the router IP, and the associated Sub-domain, BSL and SI. The egress edge routers receive this information

accordingly. When an egress edge router advertiseing MVPN Leaf A-D routes to the ingress edge router at the sender side, it includes the anchor router IP, the anchor router BIER Label, together with the egress edge router's Sub-domain, BFR-prefix and BFR-id, just as the PTA defined in [I-D.ietf-bier-mvpn].

For a deployment where more than one (typically two) anchor routers exist in the area, it is expected to use only one BIER sub-domain for the ease of configuration, while supporting the anchor routers with different BIER labels or with same BIER label (anycast label). The BIER label of an anchor is selected from SRGB and called a BIER SRGB-label. Each of the routers in the area do not have to allocate a local label (from SRLB) for a specific (Sub-domain, BSL, SI) tuple when building the BIER forwarding table. Instead, it uses the BIER SRGB-label for building the BIER forwarding table of the BIER label itself. More than one BIER SRGB labels for the same (Sub-domain, BSL, SI) tuple are allowed, each forming a forwarding table, and the local-allocated (from SRLB) BIER label forwarding table of the same (Sub-domain, BSL, SI) tuple can coexist as well.

Procedures of building the BIER SRGB label forwarding table are outside the scope of this document.

For many areas, it is not required to have a universe-unique sub-domain number or same sub-domain with universe-unique SI number from 0 to 255. For example, it is allowed for area 2 having a sub-domain 0 and SI from 0 to 10, while area 3 having a sub-domain 0 and SI from 0 to 10 too, only if their anchor routers are not the same.

The anchor information of Hybrid SR and BIER MPLS is carried in a specific PTA as below.

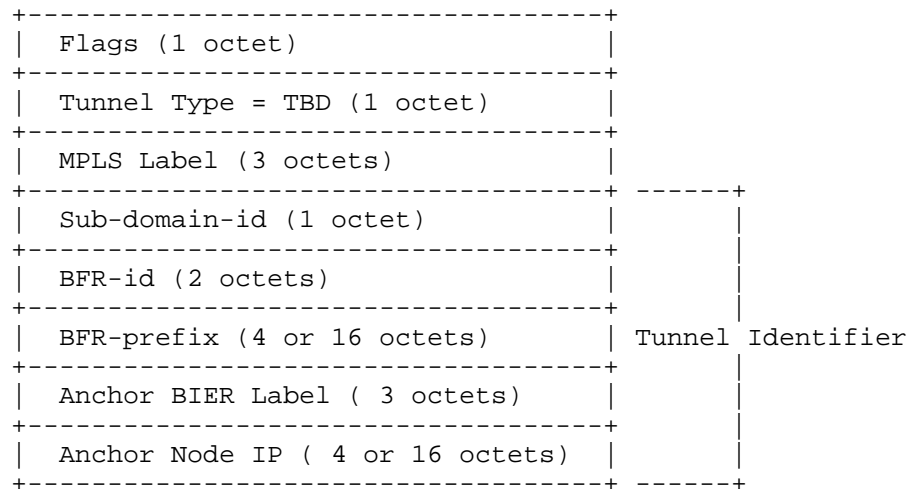


Figure 3: PTA for Hybrid SR and BIER MPLS Tunnel

4.2. MVPN Forwarding State and Forwarding Procedure

Ingress edge router has a per-flow forwarding state, indicating forwarding to every anchor router(s) of an egress area, and a BitString representing the final egress edge routers.

- o (VRF, S, G, Anchor Node SID, Anchor BIER Label of a <SD,BSL,SI>, SD, BSL, SI, BitString of a <SD,BSL,SI>).

Ingress edge router can have its own policy about how to reach some anchor router.

Each of the anchor router(s) has a per-SRGB-label BIER forwarding state, but don't have any per-VPN or per-flow state. When an anchor router receives a BIER packet encapsulated in the Segment Routing label, it pops the Segment Routing label, sees the BIER SRGB-label, and performs hop-by-hop BIER replication with BIER SRGB-label MPLS encapsulation. The hop-by-hop BIER forwarding can further change to on-hop replications directly to the egress edge routers over Segment Routing tunnels, by building BIER forwarding table over Segment Routing on anchor router(s) and egress edge routers only.

Each egress edge router has a per-flow forwarding state, indicating forwarding a packet to its interfaces connected to CE or receivers. Egress edge router can use the upstream-assigned vpnlabel to differentiate the local VRF.

5. MVPN Using SRv6 and BIER-IPv6 Encapsulation

MVPN service using SRv6 and BIER IPv6 Encapsulation is also possible by using the [I-D.xie-bier-6man-encapsulation], which allows BIER packets to run on a SRv6 tunnel.

Procedures of building the BIER IPv6 BIFT-ID forwarding table are outside the scope of this document.

5.1. Anchor information Advertisement and Usage

The anchor information of Hybrid SPv6 and BIER IPv6 is carried in a specific PTA as below.

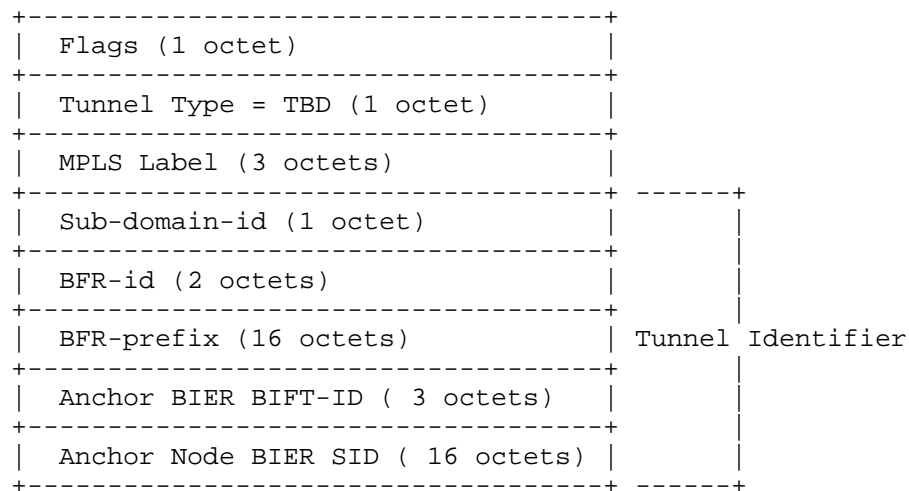


Figure 4: PTA for Hybrid SRv6 and BIER IPv6 Tunnel

5.2. MVPN Forwarding State and Forwarding Procedure

Ingress edge router has a per-flow forwarding state, indicating forwarding to every anchor router(s) of an egress area.

- o (VRF, S, G, Anchor Node BIER SID, Anchor BIER BIFT-ID of a <SD,BSL,SI>, SD, BSL, SI, BitString of a <SD,BSL,SI>).

Ingress edge router can have its own policy about how to reach some anchor router.

Each of the anchor router(s) has a per-BIFT-ID BIER forwarding state, but doesn't have any per-VPN or per-flow state. When an anchor router receives a BIER packet encapsulated in the SRv6 SRH header, it

first pops the SRH, and then sees the BIER specific Multicast address, and then performs the hop-by-hop BIER replication by using the BIFT-ID and other BIER header fields as described in [I-D.xie-bier-6man-encapsulation].

Egress edge router has a per-flow forwarding state, indicating forwarding a packet to its interfaces connected to CE or receivers. Egress edge router can use the upstream-assigned vpnlabel to differentiating the local VRF.

6. Security Considerations

The procedures of this document do not, in themselves, provide privacy, integrity, or authentication for the control plane or the data plane.

7. IANA Considerations

Allocation is expected from IANA for two new tunnel type codepoints for "Hybird SR-MPLS and BIER MPLS Tunnel" and "Hybird SRv6 and BIER IPv6 Tunnel" from the "P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types" registry.

8. Acknowledgements

TBD.

9. References

9.1. Normative References

[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-use-cases]

Kumar, N., Asati, R., Chen, M., Xu, X., Dolganow, A., Przygienda, T., Gulko, A., Robinson, D., Arya, V., and C. Bestler, "BIER Use Cases", draft-ietf-bier-use-cases-06 (work in progress), January 2018.

[I-D.ietf-spring-segment-routing]

Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.

- [I-D.xie-bier-6man-encapsulation]
Xie, J., Yan, G., McBride, M., and Y. Xia, "Encapsulation for BIER in Non-MPLS IPv6 Networks", draft-xie-bier-6man-encapsulation-00 (work in progress), April 2018.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [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>>.
- [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <<https://www.rfc-editor.org/info/rfc8296>>.

9.2. Informative 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>>.

Authors' Addresses

Liang Geng
China Mobile
Beijing 10053

Email: gengliang@chinamobile.com

Lei Wang
China Mobile
Beijing 10053

Email: wangleiyjy@chinamobile.com

Jingrong Xie
Huawei Technologies

Email: xiejingrong@huawei.com

Mike McBride
Huawei Technologies

Email: mmcbride7@gmail.com

Gang Yan
Huawei Technologies

Email: yangang@huawei.com

BIER Workgroup
Internet Draft
Intended status: Standard Track

H. Bidgoli, Ed.
A. Dolganow
J. Kotalwar
Nokia

Expires: January 3, 2019

July 2, 2018

M-LDP Signaling Through BIER Core
draft-hj-bier-mldp-signaling-00

Abstract

Bit Index Explicit Replication (BIER) is an architecture that provides multicast forwarding through a "BIER domain" without requiring intermediate routers to maintain multicast related per-flow state. Neither does BIER require an explicit tree-building protocol for its operation. A multicast data packet enters a BIER domain at a "Bit-Forwarding Ingress Router" (BFIR), and leaves the BIER domain at one or more "Bit-Forwarding Egress Routers" (BFERs). The BFIR router adds a BIER header to the packet. Such header contains a bit-string in which each bit represents exactly one BFER to forward the packet to. The set of BFERs to which the multicast packet needs to be forwarded is expressed by the according set of bits switched on in BIER packet header.

This document describes the procedure needed for mLDP tunnels to be signaled and stitched through a BIER core. Allowing LDP routers to run traditional Multipoint LDP services through a BIER core.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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." The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on October 8, 2017.

Copyright Notice

Copyright (c) 2018 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 carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
2.1. Definitions	3
3. M-LDP Signaling Through BIER domain	5
3.1. Ingress BBR procedure	5
3.2. Assigning the BIER sub-domain Tree Label	6
3.2.1. Method 1: IBBR as extraction point to SDN controller	6
3.2.2. Method 2, EBBR assigning the BTL	7
3.2.2.1. BIER packet construction at IBBR for Method 2	7
3.2.2.2. Signaling mLDP through the BIER domain procedure	8
3.3. SDN Controller procedure for updating BBRs with BTL	8
3.4. BGP procedure for signaling BTL for method 2	9
3.5. EBBR procedure method	9
3.6. Label release	10
3.6.1 Label release in method 1	10
3.6.2 Label release in method 2	10
4. Datapath Forwarding	10
4.1. BFIR tracking of FEC	10
4.2. Datapath traffic flow	10
5. Recursive FEC	11
6. IANA Considerations	11
7. Security Considerations	11
8. References	11
8.1. Normative References	11
8.2. Informative References	11

7. Acknowledgments	12
Authors' Addresses	12

1. Introduction

This draft extends draft-ietf-bier-pim-signaling to mLDP.

Some operators would like to deploy BIER technology in some segment of their network. This draft explains a method to signal mLDP services and stitch it to a BIER domain, with minimal disruption and operational impact to the mLDP domain.

This draft explains the procedures needed to signal and uniquely identify a mLDP P2MP LSP in a BIER domain.

2. Conventions used in this document

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].

2.1. Definitions

Some of the terminology specified in [I-D.draft-ietf-bier-architecture-05] is replicated here and extended by necessary definitions:

BIER:

Bit Index Explicit Replication (The overall architecture of forwarding multicast using a Bit Position).

BFR:

Bit Forwarding Router (A router that participates in Bit Index Multipoint Forwarding). A BFR is identified by a unique BFR-prefix in a BIER domain.

BFIR:

Bit Forwarding Ingress Router (The ingress border router that inserts the Bit Map into the packet). Each BFIR must have a valid BFR-id assigned. BFIR is term used for dataplain packet forwarding.

BFER:

Bit Forwarding Egress Router. A router that participates in Bit Index Forwarding as leaf. Each BFER must be a BFR. Each BFER must have a valid BFR-id assigned. BFER is term used for dataplain packet forwarding.

BBR:

BIER Boundary router. The router between the LDP domain and BIER domain. Maintains mLDP adjacency for all routers attached to it on the mLDP domain and terminates the mLDP adjacency toward the BIER domain.

IBBR:

Ingress BIER Boundary Router. The ingress router from signaling point of view. It maintains mLDP adjacency toward the LDP domain and determines if the mLDP FEC needs to be signaled across the BIER domain. If so it terminates the mLDP adjacency toward the BIER domain and signals the mLDP FEC through the BIER core. The router also signals the FEC withdraw or release.

EBBR:

Egress BIER Boundary Router. The egress router in BIER domain from signaling point of view. It terminates the BIER packet sends the mLDP FEC to LDP module to be signaled through the LDP domain.T:

BFT:

Bit Forwarding Tree used to reach all BFERS in a domain.

BIFT:

Bit Index Forwarding Table.

BIER sub-domain:

A further distinction within a BIER domain identified by its unique sub-domain identifier. A BIER sub-domain can support multiple BitString Lengths.

BFR-id:

An optional, unique identifier for a BFR within a BIER sub-domain.

3. M-LDP Signaling Through BIER domain

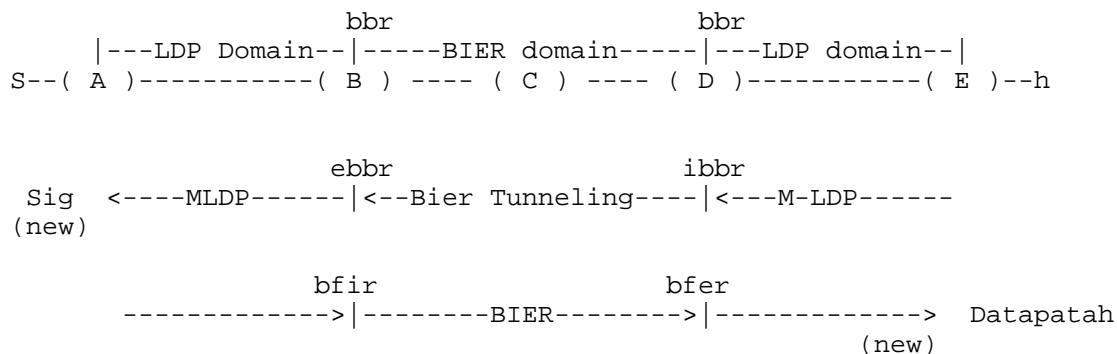


Figure 1: bier boundary router

As per figure 1, the procedures of mLDP signaling is done at the BIER boundary routers. The BIER boundary router (BBR) are connected to LDP capable routers toward the LDP domain and BIER routers toward the BIER domain. LDP routers in LDP domain continue to send LDP signaling messages to the BBR. The BBR will create LDP adjacency between all the LDP routers attach to it on the LDP domain. That said the BBR does not propagate the LDP Signaling packets natively into the BIER domain. Instead when it determines that the label mapping or withdraw message needs to be signaled through the BIER domain, it will execute the procedures in this document to uniquely identify and stitch the P2MP LSP through the BIER domain. These procedures are achievable via an SDN Controller or signaling of the the mLDP FEC through the BIER domain, depending on the method. For the latter method this signaling is not for creating a mLDP adjacency between the two disjoint ldp domains through the BIER network.

The terminology ingress BBR (ibbr) and egress BBR (ebbr) are relative from signaling point of view.

To represent the mLDP P2MP FEC uniquely within the BIER domain there needs to be a BIER TREE Label (BTL). BTL in essence is an MPLS label that represents the P2MP LSP within the BIER domain uniquely. There could be multiple methods used for assigning a BTL to a mLDP P2MP LSP, this is explained in upcoming sections.

3.1. Ingress BBR procedure

IBBR will create LDP adjacency to all LDP routers attach to it toward the LDP domain.

When a mLDP label mapping or withdraw arrives, the IBBR first determines whether the root of the FEC is reachable through the BIER domain. As an example, this root is located on a disjoint LDP domain that is reachable through the BIER domain. If so IBBR will forward the FEC to a BIER label assigning entity to allocated a BIER domain label (BTL) which would uniquely represent this P2MP FEC in the BIER domain. As it will be explained in upcoming sections this "BIER label assigning entity" could be an SDN controller or EBBR.

On forwarding plane the IBBR will track all the LDP interfaces on the attach LDP domain which are signaling the same FEC. It creates a stitching point (ILM entry) between the assigned BTL and labels received via the label mapping message on LDP interfaces in LDP domain. This stitching could be a swap or pop and push. With the same token if a label withdraw arrives the IBBR will remove the stitch mapping and forward the FEC and the action to the "BIER label assigning entity" for appropriate action.

3.2. Assigning the BIER sub-domain Tree Label

There could be 2 methods for assigning a BTL. First via a SDN controller and second via the EBBR. In either method for scalability the BTL needs to be assigned from a label pool represented by EBBR prefixID and its BIER sub-domain <EBBR, SD>.

3.2.1. Method 1: IBBR as extraction point to SDN controller

In the first method IBBR will terminate the MDLP signaling from LDP domain. For label mapping/withdraw signaling packets, IBBR will extract the FEC and the action and forward it to the SDN Controller with its own BIER Prefix. The message format is <<IBBR,SD>, FEC, action>.

The SDN Controller has the entire view of the end to end network or at minimum the view of the BIER domain network. BGP-LS could be used to build this network view. The controller will examine the FEC and find the EBBR closes to the root. The procedure to find the EBBR closest to the root is described in ietf-draft-bier-pim-signaling. After the SDN Controller determines the EBBR it will determine whether this is the first occurrence of this specific mLDP FEC, if so it will assign a BTL from the <EBBR, SD> pool to the FEC to uniquely represent the P2MP tree in the BIER domain.

From this point on the SDN Controller keeps track of all new IBBRs which are interested in this P2MP FEC. The SDN Controller will create

a mapping <<<EBBR, SD> (BTL), FEC, action>>, IBBRs>. The SDN Controller will update the relevant BBRs as described in the SDN Controller section.

3.2.2. Method 2, EBBR assigning the BTL

Alternatively the IBBR can signal the <FEC, action> to EBBR via bier in-band signaling in BIER domain. This signaling could be a new BIER TLV or using the mLDP packet as a signaling packet, in par with draft-ietf-bier-pim-signaling.

IBBR will use the ROOT of the mLDP FEC to find EBBR. The procedure to find EBBR is identical to ietf-draft-bier-pim-signaling. After identifying the EBBR, IBBR will encapsulate the mLDP signaling in a BIER packet with the correct EBBR bit set in the bier header and forward the signaling packet into the BIER sub-domain.

The EBBR will examine the signaling packet and will allocate a BTL from its <EBBR, SD> pool. The EBBR then constructs <<<<EBBR, SD> (BTL), FEC, action>, IBBRs> mapping. EBBR needs to signal the <<<<EBBR, SD> (BTL), FEC> to the relevant IBBRs. This EBBR signaling can be done via SDN Controller or BGP, explained in upcoming sections.

3.2.2.1. BIER packet construction at IBBR for Method 2

Assuming the mLDP packet is forwarded from IBBR to EBBR for signaling, the BIER header will be encoded with the BFR-id of the IBBR(with appropriate bit set in the bitstring) and the mLDP signaling packet is then encapsulated in the packet.

```

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
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     BIFT-id                                     | TC | S |         TTL         |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Nibble | Ver | BSL |                                     Entropy                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
| OAM | Rsv | DSCP | Proto |                                     BFIR-id                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     BitString (first 32 bits)                                     ~
+-----+-----+-----+-----+-----+-----+-----+-----+
~
+-----+-----+-----+-----+-----+-----+-----+-----+
~                                     BitString (last 32 bits)                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

BIERHeader.Proto = IPv4 or IPv6

BIERHeader.BitString= Bit corresponding to the BFR-ID of the EBBR

BIERHeader.BFIR-id = BFR-Id of the BER originating the encapsulated LDP packet, i.e. the IBBR.

Rest of the values in the BIER header are determined based on the network (MPLS/non-MPLS), capabilities (BSL), and network configuration.

3.2.2.2. Signaling mLDP through the BIER domain procedure

Throughout the BIER domain the BIER forwarding procedure is in par with RFC 8279. No BIER transit router will examine the BIER packet encapsulating the mLDP signaling packet.

The packet will be forwarded through the BIER domain until it reaches the BBR with matching BFR-ID as in the BIERHeader.Bitstring. This BBR (EBBR) will remove the BIER header and examine the mLDP signaling packet farther.

3.3. SDN Controller procedure for updating BBRs with BTL

The BBRs can use openflow message to send the mLDP FEC info to the SDN Controller. On the other direction the controller can use BGP SR-TE signaling (or pcep) to download the mapping information to the BBRs. Detail are outside the scope of this document but a new BGP SR-TE NRLI and TLVs would be needed.

In either method the SDN Controller will track the <<<<EBBR, SD> (BTL), FEC, action>, List of <IBBRs>> and download the <<EBBR, SD> (BTL), FEC, action> to the relevant BBRs.

In method 1 the relevant BBRs are EBBR and the IBBRs that are interested in the P2MP LSP. It should be noted since in this method EBBR does not know about all the IBBRs which are interested in a mLDP FEC, as such the SDN Controller should download to EBBR the <<EBBR, SD> (BTL), FEC, action> mapping with the list of all interested IBBRs. In addition to this, every time an IBBR receives the mLDP FEC it should be signaled to SDN Controller via described method and added to the list. SDN Controller should update the EBBR with that info.

In method 2 the EBBR has assigned this mapping and is aware of all IBBRs that are interested in the FEC. EBBR should signal the <<<<EBBR, SD> (BTL), FEC, action>, List of <IBBRs>> to SDN Controller. SDN Controller downloads the mapping to only IBBRs that are interested in the P2MP lsp. As an example the list of <IBBRs>. EBBR will update the SDN controller with each arriving new IBBR that

wants to join the P2MP LSP. The SDN controller will update these new IBBRs with the relevant mapping.

3.4. BGP procedure for signaling BTL for method 2

As described in the "Method 2, EBBR assigning the BTL" section, BGP can be used to signal the <<EBBR, SD> (BTL), FEC> to IBBR. This signaling can be via MP-BGP MVPN address family with a new NLRI route type Intra-AS BTL Route

```

+-----+
|          SD      (8 octets)          |
+-----+
|          EBBR Router's IP Addr       |
+-----+

```

The PMSI tunnel attribute can be used to signal the (BTL, FEC) to IBBRs. A new "BTL" tunnel type needs to be assigned. The PMSI TUNNEL ATTRIBUTE MPLS label field can be used to encoded BTL. The Tunnel Identifier will contain the FEC.

In addition BGP SR-TE could be used, where the EBBR is generating the NLRI. As per SDN controller section this NLRI is a new NLRI, and the BTL will be part of the SID list (single SID).

3.5. EBBR procedure method

After identifying the BTL for the P2MP FEC (via method 1 or method 2) EBBR will assign a up stream label for the FEC and signal it toward root via mLDP signaling.

With same token the EBBR creates a multicast state with incoming interface as same interface that mLDP signaling packet was forwarded and outgoing interfaces of IBBRs BFIR-ids interested in the P2MP FEC.

EBBR will stitch the upstream label to the downstream BTL with out going interfaces the IBBRs interested in this particular P2MP tree and identified via IBBRs BFIR-id. The EBBR will also build a BIER reverse path forwarding table, using the IBBR BFIR-id. This is explained in section 4.1.

It should be noted EBBR will maintain LDP adjacency toward the LDP domain and all LDP routers which are connected to it.

At this point the end-to-end multicast traffic flow setup is complete.

3.6. Label release

Assuming label release is originated from a disjoint LDP domain, the EBBR needs to signal the release of BTL to all IBBRs interested in this particular mLDP FEC. Also the EBBR will remove the ILM entry for this FEC base on the label release.

3.6.1 Label release in method 1

In method 1 the EBBR can signal the label release to the SDN Controller <<<<EBBR, SD> (BTL), FEC, action>, the SDN Controller will find the list of IBBRs interested in this FEC and will update them accordingly. The IBBR will in addition send a label release to all mLDP neighbors on LDP domain that were signaling this FEC. The EBBR and IBBR will remove the ILM entry for this FEC and the SDN Controller will remove the entry and release the BTL for this FEC as well.

3.6.2 Label release in method 2

In method 2 the same procedure as method 1 can be followed when the SDN Controller is used for signaling. In case of BGP signaling of BTL, EBBR will automatically withdraw the BTL route.

4. Datapath Forwarding

4.1. BFIR tracking of FEC

As explained before the BFIR has a ILM entry which stitches arriving P2MP label to the BIER sub-domain Tree Label.

The BFIR (EBBR) also track all the interested BFERs via arriving binding <<<EBBR, SD> (BTL), FEC, action>>, list of (IBBRs)> from SDN Controller (method 1) or the FEC signaling in (method 2). BFIR should build its multicast tree with incoming interface (IIF) as LDP interface (in LDP domain) and out going interfaces OIFs set as the <SD, BFR-IDs> of the interested BFERs (in BIER Domain).

4.2. Datapath traffic flow

On BFIR when the MPLS label for P2MP LSP arrives a lookup in ILM table is done. Base on the arriving label BFIR will find the stitching forwarding entry. BFIR will swap the incoming MPLS label with the assigned BTL. The swap action can also be a pop of mpls domain label and push of the BTL. BFIR will go through all the BTL's out going interface, (i.e. the IBBRs BFIF-id interested in this P2MP lsp). BFIR will put the corresponding BIER header with bit index set for all IBBRs interested in this P2MP LSP. BFIR will set the

BIERHeader.Proto = MPLS and will forward the BIER packet into BIER domain.

In the BIER domain normal BIER forwarding procedure will be done, as per RFC 8279

The IBBRs will receive the BIER packet, will look at the protocol of BIER header (MPLS) and find the EBBR label pool base on the arriving packet BFR-ID and its sub-domain. BFER will remove the BIER header and will do a lookup in the <EBBR, SD> ILM for BTL. The BTL entry could be swap to the MPLS domain P2MP label or a pop of BST and push of MPLS domain P2MP. The MPLS domain will forward the packet as per MPLS forwarding procedure to all the MPLS OIFs on the IBBR.

5. Recursive FEC

The above procedures also will work with a mLDP recursive FEC. The root used to determine the EBBR is the outer root of the FEC. The entire recursive FEC needs to be preserve when it is sent from IBBR to EBBR via the controller or the inband BIER signaling.

6. IANA Considerations

This document contains no actions for IANA.

7. Security Considerations

TBD

8. References

8.1. Normative References

[BIER_ARCH] Wijnands, IJ., Rosen, E., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast using Bit Index Explicit Replication", internet-draft draft-ietf-bier-architecture-08, October 2016.

8.2. Informative References

[BIER_MVPN] Rosen, E., Ed., Sivakumar, M., Wijnands, IJ., Aldrin, S., Dolganow, A., and T. Przygienda, "Multicast VPN Using Bier", internet-draft draft-ietf-bier-mvpn-08, January 2017.

[ISIS_BIER_EXTENSIONS] Ginsberg, L., Przygienda, T., Aldrin, S., and Z. Zhang, "BIER Support via ISIS", internet-draft draft-ietf-bier-isis-extensions-06.txt, March 2017.

[OSPF_BIER_EXTENSIONS] Psenak, P., Kumar, N., Wijnands, IJ., Dolganow, A., Przygienda, T., Zhang, Z., and S. Aldrin, "OSPF Extensions for Bit Index Explicit Replication", internet-draft draft-ietf-ospf-bier-extensions-09.txt, March 2017.

7. Acknowledgments <Add any acknowledgements>

Authors' Addresses

Hooman Bidgoli (editor)
Nokia
600 March Rd.
Ottawa, Ontario K2K 2E6
Canada

Email: hooman.bidgoli@nokia.com

Jayant Kotalwar
Nokia
380 N Bernardo Ave,
Mountain View, CA 94043
US

Email: jayant.kotalwar@nokia.com

Andrew Dolganow
Nokia
750D Chai Chee Rd
06-06, Viva Business Park
Singapore 469004

Email: Andrew.dolganow@nokia.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 9 July 2022

P. Pfister
I.J. Wijnands
S. Venaas
Cisco Systems
C. Wang

Z. Zhang
ZTE Corporation
M. Stenberg
5 January 2022

BIER Ingress Multicast Flow Overlay using Multicast Listener Discovery
Protocols
draft-ietf-bier-mld-06

Abstract

This document specifies the ingress part of a multicast flow overlay for BIER networks. Using existing multicast listener discovery protocols, it enables multicast membership information sharing from egress routers, acting as listeners, toward ingress routers, acting as queriers. Ingress routers keep per-egress-router state, used to construct the BIER bit mask associated with IP multicast packets entering the BIER domain.

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 <https://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 9 July 2022.

Copyright Notice

Copyright (c) 2022 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Overview	4
4. Applicability Statement	4
5. Querier and Listener Specifications	5
5.1. Configuration Parameters	5
5.2. MLDv2 instances.	6
5.2.1. Sending Queries	6
5.2.2. Sending Reports	7
5.2.3. Receiving Queries	8
5.2.4. Receiving Reports	8
5.3. Packet Forwarding	8
6. BIER MLD/IGMP Extension Type	9
7. Security Considerations	10
8. IANA Considerations	10
9. Acknowledgements	11
10. References	11
10.1. Normative References	11
10.2. Informative References	11
Appendix A. BIER Use Case in Data Centers	13
A.1. Convention and Terminology	14
A.2. BIER in data centers	15
A.3. A BIER MLD solution for Virtual Network information	15
Authors' Addresses	16

1. Introduction

The Bit Index Explicit Replication (BIER - [RFC8279]) forwarding technique enables IP multicast transport across a BIER domain. When receiving or originating a packet, ingress routers have to construct a bit mask indicating which BIER egress routers located within the same BIER domain will receive the packet. A stateless approach would consist of forwarding all incoming packets toward all egress routers, which would in turn make a forwarding decision based on local information. But any more efficient approach would require ingress routers to keep some state about egress routers multicast membership information, hence requiring state sharing from egress routers toward

ingress routers.

This document specifies how to use the Multicast Listener Discovery protocol version 2 [RFC3810] (resp. the Internet Group Management protocol version 3 [RFC3376]) as the ingress part of a BIER multicast flow overlay (BIER layering is described in [RFC8279]) for IPv6 (resp. IPv4). It enables multicast membership information sharing from egress routers, acting as listeners, toward ingress routers, acting as queriers. Ingress routers keep per-egress-router state, used to construct the BIER bit mask associated with IP multicast packets entering the BIER domain.

This document defines an MLDv2 and IGMPv3 extension type, using the extension scheme defined in [I-D.ietf-pim-igmp-mld-extension], that is used to provide BIER specific information about the message originator.

This specification is applicable to both IP version 4 and version 6. It therefore specifies two separate mechanisms operating independently. For the sake of simplicity, the rest of this document uses IPv6 terminology. It can be applied to IPv4 by replacing 'MLDv2' with 'IGMPv3', and following specific requirements when explicitly stated.

2. Terminology

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.

The terms "Bit-Forwarding Router" (BFR), "Bit-Forwarding Egress Router" (BFER), "Bit-Forwarding Ingress Router" (BFIR), "BFR-id" and "BFR-Prefix" are to be interpreted as described in [RFC8279].

Additionally, the following definitions are used:

BIER Multicast Listener Discovery (BMLD): The modified version of MLD specified in this document.

BMLD Querier: A BFR implementing the Querier part of this specification. A BMLD Node MAY be both a Querier and a Listener.

BMLD Listener: A BFR implementing the Listener part of this specification. A BMLD Node MAY be both a Querier and a Listener.

3. Overview

This document proposes to use the mechanisms described in MLDv2 in order to enable multicast membership information sharing from BFERs toward BFIRs within a given BIER domain. BMLD queries (resp. reports) are sent over BIER toward all BMLD Nodes (resp. BMLD Queriers) using modified MLDv2 messages which IP destination is set to a configured 'all BMLD Nodes' (resp. 'all BMLD Queriers') IP multicast address.

By running MLDv2 instances with per-listener explicit tracking, BMLD Queriers are able to map BMLD Listeners with MLDv2 membership states. This state is then used to construct the set of BFERs associated with each incoming IP multicast data packet.

4. Applicability Statement

BMLD runs on top of a BIER Layer and provides the ingress part of a BIER multicast flow overlay, i.e, it specifies how BFIRs construct the set of BFERs for each ingress IP multicast data packet. The BFER part of the Multicast Flow Overlay is out of scope of this document.

The BIER Layer MUST be able to transport BMLD messages toward all BMLD Queriers and Listeners. Such packets are IP multicast packets with a BFR-Prefix as source address, a multicast destination address, and containing a MLDv2 message.

BMLD only requires state to be kept by Queriers, and is therefore more scalable than PIMv2 [RFC7761] in terms of overall state, but is also likely to be less scalable than PIMv2 in terms of the amount of control traffic and the size of the state that is kept by individual routers.

This specification is applicable to both IP version 4 and version 6. It therefore specifies two separate mechanisms operating independently. For the sake of simplicity, this document uses IPv6 terminology. It can be applied to IPv4 by replacing 'MLDv2' with 'IGMPv3', and following specific requirements when explicitly stated.

If multiple BFIRs have connectivity to the same source, a mechanism is needed to determine which BFIR should be the forwarder, that is not specified in this document. As a special case, if BIER is used end-to-end such that sources would be directly connected to the BFIRs, then an election mechanism is needed if there are multiple BFIRs on the same link as the source. One option is to utilize PIM DR Election where the DR is the BIER forwarder, but other election mechanisms could be used. In order to allow quick failover, the BFIRs that are not forwarders should still track BFER interest so that they have the correct state in case they become forwarders.

5. Querier and Listener Specifications

Routers desiring to receive IP multicast traffic (e.g., for their own use, or for forwarding) MUST behave as BMLD Listeners. Routers receiving IP multicast traffic from outside the BIER domain, or originating multicast traffic, MUST behave as BMLD Queriers.

BMLD Queriers (resp. BMLD Listeners) MUST act as MLDv2 Queriers (resp. MLDv2 Listeners) as specified in [RFC3810] unless stated otherwise in this section.

5.1. Configuration Parameters

Both Queriers and Listeners MUST operate as BFIRs and BFERs within the BIER domain in order to send and receive BMLD messages. They MUST therefore be configured accordingly, as specified in [RFC8279].

All Listeners MUST be configured with an 'all BMLD Queriers' multicast address and the BFR-ids of all the BMLD Queriers. This is used by Listeners to send BMLD reports over BIER toward all Queriers. All Queriers MUST be configured to accept BMLD reports sent to this address.

All Queriers MUST be configured with an 'all BMLD Nodes' multicast address and the BFR-ids of all the Queriers and Listeners. This information is used by Queriers to send BMLD queries over BIER toward all BMLD Nodes. All BMLD Nodes MUST be configured to accept BMLD queries sent to this address.

It may be cumbersome to configure the exact set of BFR-ids for Queriers and Listeners. One MAY configure the set of BFR-ids to contain any potentially used BFR-id, perhaps having all bit positions set. There is no harm in configuring unused BFR-ids. Configuring the BFR-ids of additional routers would in most cases cause no harm, as a router would drop the BMLD message unless it is configured as a Querier or a Listener.

Note that BMLD (unlike MLDv2) makes use of per-instance configured multicast group addresses rather than well-known addresses so that multiple instances of BMLD (using different group addresses) can be run simultaneously within the same BIER domain. Configured group addresses MAY be obtained from allocated IP prefixes using [RFC3306]. One MAY choose to use the well-known MLDv2 addresses in one instance, but different instances MUST use different addresses.

IP packets coming from outside of the BIER domain and having a destination address set to the configured 'all BMLD Queriers' or the 'all BMLD Nodes' group address MUST be dropped. It is RECOMMENDED that these configured addresses have a limited scope, enforcing this behavior by scope-based filtering on BIER domain's egress interfaces.

5.2. MLDv2 instances.

BMLD Queriers MUST run a MLDv2 Querier instance with per-host tracking, which means they keep track of the MLDv2 state associated with each BMLD Listener. For that purpose, Listeners are identified by their respective BFR-Prefix, used as IP source address in all BMLD reports.

BMLD Listeners MUST run a MLDv2 Listener instance expressing their interest in the multicast traffic they are supposed to receive for local use or forwarding.

BMLD Listeners and Queriers MUST NOT run the MLDv1 (IGMPv2 and IGMPv1 for IPv4) backward compatibility procedures.

5.2.1. Sending Queries

BMLD Queries are IP packets sent over BIER by BMLD Queriers:

- * Toward all BMLD Nodes (i.e., providing to the BIER Layer the BFR-ids of all BMLD Nodes).
- * Without the IPv6 router alert option [RFC2711] in the hop-by-hop extension header [RFC8200] (or the IPv4 router alert option [RFC2113] for IPv4).
- * With the IP destination address set to the 'all BMLD Nodes' group address.
- * With a deterministic IP source address. It is RECOMMENDED that the address is a BFR-Prefix of the sender, but it MAY be another value. This address is only used for querier election.

- * With a TTL value large enough such that the packet can be received by all BMLD Nodes, depending on the underlying BIER layer (whether it decrements the IP TTL or not) and the size of the network. The default value is 64.
- * The extension type defined in Section 6 MUST be included once, specifying the Sub-domain-id, BFR-id and BFR-Prefix of the sender. This information may be useful for logging and debugging.

5.2.2. Sending Reports

BMLD Reports are IP packets sent over BIER by BMLD Listeners:

- * Toward all BMLD Queriers (i.e., providing to the BIER layer the BFR-ids of all BMLD Queriers).
- * Without the IPv6 router alert option [RFC2711] in the hop-by-hop extension header [RFC8200] (or the IPv4 router alert option [RFC2113] for IPv4).
- * With the IP destination address set to the 'all BMLD Queriers' group address.
- * With a deterministic IP source address. It is RECOMMENDED that the address is a BFR-Prefix of the sender.
- * With a TTL value large enough such that the packet can be received by all BMLD Queriers, depending on the underlying BIER layer (whether it decrements the IP TTL or not) and the size of the network. The default value is 64.
- * The extension type defined in Section 6 MUST be included once, specifying the Sub-domain-id, BFR-id and BFR-Prefix of the sender. This information is used to create the necessary forwarding state for requested flows, and may be useful for logging and debugging.

Since the reports may contain a large number of records, they may become larger than the maximum BIER payload that can be delivered to all the BMLD Queriers. Hence an implementation will need to either use a small default maximum size, allow configuration of a maximum size, or rely on MTU discovery. MTU discovery may be done for a sub-domain using BIER MTU Discovery [I-D.ietf-bier-mtud] or for the set of BMLD Queriers using Path MTU Discovery [I-D.ietf-bier-path-mtu-discovery].

5.2.3. Receiving Queries

BMLD Queriers and Listeners MUST check the destination address of all the IP packets that are received or forwarded over BIER whenever their own BIER bit is set in the packet. If the destination address is equal to the 'all BMLD Nodes' group address the packet is processed as specified in this section.

If the IPv6 (resp. IPv4) packet contains an ICMPv6 (resp. IGMP) message of type 'Multicast Listener Query' (resp. of type 'Membership Query'), and include the extension defined in Section 6), it is processed by the MLDv2 (resp. IGMPv3) instance run by the BMLD Querier. It MUST be dropped otherwise.

During the MLDv2 processing, the packet MUST NOT be checked against the MLDv2 consistency conditions (i.e., the presence of the router alert option, the TTL equaling 1 and, for IPv6 only, the source address being link-local).

5.2.4. Receiving Reports

BMLD Queriers MUST check the destination address of all the IP packets that are received or forwarded over BIER whenever their own BIER bit is set. If the destination address is equal to the 'all BMLD Queriers' the packet is processed as specified in this section.

If the IPv6 (resp. IPv4) packet contains an ICMPv6 (resp. IGMP) message of type 'Multicast Listener Report Message v2' (resp. 'Version 3 Membership Report'), and include the extension defined in Section 6), it is processed by the MLDv2 (resp. IGMPv3) instance run by the BMLD Querier. It MUST be dropped otherwise.

During the MLDv2 processing, the packet MUST NOT be checked against the MLDv2 consistency conditions (i.e., the presence of the router alert option, the TTL equaling 1 and, for IPv6 only, the source address being link-local).

5.3. Packet Forwarding

BMLD Queriers configure the BIER Layer using the information obtained using BMLD, and the extension Section 6), to track membership state, including the Sub-domain-id, BFR-id and BFR-Prefix of the members.

More specifically, the membership state associated with each BMLD Listener is provided to the BIER layer such that whenever a multicast packet enters the BIER domain, if that packet matches the membership information from a BMLD Listener, its Sub-domain-id and BFR-id is added to the set of Sub-domains and BFR-ids the packet should be forwarded to by the BIER-Layer.

6. BIER MLD/IGMP Extension Type

A new MLD/IGMP extension type adds BIER specific information to IGMP/MLD messages, using the extension scheme defined in [I-D.ietf-pim-igmp-mld-extension]). The BIER specific information is the same as the PTA tunnel identifier in [RFC8556] and is shown in Figure 1. Note that, as defined in the MLD (resp. IGMP), existing implementations are supposed to ignore this additional data.

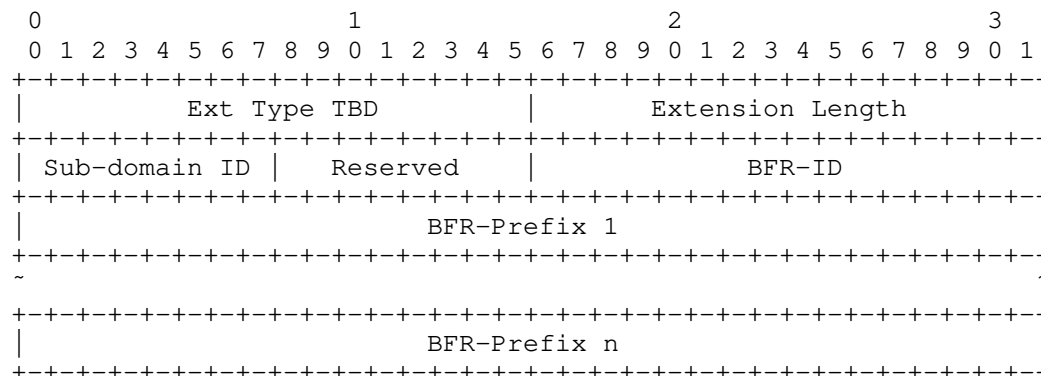


Figure 1: MLD/IGMP Extension Type for BIER

- * Ext Type: Assigned by IANA, identifying this BIER extension.
- * Extension Length: The length in octets of the data after this field. If there are n IPv4 prefixes, the length would be $4 + 4 * n$, if there are n IPv6 prefixes, the length would be $4 + 16 * n$.
- * Sub-domain-id: A single octet containing a BIER sub-domain-id (see [[RFC8279]]). This indicates the BIER sub-domain of the router originating the message.
- * Reserved: A single octet, MUST be set to 0 when sending and ignored when receiving.

- * BFR-id: A two-octet field containing the BFR-id, in the specified sub-domain, of the router originating the message.
- * BFR-prefix: The BFR-prefix (see [[RFC8279]]) of the router that is originating the message. The BFR-prefix will either be a /32 IPv4 address or a /128 IPv6 address.

This extension type MUST be present once in all IGMP and MLD messages when originated with a BIER header to identify the BIER originator. It is expected that any BIER router originating IGMP/MLD messages in BIER supports this specification. Any IGMP/MLD messages that do not contain the extension Section 6) MUST be dropped by the decapsulating router with no processing other than potentially logging or debugging. It is expected that any BIER router processing IGMP/MLD messages with BIER encapsulation supports this specification. If they do not, they will likely ignore the report since they cannot identify the BIER receiver, but they may be able to derive some of the receiver information from the BIER header.

7. Security Considerations

BMLD makes use of IGMPv3/MLDv2 messages transported over BIER in order to configure the BIER Layer of BFIRs. BMLD messages MUST be secured, either by relying on physical or link-layer security, by securing the IP packets (e.g., using IPsec [RFC4301]), or by relying on security features provided by the BIER Layer.

By spoofing the IP source address, an attacker could become the IGMP/MLD querier. Once one becomes the querier, several attack vectors are possible. This is similar to regular IGMP/MLD without BIER encapsulation.

An attacker could send reports with the BIER IGMP/MLD extension Section 6) specifying a BFR-ID and BIER prefix identifying another router. This would allow the attacker to:

- * Redirect undesired traffic toward the spoofed router by subscribing to undesired multicast traffic.
- * Prevent desired multicast traffic from reaching the spoofed router by unsubscribing to some desired multicast traffic.

8. IANA Considerations

This document requests that IANA assigns a new type called BIER information in the registry defined in [I-D.ietf-pim-igmp-mld-extension].

9. Acknowledgements

Comments concerning this document are very welcome.

10. References

10.1. Normative References

- [RFC2113] Katz, D., "IP Router Alert Option", RFC 2113, DOI 10.17487/RFC2113, February 1997, <<https://www.rfc-editor.org/info/rfc2113>>.
- [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>>.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <<https://www.rfc-editor.org/info/rfc3376>>.
- [RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <<https://www.rfc-editor.org/info/rfc3810>>.
- [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>>.
- [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>>.
- [I-D.ietf-pim-igmp-mld-extension] Sivakumar, M., Venaas, S., Zhang, Z., and H. Asaeda, "Internet Group Management Protocol version 3 (IGMPv3) and Multicast Listener Discovery version 2 (MLDv2) Message Extension", Work in Progress, Internet-Draft, draft-ietf-pim-igmp-mld-extension-05, 7 November 2021, <<https://www.ietf.org/archive/id/draft-ietf-pim-igmp-mld-extension-05.txt>>.

10.2. Informative References

- [RFC2711] Partridge, C. and A. Jackson, "IPv6 Router Alert Option", RFC 2711, DOI 10.17487/RFC2711, October 1999, <<https://www.rfc-editor.org/info/rfc2711>>.
- [RFC3306] Haberman, B. and D. Thaler, "Unicast-Prefix-based IPv6 Multicast Addresses", RFC 3306, DOI 10.17487/RFC3306, August 2002, <<https://www.rfc-editor.org/info/rfc3306>>.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<https://www.rfc-editor.org/info/rfc4301>>.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <<https://www.rfc-editor.org/info/rfc5015>>.
- [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, <<https://www.rfc-editor.org/info/rfc7348>>.
- [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>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8556] Rosen, E., Ed., Sivakumar, M., Przygienda, T., Aldrin, S., and A. Dolganow, "Multicast VPN Using Bit Index Explicit Replication (BIER)", RFC 8556, DOI 10.17487/RFC8556, April 2019, <<https://www.rfc-editor.org/info/rfc8556>>.
- [I-D.ietf-bier-mtud]
Venaas, S., Wijnands, I., Ginsberg, L., and M. Sivakumar, "BIER MTU Discovery", Work in Progress, Internet-Draft,

draft-ietf-bier-mtud-00, 27 February 2019,
 <<https://www.ietf.org/archive/id/draft-ietf-bier-mtud-00.txt>>.

[I-D.ietf-bier-path-mtu-discovery]

Mirsky, G., Przygienda, T., and A. Dolganow, "Path Maximum Transmission Unit Discovery (PMTUD) for Bit Index Explicit Replication (BIER) Layer", Work in Progress, Internet-Draft, draft-ietf-bier-path-mtu-discovery-11, 4 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-bier-path-mtu-discovery-11.txt>>.

Appendix A. BIER Use Case in Data Centers

In current data center virtualization, virtual eXtensible Local Area Network (VXLAN) [RFC7348] is a kind of network virtualization overlay technology which is overlaid between NVEs and is intended for multi-tenancy data center networks, whose reference architecture is illustrated as per Figure 2.

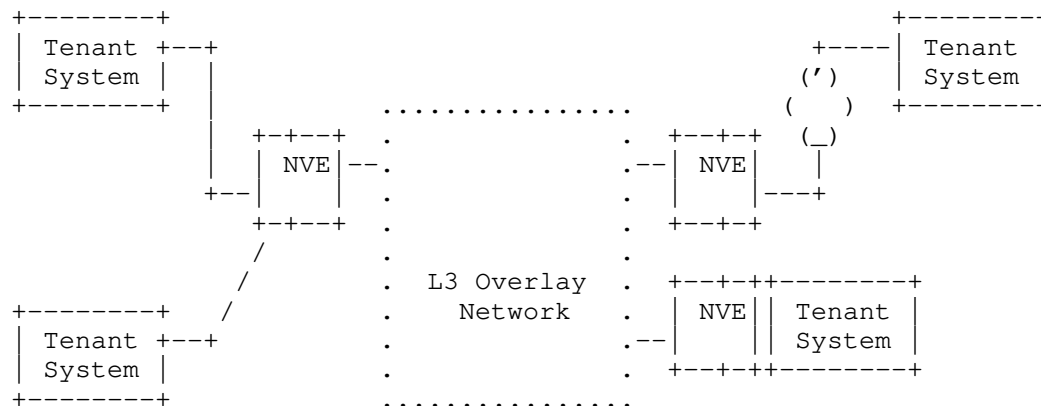


Figure 2: NVO3 Architecture

And there are two kinds of most common methods about how to forward BUM packets in this virtualization overlay network. One is using PIM as underlay multicast routing protocol to build explicit multicast distribution tree, such as PIM-SM [RFC7761] or PIM-BIDIR [RFC5015] multicast routing protocol. Then, when BUM packets arrive at NVE, it requires NVE to have a mapping between the VXLAN Network Identifier and the IP multicast group. According to the mapping, NVE can encapsulate BUM packets in a multicast packet which group address is the mapping IP multicast group address and steer them through explicit multicast distribution tree to the destination NVEs. This method has two serious drawbacks. It need the underlay network

supports complicated multicast routing protocol and maintains multicast related per-flow state in every transit nodes. What is more, how to configure the ratio of the mapping between VNI and IP multicast group is also an issue. If the ratio is 1:1, there should be 16M multicast groups in the underlay network at maximum to map to the 16M VNIs, which is really a significant challenge for the data center devices. If the ratio is n:1, it would result in inefficiency bandwidth utilization which is not optimal in data center networks.

The other method is using ingress replication to require each NVE to create a mapping between the VXLAN Network Identifier and the remote addresses of NVEs which belong to the same virtual network. When NVE receives BUM traffic from the attached tenant, NVE can encapsulate these BUM packets in unicast packets and replicate them and tunnel them to different remote NVEs respectively. Although this method can eliminate the burden of running multicast protocol in the underlay network, it has a significant disadvantage: large waste of bandwidth, especially in big-sized data center where there are many receivers.

BIER [RFC8279] is an architecture that provides optimal multicast forwarding through a "BIER domain" without requiring intermediate routers to maintain any multicast related per-flow state. BIER also does not require any explicit tree-building protocol for its operation. A multicast data packet enters a BIER domain at a "Bit-Forwarding Ingress Router" (BFIR), and leaves the BIER domain at one or more "Bit-Forwarding Egress Routers" (BFERs). The BFIR router adds a BIER header to the packet. The BIER header contains a bit-string in which each bit represents exactly one BFER to forward the packet to. The set of BFERs to which the multicast packet needs to be forwarded is expressed by setting the bits that correspond to those routers in the BIER header. Specifically, for BIER-TE, the BIER header may also contain a bit-string in which each bit indicates the link the flow passes through.

The following sub-sections try to propose how to take full advantage of overlay multicast protocol to carry virtual network information, and create a mapping between the virtual network information and the bit-string to implement BUM services in data centers.

A.1. Convention and Terminology

The terms about NVO3 are defined in [RFC7365]. The most common terminology used in this appendix is listed below.

NVE: Network Virtualization Edge, which is the entity that implements the overlay functionality. An NVE resides at the boundary between a Tenant System and the overlay network.

VXLAN: Virtual eXtensible Local Area Network

VNI: VXLAN Network Identifier

Virtual Network Context Identifier: Field in an overlay encapsulation header that identifies the specific VN the packet belongs to.

A.2. BIER in data centers

This section tries to describe how to use BIER as an optimal scheme to forward the broadcast, unknown and multicast (BUM) packets when they arrive at the ingress NVE in data centers.

The principle of using BIER to forward BUM traffic is that: firstly, it requires each ingress NVE to have a mapping between the Virtual Network Context Identifier and the bit-string in which each bit represents exactly one egress NVE to forward the packet to. And then, when receiving the BUM traffic, the BFIR/Ingress NVE maps the receiving BUM traffic to the mapping bit-string, encapsulates the BIER header, and forwards the encapsulated BUM traffic into the BIER domain to the other BFERs/Egress NVEs indicated by the bit-string.

Furthermore, as for how each ingress NVE knows the other egress NVEs that belong to the same virtual network and creates the mapping is the main issue discussed below. Basically, BIER Multicast Listener Discovery is an overlay solution to support ingress routers to keep per-egress-router state to construct the BIER bit-string associated with IP multicast packets entering the BIER domain. The following section tries to extend BIER MLD to carry virtual network information (such as Virtual Network Context identifier), and advertise them between NVEs. When each NVE receives these information, they create the mapping between the virtual network information and the bit-string representing the other NVEs belonged to the same virtual network.

A.3. A BIER MLD solution for Virtual Network information

The BIER MLD solution allows having multiple MLD instances by having unique pairs of BMLD Nodes and BMLD Querier addresses for each instance. Assume for now that we have a unique instance per VNI and that all BMLD routers are using the same mapping between VNIs and BMLD address pairs. Also for each VNI there is a multicast group used for encapsulation of BUM traffic over BIER. This group may potentially be shared by some or all of the VNIs.

Each NVE acquires the Virtual Network information, and advertises this Virtual Network information to other NVEs through the MLD messages. For a given VNI it sends BMLD reports to the BMLD nodes

address used for that VNI, for the group used for delivering BUM traffic for that VNI. This allows all NVE routers to know which other NVE routers have interest in BUM traffic for a particular VNI. If one attached virtual network is migrated, the NVE will withdraw the Virtual Network information by sending an unsolicited BMLD report. Note that NVEs also respond to periodic queries to BMLD Nodes addresses corresponding to VNIs for which they have interest.

When ingress NVE receives the Virtual Network information advertisement message, it builds a mapping between the receiving Virtual Network Context Identifier in this message and the bit-string in which each bit represents one egress NVE who sends the same Virtual Network information. Subsequently, once this ingress NVE receives some other MLD advertisements which include the same Virtual Network information from some other NVEs, it updates the bit-string in the mapping and adds the corresponding sending NVE to the updated bit-string. Once the ingress NVE removes one virtual network, it will delete the mapping corresponding to this virtual network as well as send withdraw message to other NVEs.

After finishing the above interaction of MLD messages, each ingress NVE knows where the other egress NVEs are in the same virtual network. When receiving BUM traffic from the attached virtual network, each ingress NVE knows exactly how to encapsulate this traffic and where to forward them to.

This can be used in both IPv4 network and IPv6 network. In IPv4, IGMP protocol does the similar extension for carrying Virtual Network information TLV in Version 2 membership report message.

Note that it is possible to have multiple VNIs map to the same pair of BMLD addresses. Provided VNIs that map to the same BMLD address uses different multicast groups for encapsulation, this is not a problem, because each instance is tracking interest for each multicast group separately. If multiple VNIs map to the same pair and the multicast group used is not unique, some NVEs may receive BUM traffic for which they are not interested. An NVE would drop packets for an unknown VNI, but it means wasting some bandwidth and processing. This is similar to the non-BIER case where there is not a unique multicast group for encapsulation. The improvement offered by using BMLD is by using multiple instance, hence reducing the problems caused by using the same transport group for multiple VNIs.

Authors' Addresses

Pierre Pfister
Cisco Systems
Paris
France

Email: pierre.pfister@darou.fr

IJsbrand Wijnands
Cisco Systems
De Kleetlaan 6a
1831 Diegem
Belgium

Email: ice@cisco.com

Stig Venaas
Cisco Systems
Tasman Drive
San Jose, CA 95134
United States of America

Email: stig@cisco.com

Cui(Linda) Wang

Email: lindawangjoy@gmail.com

Zheng(Sandy) Zhang
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
CA,
China

Email: zhang.zheng@zte.com.cn

Markus Stenberg
FI-00930 Helsinki
Finland

Email: markus.stenberg@iki.fi

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 26 January 2022

H. Bidgoli, Ed.
Nokia
F. Xu
Verizon
J. Kotalwar
Nokia
I. Wijnands
M. Mishra
Cisco System
Z. Zhang
Juniper Networks
25 July 2021

PIM Signaling Through BIER Core
draft-ietf-bier-pim-signaling-12

Abstract

Consider large networks deploying traditional PIM multicast service. Typically, each portion of these large networks have their own mandates and requirements. It might be desirable to deploy BIER technology in some part of these networks to replace traditional PIM services. In such cases downstream PIM states need to be signaled over the BIER Domain toward the source.

This draft specifies the procedure to signal PIM join/prune messages through a BIER Domain, as such enabling the provisioning of traditional PIM services through a BIER Domain. These procedures are valid for forwarding PIM join/prune messages to the Source (SSM) or Rendezvous Point (ASM).

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 <https://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 26 January 2022.

Copyright Notice

Copyright (c) 2021 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
2.1. Definitions	3
3. PIM Signaling Through BIER domain	4
3.1. Ingress BBR procedure	4
3.1.1. New Pim Join Attribute, BIER Information Vector	5
3.1.1.1. BIER packet construction at the IBBR	6
3.2. Signaling PIM through the BIER domain procedure	7
3.3. EBBR procedure	7
4. Datapath Forwarding	7
4.1. Datapath traffic flow	8
5. PIM-SM behavior	8
6. Applicability to MVPN	8
7. IANA Considerations	9
8. Security Considerations	9
9. Acknowledgments	9
10. References	9
10.1. Normative References	9
10.2. Informative References	10
Appendix A. Determining the EBBR	10
A.1. Link-State Protocols	10
A.2. Indirect next-hop	10
A.2.1. Static Route	11
A.2.2. Interior Border Gateway Protocol (iBGP)	11
A.3. Inter-area support	11
A.3.1. Inter-area Route summarization	12
Authors' Addresses	12

1. Introduction

It might be desirable to simplify/upgrade some part of an existing network to BIER technology, removing any legacy multicast protocols like PIM. This simplification should be done with minimum interruption or disruption to the other parts of the network from singling, services and software upgrade point of view. To do so this draft specifies procedures for signaling multicast join and prune messages over the BIER domain, this draft is not trying to create FULL PIM adjacency over a BIER domain between two PIM nodes. The PIM adjacency is terminated at BIER edge routers and only join/prune signaling messages are transported over the BIER network. It just so happened that this draft chose signaling messages to be in par with PIM join/prune messages. These signaling messages are forwarded upstream toward the BIER edge router on path to the Source or Rendezvous point. These signaling messages are encapsulated in a BIER header.

2. Conventions used in this document

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.

2.1. Definitions

An understanding of the BIER architecture [RFC8279] and the related terminology is expected. The following are some of the new definitions used in this draft.

BBR:

BIER Boundary router. A router between the PIM domain and BIER domain. Maintains PIM adjacency for all routers attached to it on the PIM domain and terminates the PIM adjacency toward the BIER domain.

IBBR:

Ingress BIER Boundary Router. An ingress router from signaling point of view. It maintains PIM adjacency toward the PIM domain and signals join/prune messages across the BIER domain to EBBR as needed.

EBBR:

Egress BIER Boundary Router. An egress router in BIER domain from signaling point of view. It maintains PIM adjacency to all upstream PIM routers. It terminates the BIER signaling packets and creates necessary PIM join/prune messages into PIM Domain.

3. PIM Signaling Through BIER domain

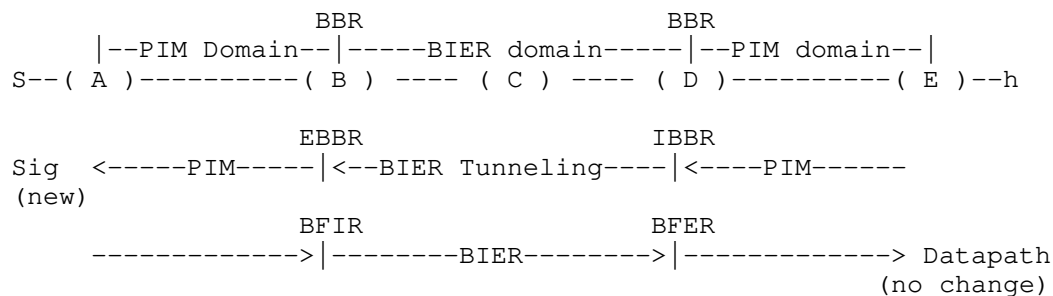


Figure 1: BIER boundary router

Figure 1 illustrates the operation of the BIER Boundary router (BBR). BBRs are connected to PIM routers in the PIM domain and BIER routers in the BIER domain. PIM routers in PIM domain continue to send PIM state messages to the BBR. The BBR will create PIM adjacency between all the PIM routers attached to it on the PIM domain. Each BBR determines if a BIER Signaling Join or Prune message needs to be transmitted through the BIER domain. This draft has chosen these BIER signaling messages to be PIM join/prune message, as such an implementation could chose to tunnel actual PIM join/prune messages through BIER network. This tunneling is only done for signaling purposes and not for creating a PIM adjacency between the two disjoint PIM domains through the BIER domain.

The terminology ingress BBR (IBBR) and egress BBR (EBBR) is relative only from a signaling point of view.

The egress BBR will determine if the arriving BIER packet is a signaling packet and if so it will generate a PIM join/prune packet toward its attached PIM domain.

The new procedures in this draft are only applicable to signaling and there are no changes from datapath point of view.

3.1. Ingress BBR procedure

The IBBR maintains a PIM adjacency [RFC7761] with any PIM router attached to it on the PIM domain.

When a PIM Join or Prune message is received, the IBBR determines whether the Source or RP is reachable through the BIER domain. The EBBR is the BFR through which the Source or RP is reachable. In PIM terms [RFC7761], the EBBR is the RPF Neighbor, and the RPF Interface is the BIER "tunnel" used to reach it. The mechanisms used to find the EBBR are outside the scope of this document and there can be many mechanism depending on if the source or RP are in same area or autonomous system (AS) or in different area or AS -- some examples are provided in Appendix A.

If the lookup for source or rendezvous point results into multiple EBBRs, different IBBRs could choose different EBBRs for the same flow. As long as a unique IBBR chooses a unique EBBR for the same flow. On downstream these EBBRs will send traffic to their corresponding IBBRs.

After discovering the EBBR and its BFR-id, the IBBR MUST use the BIER Information Vector (Section 3.1.1) which is a PIM Join Attribute type [RFC5384]. The EBBR uses this attribute to obtain the necessary BIER information to build its multicast state. The signaling packet, in this case a PIM Join/Prune message, is encapsulated in the BIER Header and forwarded through the BIER domain to the EBBR. The source address of the PIM packets MUST be set to IBBR local BFR-prefix. The destination address MUST be set to ALL-PIM-ROUTERS [RFC7761].

The IBBR will track all the PIM interfaces on the attached PIM domain which are interested in a certain (S,G). It creates multicast states for arriving join messages from PIM domain, with incoming interface as BIER "tunnel" interface and outgoing interface as the PIM domain interface(s) on which PIM Join(s) were received on.

3.1.1. New Pim Join Attribute, BIER Information Vector

The new PIM Join Attribute " BIER Information Vector" is defined as follow based on [RFC5384]

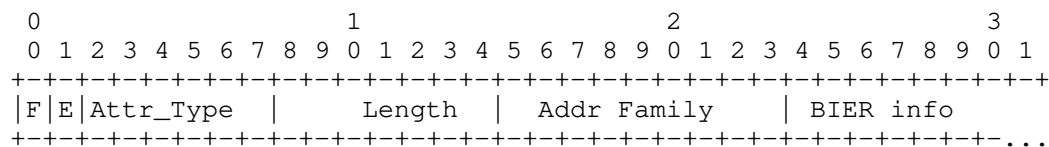


Figure 2: PIM Join Attribute

F bit: Transitive Attribute, as specified in [RFC5384]. MUST be set to zero as this attribute is always non-transitive. If EBBR receives this attribute type with the F bit set it must discard the Attribute.

E bit: End of Attributes, as specified in [RFC5384]

Attr_Type: TBD assign by IANA.

Length: Length of the value field, as specified in [RFC5384]. MUST be set to the length of the BIER Info field + 1. For IPv4 the length is 8, and 20 for IPv6. Incorrect length value compare to the Addr Family must be discarded.

Addr Family: PIM address family as specified in [RFC7761].
Unrecognized Address Family must be discarded.

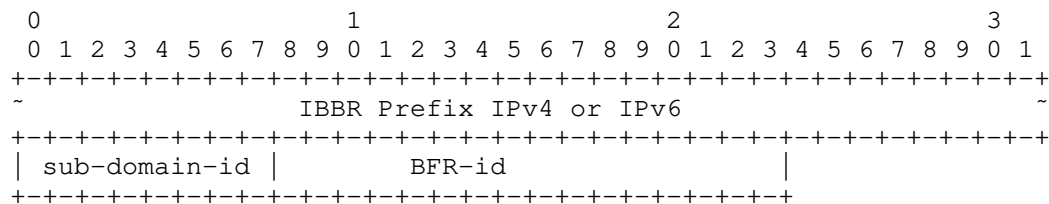


Figure 3: PIM Join Attribute detail

BIER Info: IBBR's BFR-prefix (IPv4 or IPv6), sub-domain-id, BFR-id

3.1.1.1. BIER packet construction at the IBBR

The BIER header will be encoded with the BFR-id of the IBBR (with appropriate bit set in the BitString) and the PIM signaling packet is then encapsulated in the packet.

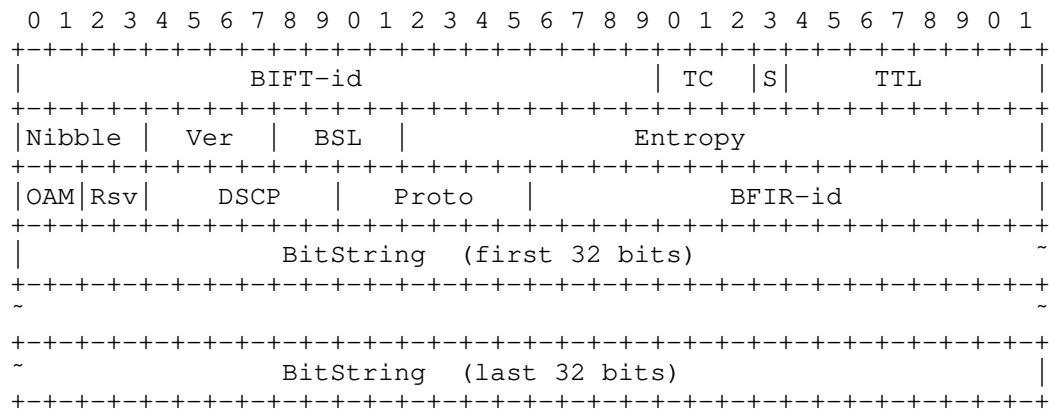


Figure 4: BIER header

BIERHeader.Proto = PIM Address Family

BIERHeader.BitString= Bit corresponding to the BFR-id of the EBBR

BIERHeader.BFIR-id = BFR-Id of the BBR originating the encapsulated signaling packet, i.e. the IBBR.

Rest of the values in the BIER header are determined based on the network (MPLS/non-MPLS), capabilities (BSL), and network configuration.

3.2. Signaling PIM through the BIER domain procedure

Throughout the BIER domain the BIER forwarding procedure is according to [RFC8279]. No BIER router will examine the BIER the signaling packet. As such there is no multicast state built in the BIER domain.

The packet will be forwarded through the BIER domain until it reaches the EBBR indicated by the BIERHeader.Bitstring. Only this targeted EBBR router will remove the BIER header and examine the PIM IPv4 or IPv6 signaling packet further as per EBBR Procedure section.

3.3. EBBR procedure

EBBR removes the BIER Header and determine this is a signaling packet. The Received signaling packet, PIM join/prune message, is processed as if it were received from neighbors on a virtual interface, (i.e. as if the pim adjacency was present, regardless of the fact that there is no adjacency).

The EBBR will build a forwarding table for the arriving (S,G) using the obtained BFIR-id and the Sub-Domain information from BIER Header and/or the PIM join Attributes added to the signaling packet. In short it tracks all IBBRs interested in this (S,G). For a specific Source and Group, EBBR SHOULD track all the interested IBBRs via signaling messages arriving from the BIER Domain. BFER builds its (s,g) forwarding state with incoming interface (IIF) as the Reverse Path Forwarding (RPF) interface (in attached PIM domain) towards the source or rendezvous point. The outgoing interfaces include a virtual interface that represent BIER forwarding to tracked IBBRs.

The EBBR maintains a PIM adjacency [RFC7761] with any PIM router attached to it on the PIM domain. At this point the end-to-end multicast traffic flow setup is complete.

4. Datapath Forwarding

4.1. Datapath traffic flow

When multicast data traffic arrives on the BFIR (EBBR) it forwards the traffic, through the BIER domain, to all interested IBBRs following the procedures specified in [RFC8279]. The BFER(s) (IBBR(s)) also follow the procedures in [RFC8279] and forward the multicast packet through its outgoing interface(s).

5. PIM-SM behavior

The procedures described in this document can be used with Any-Source Multicast (ASM) as long as a static Rendezvous Point (RP) or embedded RP for IPv6 is used [RFC3956].

It should be noted that this draft only signals PIM Joins and Prunes through the BIER domain and not any other PIM message types including PIM Hellos or Asserts. As such functionality related to these other type of messages will not be possible through a BIER domain with this draft and future drafts might cover these scenarios. As an example DR selection should be done in the PIM domain or if the PIM routers attached to IBBRs are performing DR selection there needs to be a dedicated PIM interface between these routers. The register messages are unicasts encapsulated from the source to RP as such they are forwarded without these procedures.

In case of PIM ASM Static RP or embedded RP for IPv6 the procedure for leaves joining RP is the same as above. It should be noted that for ASM, the EBBRs are determined with respect to the RP instead of the source.

6. Applicability to MVPN

With just minor changes, the above procedures apply to MVPN as well, with BFIR/BFER/EBBR/IBBR being VPN PEs. All the PIM related procedures, and the determination of EBBR happens in the context of a VRF, following procedures for PIM-MVPN.

When a PIM packet arrives from PIM domain attached to the VRF (IBBR), and it is determined that the source is reachable via the VRF through the BIER domain, a PIM signaling message is sent via BIER to the EBBR. In this case usually the PE terminating the PIM-MVPN is the EBBR. A label is imposed before the BIER header is imposed, and the "proto" field in the BIER header is set to 1 (for "MPLS packet with downstream-assigned label at top of stack"). The label is advertised by the EBBR/BFIR to associate incoming packets to its correct VRF. In many scenarios a label is already bound to the VRF loopback address on the EBBR/BFIR and it can be used.

When a multicast data packet is sent via BIER by an EBBR/BFIR, a label is imposed before the BIER packet is imposed, and the "proto" field in the BIER header is set to 1 (for "MPLS packet with downstream-assigned label at top of stack"). The label is assigned to the VPN consistently on all VRFs [draft-zzhang-bess-mvpn-evpn-aggregation-label-01].

If the more complicated label allocation scheme is needed for the data packets as specified in [draft-zzhang-bess-mvpn-evpn-aggregation-label-01], then additional PMSI signaling is needed as specified in [RFC6513].

To support per-area subdomain in this case, the ABRs would need to become VPN PEs and maintain per-VPN state so it is unlikely practical.

7. IANA Considerations

IANA is requested to assign a value (TBD) to the BIER Information Vector PIM Join Attribute from the PIM Join Attribute Types registry.

8. Security Considerations

The procedures of this document do not, in themselves, provide privacy, integrity, or authentication for the control plane or the data plane. For a discussion of the security considerations regarding the use of BIER, please see [RFC8279] and [RFC8296]. The security consideration for [RFC7761] also apply.

9. Acknowledgments

The authors would like to thank Eric Rosen, Stig Venaas for thier reviews and comments.

10. References

10.1. Normative References

- [RFC2119] "S. Brandner, "Key words for use in RFCs to Indicate Requirement Levels"", March 1997.
- [RFC5384] "A. Boers, I. Wijnands, E. Rosen, "PIM Join Attribute Format"", November 2008.
- [RFC7761] "B.Fenner, M.Handley, H. Holbrook, I. Kouvelas, R. Parekh, Z.Zhang "PIM Sparse Mode"", March 2016.

- [RFC8174] "B. Leiba, "ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words"", May 2017.
- [RFC8279] "Wijnands, IJ., Rosen, E., Dolganow, A., Przygienda, T. and S. Aldrin, "Multicast using Bit Index Explicit Replication"", October 2016.
- [RFC8296] "IJ. Wijnands, E. Rosen, A. Dolganow, J. Yantsura, S. Aldrin, I. Meilik, "Encapsulation for BIER"", January 2018.

10.2. Informative References

- [draft-zzhang-bess-mvpn-evpn-aggregation-label-01]
"Z. Zhang, E. Rosen, W. Lin, Z. Li, I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common labels"", April 2018.
- [RFC3956] "P. Savola, B. Haberman "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address"". "
- [RFC6513] "E. Rosen, R. Aggarwal, "Multicast in MPLS/BGP IP VPNs"", November 2008.

Appendix A. Determining the EBBR

This appendix provides some examples of routing procedures that can be used to determine the EBBR at the IBBR.

A.1. Link-State Protocols

On IBBR SPF procedures can be used to find the EBBR closest to the source.

Assuming the BIER domain consists of all BIER forwarding routers, SPF calculation can identify the router advertising the prefix for the source. A post process can find the EBBR by walking from the advertising router back to the IBBR in the reverse direction of shortest path tree branch until the first BFR is encountered.

A.2. Indirect next-hop

Alternatively, the route to the source could have an indirect next-hop that identifies the EBBR. These methods are explained in the following sections.

A.2.1. Static Route

A static route to the source can be configured on the IBBR with the next-hop set as the EBBR's BFR-prefix.

A.2.2. Interior Border Gateway Protocol (iBGP)

Consider the following topology:

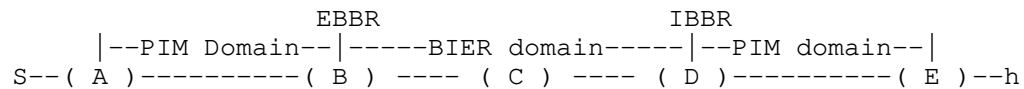


Figure 5: Static Route

Suppose BGP is enable between EBBR (B) and IBBR (D) and the PIM Domain routes are redistributed to the BIER domain via BGP, performing next-hop-self for these routes. This would include the Multicast Source IP address (S). In such case BGP should use the same next-hop as the EBBR BIER prefix. This will ensure that all PIM domain routes, including the Multicast Source IP address (S) are resolve via EBBR's BIER prefix address. When the host (h) triggers a PIM join message to IBBR (D), IBBR tries to resolve (S). It resolves (S) via BGP installed route and realizes its next-hop is EBBR (B).

A.3. Inter-area support

If each area has its own BIER sub-domain, the above procedure for post-SPF could identify one of the ABRs and the EBBR. If a sub-domain spans multiple areas, then additional procedures as described in A.2 is needed.

A.3.1. Inter-area Route summarization

In a multi-area topology, a BIER sub-domain can span a single area. Suppose this single area is constructed entirely of BIER capable routers and the ABRs are the BIER Boundary Routers attaching the BIER sub-domain in this area to PIM domains in adjacent areas. These BBRs can summarize the PIM domain routes via summary routes, as an example for OSPF, a type 3 summary LSAs can be used to advertise summary routes from a PIM domain area to the BIER area. In such scenarios the IBBR can be configured to look up the Source via IGP database and use the summary routes and its Advertising Router field to resolve the EBBR. The IBBR needs to ensure that the IGP summary route is generated by a BFR. This can be achieved by ensuring that BIER Sub-TLV exists for this route. If multiple BBRs (ABRs) have generated the same summary route the lowest Advertising Router IP can be selected or a vendor specific hashing algorithm can select the summary route from one of the BBRs.

Authors' Addresses

Hooman Bidgoli (editor)
Nokia
Ottawa
Canada

Email: hooman.bidgoli@nokia.com

Fengman Xu
Verizon
Richardson,
United States of America

Email: fengman.xu@verizon.com

Jayant Kotalwar
Nokia
Mountain View,
United States of America

Email: jayant.kotalwar@nokia.com

IJsbrand Wijnands
Cisco System
Diegem
Belgium

Email: ice@cisco.com

Mankamana Mishra
Cisco System
Milpitas,
United States of America

Email: mankamis@cisco.com

Zhaohui Zhang
Juniper Networks
Boston,
United States of America

Email: zzhang@juniper.com

BIER
Internet-Draft
Intended status: Standards Track
Expires: December 23, 2018

A. Przygienda
Z. Zhang
Juniper Networks
Jun 21, 2018

BIER Migration Frameworks
draft-przygienda-bier-migration-options-00

Abstract

BIER is a new architecture for the forwarding and replication of multicast data packets. This document defines possible approaches to introduce BIER into networks consisting of a mixture of BFRs and non-BFRs and their respective preconditions and properties.

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

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 <https://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 December 23, 2018.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. 'Naked' MT	3
2.1. Preconditions	3
2.2. Properties	3
3. RFC8279 Section 6.9	4
3.1. Preconditions	4
3.2. Properties	5
4. BIER Specific Algorithm Based Solutions	6
4.1. Preconditions	6
4.2. Properties	6
5. Controller Based Solutions	7
5.1. Preconditions	7
5.2. Properties	7
6. IANA Considerations	7
7. Security Considerations	7
8. Normative References	7
Authors' Addresses	9

1. Introduction

BIER [RFC8279] is a new architecture for the forwarding of multicast data packets. It allows replication through a "multicast domain" and it does not precondition construction of a multicast distribution tree, nor does it precondition intermediate nodes to maintain any per-flow state.

Given that BIER encompasses a novel switching path it can be reasonably expected that in many deployment scenarios, at least initially, a mixture of BFRs and non-BFR (i.e. routers having all or some of the interfaces not being capable of BIER forwarding) will be used and represent what we will call "mixed environments". [RFC8279] offers several suggestions how a mixture of such routers can be handled in the network. The purpose of this memo is to cover other possible deployment options with explanation what preconditions are necessary to apply each of those and what properties and requirements they bring in operational considerations respectively.

The presented sequence of possible solutions follows very loosely an ordering starting with the ones that use "least" amount of additional technologies beside BIER to deploy a "mixed environment". This

serves subsequently to facilitate the introduction of consecutive, more interdependent solutions. Nevertheless, this does not imply that any of the solutions is better or simpler. The "optimal" solution will depend every time on operational realities of the network performing a migration towards BIER deployment.

Any tunnelling technology used when deploying BIER in a "mixed environment" must ensure that in case the tunnel carries other types of traffic beside BIER the tunnel termination point MUST be capable of identifying BIER frames by some means. In case of tunnel carrying only Ethernet frames or MPLS encapsulated traffic [RFC8296] allows to distinguish BIER from other frames.

This document uses terminology defined in [RFC8279].

2. 'Naked' MT

Strictly speaking BIER can be deployed in "mixed environments" without any additional extensions or new technologies in its basic form. Proper use of multi-topology [RFC5120] configuration in IGP will allow separation of BIER capable routers and interfaces in the topology, possibly connected via IGP tunnels to create at minimum a graph of BFRs.

2.1. Preconditions

- o BIER IGP signalling via [I-D.ietf-bier-ospf-bier-extensions] or [RFC8401] and
- o implementation of multi-topology and
- o any kind of tunneling technology that can be viewed as adjacency in IGP.

2.2. Properties

- o Multi-topology has been standardized and used for many years in IGPs and other signalling protocols.
- o The use of multi-topology allows for multicast and unicast traffic to follow (per subdomain) different paths if necessary in case such a behavior is desired operationally.
- o Normal IGP computation results are used as BIER nexthops, i.e. normal SPF nexthops or even TE computation nexthops and techniques like [RFC3906] are applicable.

- o Reconfiguring multi-topology preconditions the touching of both sides of a link in the multi-topology and recomputation of BIER nexthops for the given topology on all routers. On changes in topology the tunnels may need to be reprovisioned depending on technology and protection scheme used.
- o Physical links configured as members of several multi-topologies can be "shared" between subdomains for e.g. protection purposes, i.e. if multi topologies used for different sub-domains are using same physical links, the links will be used by the according sub-domains as well. By adjusting IGP metrics the traffic can be kept separate per subdomain with the possibility of a "fail-over" onto the links with high IGP metric in case of failures. It is even possible to use the same physical topology with each multi-topology carrying different metrics to make different links having different preference for each sub-domain and "separate" traffic per sub-domain that way.
- o Since multi-topology membership is a "per interface" property it allows to manage "partial BFR" routers, i.e. routers where only a subset of interfaces is BIER capable.
- o Multi-topology solution can be combined in case of "mixed environment" with any other solution described in this document that is multi-topology aware.
- o If tunnel metrics are chosen based on purely IGP metrics the solution may load-balance between hop-by-hop BIER path and tunnels which can lead to different timing behavior on each path albeit in case of BIER entropy encompassing a logical flow this should be benign.
- o Multi-topology provides inherently separate routing tables and according statistics.

3. RFC8279 Section 6.9

This section deals with the "re-parenting" solution outlined in Section 6.9 of [RFC8279]. We will deal with the modified step 2) solution in Section 4.

3.1. Preconditions

- o BIER IGP signalling via [I-D.ietf-bier-ospf-bier-extensions] or [RFC8401] and
- o pre-provisioned "static" tunnels that allows "re-parenting" in any possible failure scenario and/or

- o a "dynamic tunneling" technology that can use a unicast tunnel between any pair of nodes in the domain without configuration or setup, e.g. "soft" GRE [RFC2784], LDP [RFC3036] in Downstream Unsolicited mode or Segment Routing [I-D.ietf-spring-segment-routing] are assumed to be deployed through the whole BIER domain.

3.2. Properties

- o When used with dynamic tunnels the solution can automatically "bridge" disconnected areas without necessity to provision multi topology or static tunnel configuration, i.e. this solution can deal with any arbitrary breakage of topology as long the network does not become partitioned. It is equivalent to node protection [RFC5286].
- o IGP's do not have to be aware of the tunnels.
- o BIER traffic strictly follows unicast path only (assuming that the "dynamic tunnels" are following IGP unicast nexthops as well) and with that
 - * all BIER capable routers MUST have enough scale to carry unicast load and
 - * if the unicast next-hop is a non-BIER capable router the router upstream will ingress replicate to all the children on the unicast tree of that next-hop and
 - * BIER may load balance between tunneled and BIER native forwarding paths which can lead to different timing behavior albeit in case of BIER entropy encompassing a logical flow this should be benign.
- o All interfaces on BFRs MUST be capable of BIER forwarding.
- o Dynamic tunneling topologies do not provide extensive OAM normally albeit they may provide node and link failure protection. On the other hand, some "dynamic tunnelling" technologies like segment routing will hold minimum amount of state in the network, i.e. no per-tunnel specific state while providing coverage for any non-partitioning failure.
- o If a tunnel is used to reach the next BFR, the tunnel's own node/link protection provides FRR.
- o Each change in dynamic tunnel signalling (such as LDP) may lead to recomputation of BIFT entries.

4. BIER Specific Algorithm Based Solutions

BIER can support a multitude of BIER Algorithms (BAR) as specified in IGP drafts and [I-D.ietf-bier-bar-ipa] to operate in "mixed environments" and take into consideration BIER specific constraints and properties. While doing that BFRs signal which algorithm they use so the distributed computation delivers consistent results on all BFRs. In its simplest form BAR can define an SPF where non-BFRs are not being put on the candidate list which we denote for the moment as BAR=1 and consider further.

4.1. Preconditions

- o BIER IGP signalling via [I-D.ietf-bier-ospf-bier-extensions] or [RFC8401] and
- o Implementation of non-zero BAR values and
- o any kind of tunneling technology that can be viewed as an adjacency in IGP.

4.2. Properties

- o BAR allows for multicast and unicast traffic to follow different paths if necessary in case such a behavior is desired operationally.
- o BAR could take into account different limitations like e.g. maximum possible fan-out degree on nodes or inter-dependency of sub-domains in same BIER domain.
- o Normal IGP computation can be used easily to compute BAR BIER nexthops while preserving all unicast node and link-protection schemes.
- o Reconfiguring BAR preconditions the touching of all participating BFR.
- o BAR can allow to manage "partial BFR" routers, i.e. routers where only a subset of interfaces is BIER capable if additional information is advertised with BIER sub-TLVs.
- o All interfaces on BFRs MUST be capable of BIER forwarding unless the static tunnels can be "homed" on BIER capable interfaces only.

5. Controller Based Solutions

Ultimately, the according BIRTs and BIFTs can be precomputed by an off-line controller via any algorithm desirable (in a sense similar to Section 4 but being able to take other metrics and constraints in the computation than distributed by IGP possibly) and downloaded.

5.1. Preconditions

- o Controller computing BIRTs and/or BIFTs and downloading them into all BIER nodes and
- o Preferably signalling of a special BAR value on each router to ensure that it is configured to use the according controller downloaded tables.

5.2. Properties

- o Controller based solution can take into account many constraints and metrics that are not distributed network-wide such as provisioning constraints depending on time of day.
- o Centralized controller computation cannot normally react quickly to node or link failures due to delays involved. It is possible that a centralized computation precomputes and installs according link- and node-protection BIER next-hops and installs those in the forwarding path. Depending on delays two set of tables may be necessary where after download to all routers a 'fast switch-over' is performed to minimize holes and traffic losses.

6. IANA Considerations

None.

7. Security Considerations

General BIER security considerations apply and this document does not introduce any new security relevant topics.

Controller based solutions may introduce new security considerations.

8. Normative References

- [I-D.ietf-bier-bar-ipa]
Zhang, Z., Przygienda, T., Dolganow, A., Bidgoli, H., Wijnands, I., and A. Gulko, "BIER Underlay Path Calculation Algorithm and Constraints", draft-ietf-bier-bar-ipa-01 (work in progress), April 2018.

- [I-D.ietf-bier-ospf-bier-extensions]
Psenak, P., Kumar, N., Wijnands, I., Dolganow, A., Przygienda, T., Zhang, Z., and S. Aldrin, "OSPFv2 Extensions for BIER", draft-ietf-bier-ospf-bier-extensions-18 (work in progress), June 2018.
- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.
- [RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, DOI 10.17487/RFC2784, March 2000, <<https://www.rfc-editor.org/info/rfc2784>>.
- [RFC3036] Andersson, L., Doolan, P., Feldman, N., Fredette, A., and B. Thomas, "LDP Specification", RFC 3036, DOI 10.17487/RFC3036, January 2001, <<https://www.rfc-editor.org/info/rfc3036>>.
- [RFC3906] Shen, N. and H. Smit, "Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels", RFC 3906, DOI 10.17487/RFC3906, October 2004, <<https://www.rfc-editor.org/info/rfc3906>>.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <<https://www.rfc-editor.org/info/rfc5120>>.
- [RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <<https://www.rfc-editor.org/info/rfc5286>>.
- [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>>.

- [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <<https://www.rfc-editor.org/info/rfc8296>>.
- [RFC8401] Ginsberg, L., Ed., Przygienda, T., Aldrin, S., and Z. Zhang, "Bit Index Explicit Replication (BIER) Support via IS-IS", RFC 8401, DOI 10.17487/RFC8401, June 2018, <<https://www.rfc-editor.org/info/rfc8401>>.

Authors' Addresses

Tony Przygienda
Juniper Networks

EMail: prz@juniper.net

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 21, 2019

D. Purkayastha
A. Rahman
D. Trossen
InterDigital Communications, LLC
T. Eckert
Huawei
October 18, 2018

Applicability of BIER Multicast Overlay for Adaptive Streaming Services
draft-purkayastha-bier-multicast-http-response-01

Abstract

HTTP Level multicast, using BIER, is described as a use case in BIER Use cases document. HTTP Level Multicast is used in today's video streaming and delivery services such as HLS, AR/VR etc., generally realized over IP Multicast. A realization of "HTTP Multicast" over "IP Multicast" is described. IP multicast is commonly used for IPTV services. DVB and BBF is also developing a reference architecture for IP Multicast service. Few problems with IPMC, such as waste of transmission bandwidth, increase in signaling when there are few users are described. Realization over BIER, through a BIER Multicast Overlay Layer, is described. How BIER Multicast Overlay operation improves over IP Multicast, such as reduction in signaling, dynamic creation of multicast groups to reduce signaling and bandwidth wastage is described. We conclude with few next steps.

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 <https://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 April 21, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Reference Deployment	3
2. Conventions used in this document	5
3. Use cases	5
4. Requirements	6
5. Realization over IP Multicast	6
5.1. Mapping to Requirements	7
5.2. Problems	8
6. Realization over BIER	8
6.1. Description of a "BIER Multicast Overlay" to support HTTP Multicast	9
6.1.1. BIER Multicast Overlay Components	9
6.1.2. BIER Multicast Overlay Operations	10
6.2. Achieving Multicast Responses	12
6.3. BIER multicast Overlay Traffic Management	13
7. Next Steps	13
8. IANA Considerations	14
9. Security Considerations	14
10. Informative References	14
Authors' Addresses	15

1. Introduction

BIER Use Cases document [I-D.ietf-bier-use-cases] describes an "HTTP Level Multicast" scenario, where HTTP Responses are carried over a BIER multicast infrastructure to multiple clients. Especially rate-adaptive HTTP solutions can benefit from the dynamic multicast group membership changes enabled by BIER. For this, the "server side NAP (Network Attachment Point), creates a list of outstanding client side NAP (Network Attachment Point) requests for the same HTTP resource. When the response is available, the list of NAPs with outstanding

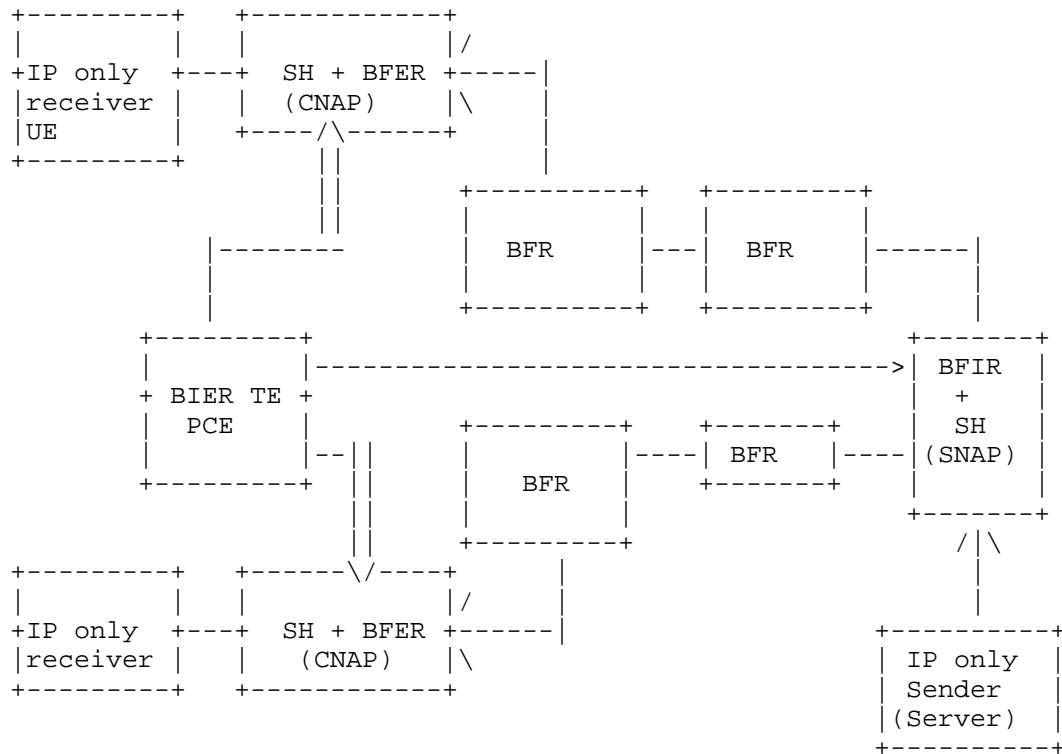
client requests are converted into the BIER or BIER-TE bitstring and used to send the HTTP response.

In this draft, we describe how this class of use cases can be realized over IP Multicast and how the operation of the use case can be improved if realized over BIER. The realization over BIER is achieved through what is called "BIER Multicast overlay" layer, i.e., the methods by which the sending BIER router knows how to send other application packets. The requirements for BIER Multicast overlay layer is described in this document. It also describes the necessary functions that form the BIER multicast overlay and the operations that enable the desired "HTTP Level Multicast" behavior. One such operation is generating the PATH ID (represents the path between BFIR and BFER) based on named service relationship and translating it to appropriate BIER header. We describe a list of protocols needed for the realization of the individual operations.

We conclude with future steps and seek input from the WG.

1.1. Reference Deployment

Let us formulate the architecture of the BIER multicast overlay for the scenario outlined in [I-D.ietf-bier-use-cases]. This overlay is shown in Figure 1 below.



[SH : Service Handler, CNAP : Client Network Attachment Point]
[SNAP : Server Network Attachment Point]
[PCE : Path Computation Element]

Figure 1: Deployment over BIER

The multicast overlay is formed by the BFIR and BFER of the BIER layer and the additional SH (Service Handler) and PCE (Path Computation Element) elements shown in the figure. When interconnecting with a non-BIER enabled IP routed peering network, a special SH, such as Border Gateway may be used.

The Service Handler and BFER can be assumed to be collocated and can be viewed as Client Network Attachment Point (CNAP). Clients sends and receives HTTP transactions through CNAP.

On the server side, the Service handling function can be part of the Server Network Attachment Point (SNAP). It includes the BFIR function and SH. SNAP is responsible for aggregating the relevant

HTTP Requests and sending one or more BIER Multicast HTTP response to multiple clients who requested the same content.

The SH function is assumed to be collocated with BFIR / BFER. The BFIR and BFER is assumed to be normal router boxes in the network. If the additional function of SH cannot be added to normal routers, then SH can be deployed as a separate function outside the routers. In such scenario an interface between SH and BFIR or BFER needs to be defined.

As part of POINT/RIFE EU Horizon 2020 project, HTTP Level Multicast use case has been executed on SDN based and ICN based underlay network, as described in the [I-D.irtf-icnrg-deployment-guidelines].

"HTTP multicast" demonstrated benefits in HTTP-level streaming video delivery, when deployed on POINT test bed with 80+ nodes. This draft [I-D.irtf-icnrg-deployment-guidelines] also describes protocol requirements to enable HTTP multicast to work on ICN underlay.

2. Conventions used in this document

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].

3. Use cases

With the extensive use of "web technology", "distributed services" and availability of heterogeneous network, HTTP has effectively transitioned into the common transport or session layer for E2E and multi-hop communication across the web that is also called Service signaling. Multi-hop when using a sequence of HTTP instance such as HTTP caches. The draft "On the use of HTTP as a Substrate" [I-D.ietf-httpbis-bcp56bis], describes how HTTP is commonly used among service instances to communicate with each other, thus abstracting the lower layer details to application developers.

Referring to the BIER Use Cases [I-D.ietf-bier-use-cases], multicast is used to scale out HLS (HTTP live streaming) to a large number of receivers that use HTTP. This is used today in solutions like DOCSIS hybrid streaming [TR_IPMC_ABR]. Multicast can speed up both live and high-demand VoD streaming. Adaptive Bit Rate IPMC [TR_IPMC_ABR] describes use of IP multicast towards the CMTS or a box beside it, where the content is converted to HTTP/TCP to stream to the receivers (e.g., homes). A server hosting the HLS content is shown as "NAP Server". The gateways acting as receivers for the multicast from the server are shown as "Client-NAP" (CNAP). Each CNAP can serve multiple clients.

HTTP request and response used in media streaming services like HLS, use HTTP response for delivery of content. In such scenarios, where semi-synchronous access to the same resource occurs (such as watching prominent videos over Netflix or similar platforms or live TV over HTTP), traffic grows linearly with the number of viewers since the HTTP-based server will provide an HTTP response to each individual viewer. This poses a significant burden on operators in terms of costs and on users in terms of likely degradation of quality.

This solution is not limited to traditional TV broadcasting. Consider a virtual reality use case where several users are joining a VR session at the same time, e.g., centered around a joint event. Hence, due to the temporal correlation of the VR sessions, we can assume that multiple requests are sent for the same content at any point, particularly when viewing angles of VR clients are similar or the same. Due to availability of virtual functions and cloud technology, the actual end point from where content is delivered may change.

4. Requirements

A realization for the "HTTP multicast" use case may have the following requirements:

- o MUST support multiple FQDN-based service endpoints to exist in the overlay
- o MUST send FQDN-based service requests at the network level to a suitable FQDN-based service endpoint via policy-based selection of appropriate path information
- o MUST allow for multicast delivery of HTTP response to same HTTP request URI
- o MUST provide direct path mobility, where the path between the egress and ingress Service Routers(SR) can be determined as being optimal (e.g., shortest path or direct path to a selected instance), is needed to avoid the use of anchor points and further reduce service-level latency

5. Realization over IP Multicast

IPTV or Internet video distribution in CDNs, uses HTTP Level Multicast and realized over IP Multicast (IPMC). Many features of the IPTV service uses IPMC Group dependent state. Besides popular features like PIM, Mldp, in a variable bit rate encoded content source, content consumption also depends on group state.

DVB released reference architecture [DVB_REF_ARCH] for an end-to-end system to deliver linear content over IP networks in a scalable and standards-compliant manner. It focuses on delivering Adaptive Bit Rate unicast content over a IP multicast network.

A Multicast gateway is deployed in a CPE, Upstream Network Edge device or Terminal and provides multicast to unicast conversion facilities for several homes. All in-scope traffic on the access network between the Multicast Gateway (e.g. network edge device) and the Terminal or home gateway device is unicast. The individual media files are encapsulated into other protocols, so that they can be recovered as discrete files, when they exit the multicast pipe, which is terminated at Multicast Gateway. Interface "L" between Multicast server and Content playback supports fetching of all specified types of Content, Conditional request, Range request, Caching etc. BBF also started similar work in October 2016, called WT-399. This work is now coordinated with DVB. BBF focuses on developing the device management model.

Assume clients that are consuming the same content (such as a TV program) and that this content has for each block (typically segments worth 2 seconds of content) a set of outstanding requests from its clients. When IP Multicast is used in the domain, such as in aforementioned pre-existing solutions like in Cablelabs/DOCSIS [TR_IPMC_ABR], all possible blocks of the content have to be mapped to some IP multicast group, and the CNAP will need to know the mapping of block to groups. For example, a live stream may have 11 different bitrates available. In the most simple Block to IP multicast group mapping scheme, there could be 11 multicast groups, one for all the blocks of one bitrate (note that this is not necessarily done in deployments of this solution, but we consider it here for the purpose of explanation).

If the multicast domain and especially the links into the CNAP has enough bandwidth, this solution work well with IP multicast. As soon as there is at least one Client connected to a CNAP for one particular content, the CNAP would join all 11 multicast groups for this content.

5.1. Mapping to Requirements

To realize "HTTP Level Multicast" over "IP Multicast", some additional functions needs to be supported in an intermediate (overlay) layer.

Support of mapping between FQDN based end points, Multicast Address.
Creating multicast group from FQDN based end points.

Control mechanism related to time when to start sending response as the multicast group is created. It is required that the source should not send response immediately to the Multicast address. Wait for some time to build the group sufficiently and then send response.

Support of IGMP signaling between User device, NAPs and Multicast Router.

5.2. Problems

If the number of clients on a CNAP for a particular program is large, the approach will work fairly well, because the likelihood that each of the 11 bitrates of a content is necessary for at least one Client is then fairly high.

When the number of receivers is not very large, IP multicast runs into two issues. If all the bitrates for the content are sent across the same group, then many of the bitrates may not be required and would have to be received unnecessarily and dropped by the CNAP. If each bitrate was sent on a different IP multicast group, the CNAP could dynamically join/leave each multicast group based on the known receivers, but that would create an extremely high and undesirable amount of IP multicast signaling protocol activity (PIM/IGMP) that is easily overloading the network

For efficiency reasons, the CNAP would need to dynamically join to only those bitrate streams where it does have outstanding requests, therefore achieving the best efficiency. This would mean in the worst case that a CNAP would need to send for each new block, aka.: every two second for every client one IGMP/PIM leave and one IGMP/PIM join towards the upstream router to get a block for an appropriate bitrate (or changed content) whenever bitrate or content on a client have changed. This high rate of control-plane signaling between CNAP and routers, and even between routers inside the multicast Domain is a major pain point and may easily prohibit deployment of these solutions because in many network devices, the performance of PIM/IGMP is not scaled for continuous change in forwarding. Even worse, the limit may not simply be the CPU performance of the routers control plane, but a limitation in the number of changes in forwarding that the forwarding plane units (NPU/ASICs) can support.

6. Realization over BIER

6.1. Description of a "BIER Multicast Overlay" to support HTTP Multicast

The Service Handler (as in Figure 1) in BIER Multicast Overlay, process the FQDN in the service request. At the service level, e.g. HTTP service, the fixed relationship among consumer and providers may be abstracted using "Service Names", and the changing relationship at the Service execution endpoints can be managed at the "multicast overlay" level, handing out the exact locations where service request or response needs to be sent to BIER layer.

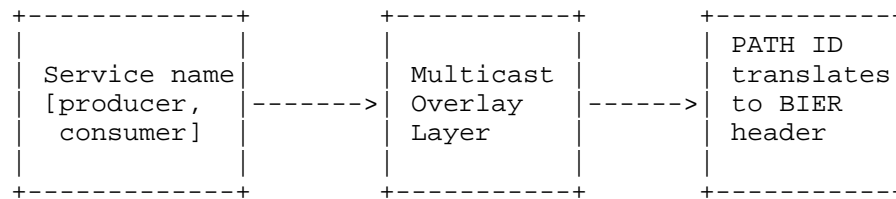


Figure 2: Service name to Path ID translation

We illustrate this using HTTP URI as service names. It should be noted, other identifiers can also be used as service name, such as an IP address. In the example illustration, other layers such as TCP, IP has been terminated at the egress point. Outside BIER domain we terminate TCP/IP session to extract the URI. The URI is processed by the "multicast overlay" layer to generate PATH IDENTIFIER, which is used as BIER header.

Path Identifier or PATH ID, is used in path-based approach, which utilizes path information provided by the source of the packet for forwarding said packet in the network. This is similar to segment routing albeit differing in the type of information provided for such source-based forwarding.

Once the BIER header is determined and added at the BFIR, the rest of the transport layers is assumed to be any underlay technology as supported by BIER. We assume TCP friendly transport, which can assure reliable delivery.

6.1.1. BIER Multicast Overlay Components

With reference to Figure 1, the following components are part of BIER Multicast Overlay Layer.

- o Service Handler (SH): The Service handler terminates transport level protocols, such as TCP, and extracts the URI. It processes the URI in order to determine the PATH ID by contacting the PCE for a suitable path resolution, which in turn is used to send the HTTP Request.
- o Optional PCE : Path Computation Element keeps track of all service execution end points through a registration process. SH interacts with the PCE to obtain PATH information by resolving the FQDN from the incoming URI at the ingress SH to a suitable PATH ID.
- o Interface functions to BFIR where the PATH ID is mapped to BIER header. An Interface to the BFER is likely not required because the BFER will only receive the traffic that they need and should be able to derive from the BIER payload which subset of its receivers need to get an HTTP encapsulated version of a particular reply.

6.1.2. BIER Multicast Overlay Operations

As shown in Figure 3, the "Multicast overlay function" includes a function called PCE (Path Computation Element function), which is responsible for selecting the correct multicast end point and possibly realizing path policy enforcement. The result of the selection is a BIER path identifier, which is delivered to the SH upon initial path computation request (or provided to the ingress router BFIR to be added as BIER header) (i.e., when sending a request to or response for a specific URL for the first time). The path identifier is utilized for any future request for a given URL-based request.

All service end points indicate availability to the PCE through a registration procedure, the PCE will instruct all SHs to invalidate previous path identifiers to the specific URL that might exist. This may result in an a renewed path computation request at the next service request forwarding. Through this, the newly registered service endpoint might be utilized if the policy-governed path computation selects said service instance. Otherwise, a previously resolved PATH ID for the URI determined at the ingress SH is being used instead, removing any resolution latency to an SH-local lookup of the PATH ID.

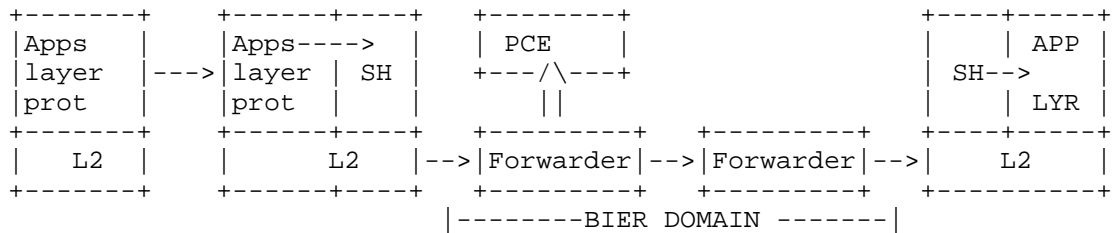


Figure 3: Protocol for Multicast Overlay Layer

In the diagram shown above, an HTTP request is sent by an IP-based device towards the FQDN of the server defined in the HTTP request.

At the client facing SH, the HTTP request is terminated at the TCP level at a local HTTP proxy. The server side SH at the egress terminates any transport protocol on the outgoing (server) side. These terminating functions are assumed to be part of the client/server SH. As a consequence, the SH obtains the destination "Service Name" from the received HTTP request.

If no local BIER forwarding information exists at the client side SH, the path computation entity (PCE) is consulted, which calculates a unicast path from the BFIR to which the client SH is connected to the BFER to which the server SH is connected. The PCE provides the forwarding information (Path ID) to the client SH, which in turn caches the result. The Client SH may forward the Path ID to BFIR, which creates the BIER header.

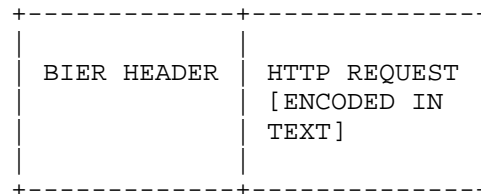


Figure 4: Encapsulation of Service Request

Ultimately, the "HTTP Request" encapsulated by BIER header, as shown in above diagram, is forwarded by the client SH towards the server-facing SH via the local BFIR. We assume a (TCP-friendly) transport protocol being used for the transmission between client and server SH. The possibility of sending one HTTP response to several CNAPs makes this a reliable multicast transport protocol. The exact nature

of this transport protocol is left for further studies. A suitable transport or Layer 2 encapsulation, as supported by BIER layer, is added to the above payload.

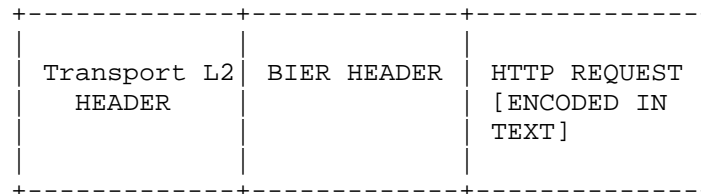


Figure 5: Transport Encapsulation of BIER payload

Upon arrival of an HTTP request at the server SH, it forwards the HTTP request as a well-formed HTTP request locally to the server, awaiting an HTTP response for the reverse direction.

If no BIER forwarding information exists for the reverse direction towards the requesting client SH, this information is requested from the PCE, similar to the operation in forward direction.

6.2. Achieving Multicast Responses

Upon arrival of any further client SH request at the server SH to an HTTP request whose response is still outstanding, the client SR is added to an internal request table. Optionally, the request is suppressed from being sent to the server.

Upon arrival of an HTTP response at the server SH, the server SH consults its internal request table for any outstanding HTTP requests to the same request. The server SH retrieves the stored BIER forwarding information for the reverse direction for all outstanding HTTP requests and determines the path information to all client SHs through a binary OR over all BIER forwarding identifiers with the same SI field. This newly formed joint BIER multicast response identifier is used to send the HTTP response across the network.

BIER makes the solution scalable. Instead of IP multicast with IGMP/PIM, BIER is being used between Server NAP (SNAP) and CNAP, the SNAP simply coalesces the forwarded HTTP requests from the CNAP, and determines for every requested block the set of CNAPs requesting it. A set of CNAPs corresponds to a set of bits in the BIER-bitstring, one bit per CNAP. The SNAP then sends the block into BIER with the appropriate bitstring set.

This completely eliminates any dynamic multicast signaling between CNAP and SNAP. It also avoids sending of any unnecessary data block, which in the IP multicast solution is pretty much unavoidable.

Furthermore, using the approach with BIER, the SNAP can also easily control how long to delay sending of blocks. For example, it may wait for some percentage of the time of a block (e.g, 50% = 1 second), therefore ensuring that it is coalescing as many requests into one BIER multicast answer as possible.

6.3. BIER multicast Overlay Traffic Management

BIER-TE (BIER Traffic Engineering [I-D.ietf-bier-te-arch]) forwards and replicates packets like BIER based on a BitString in the packet header. Where BIER forwards and replicates its packets on shortest paths towards BFER, BIER-TE allows (and requires) to also use bits in the bitstring to indicate the paths in the BIER domain across which the BIER-TE packets are to be sent. This is done to support Traffic Engineering for BIER packets via explicit hop-by-hop and/or loose hop forwarding of BIER-TE packets. A BIER-TE controller calculates explicit paths for this packet forwarding.

The Multicast Flow Overlay operates as in BIER. Instead of interacting with the BIER layer, it interacts with the BIER-TE Controller.

In this draft, "Name-based" service forwarding over BIER, is described to handle changes in service execution end points and manage adhoc relationship in a multicast group. BIER-TE is another way of doing this, while integrated with BIER architecture. The PCE function described earlier in the BIER Multicast Overlay, may become part of BIER-TE Controller. The SH function in the CNAP and SNAP communicates with BIER TE controller. SH sends the service name to the controller, which process the request using the PCE function and returns the "bitstring" to be used as BIER header for delivery of the HTTP response to multiple clients.

7. Next Steps

This Applicability Statement document describes how HTTP multicast responses can be realized over BIER. This document describes the functionalities in the multicast overlay layer to enable this functionality. We would like to get feedback and support from the WG to continue this work. We will elaborate further on specific protocols for the overlay layer and request adoption as a WG draft.

8. IANA Considerations

This document requests no IANA actions.

9. Security Considerations

The operations in Section 6 consider the forwarding of HTTP packets between ingress and egress points based on information derived from the HTTP request. The support for HTTPS is foreseen to ensure suitable encryption capability of such exchanges. Future updates to this draft will outline the support for such HTTPS-based exchanges.

10. Informative References

[DVB_REF_ARCH]

DVB, "Adaptive media streaming over IP multicast", DVB Document A176, March 2018, <https://www.dvb.org/resources/public/standards/a176_adaptive_media_streaming_over_ip_multicast_2018-02-16_draft_bluebook.pdf>.

[I-D.ietf-bier-te-arch]

Eckert, T., Cauchie, G., Braun, W., and M. Menth, "Traffic Engineering for Bit Index Explicit Replication (BIER-TE)", draft-ietf-bier-te-arch-00 (work in progress), January 2018.

[I-D.ietf-bier-use-cases]

Kumar, N., Asati, R., Chen, M., Xu, X., Dolganow, A., Przygienda, T., Gulko, A., Robinson, D., Arya, V., and C. Bestler, "BIER Use Cases", draft-ietf-bier-use-cases-07 (work in progress), July 2018.

[I-D.ietf-httpbis-bcp56bis]

Nottingham, M., "On the use of HTTP as a Substrate", draft-ietf-httpbis-bcp56bis-05 (work in progress), May 2018.

[I-D.irtf-icnrg-deployment-guidelines]

Rahman, A., Trossen, D., Kutscher, D., and R. Ravindran, "Deployment Considerations for Information-Centric Networking (ICN)", draft-irtf-icnrg-deployment-guidelines-04 (work in progress), September 2018.

[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>>.

[TR_IPMC_ABR]

CableLabs, "IP Multicast Adaptive Bit Rate Architecture
Technical Report", OC-TR-IP-MULTI-ARCH-V01-141112 C01,
October 2016, <<https://community.cablelabs.com/wiki/plugins/servlet/cablelabs/alfresco/download?id=51b3c11a-3ba4-40ab-b234-42700e0d4669;1.0>>.

Authors' Addresses

Debashish Purkayastha
InterDigital Communications, LLC
Conshohocken
USA

Email: Debashish.Purkayastha@InterDigital.com

Akbar Rahman
InterDigital Communications, LLC
Montreal
Canada

Email: Akbar.Rahman@InterDigital.com

Dirk Trossen
InterDigital Communications, LLC
64 Great Eastern Street, 1st Floor
London EC2A 3QR
United Kingdom

Email: Dirk.Trossen@InterDigital.com
URI: <http://www.InterDigital.com/>

Toerless Eckert
Huawei USA - Futurewei Technologies Inc.
2330 Central Expy
Santa Clara 95050
USA

Email: tte+ietf@cs.fau.de

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: December 30, 2018

S. Venaas
IJ. Wijnands
L. Ginsberg
Cisco Systems, Inc.
M. Sivakumar
Juniper Networks
June 28, 2018

BIER MTU Discovery
draft-venaas-bier-mtud-01

Abstract

This document defines an IGP based mechanism for discovering the MTU of a BIER sub-domain. This document defines extensions to OSPF and IS-IS, but other protocols could potentially be extended. MTU discovery is usually done for a given path, while this document defines it for a sub-domain. This allows the computed MTU to be independent of the set of receivers. Also, the MTU is independent of rerouting events within the sub-domain.

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 <https://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 December 30, 2018.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	2
3. MTU discovery procedure	3
4. IS-IS BIER Sub-Domain MTU Sub-sub-TLV	4
5. OSPF BIER Sub-Domain MTU Sub-TLV	4
6. IANA considerations	5
7. Acknowledgments	5
8. References	5
8.1. Normative References	5
8.2. Informative References	6
Authors' Addresses	6

1. Introduction

This document defines an IGP based mechanism for discovering the MTU of a BIER sub-domain. The discovered MTU indicates the largest possible BIER payload, such as an IP packet, that can be sent across any link in a BIER sub-domain. This is different from [I-D.ietf-bier-path-mtu-discovery] which performs Path MTU Discovery (PMTUD) for a set of receivers. PMTUD is based on probing, and when there are routing changes, e.g., a link going down, the actual MTU for a path may become less than was previously discovered, and there will be some delay until the next probe is performed. Also, the set of receivers for a flow may change at any time, which may cause the MTU to change. This document instead discovers a BIER sub-domain MTU, which is independent of paths and receivers within the sub-domain.

Discovering the sub-domain MTU is much simpler than discovering the multicast path MTU, and is more robust with regards to path changes as discussed above. However, the sub-domain MTU may be a lot smaller than the path MTU would have been for a given flow. The discovery mechanisms may be combined, allowing the discovery of the path MTU for certain flows as needed.

2. Conventions used in this document

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.

3. MTU discovery procedure

An interface on a router is said to be a BIER interface if the router has a BIER neighbor on the interface. That is, there is a directly connected router on that interface that is announcing a BIER prefix. Further, the BIER interface is said to belong to a given sub-domain if the router itself announces a prefix tagged with the sub-domain, and there is BIER neighbor on the interface also announcing a prefix tagged with the sub-domain.

The BIER MTU of an interface is the largest BIER encapsulated payload that can be sent out of the interface. Further, the local sub-domain MTU of a router is the minimum of all the BIER MTUs of the BIER interfaces in the sub-domain. Note that the local sub-domain MTU of a router is only defined if it has at least one BIER interface in the sub-domain.

A BIER router announces a BIER prefix in either IS-IS or OSPF as specified in [RFC8401] and [I-D.ietf-bier-ospf-bier-extensions]. They both define a BIER Sub-TLV to be included with the prefix. There is one BIER Sub-TLV included for each sub-domain. This document defines how a router includes its local sub-domain MTU in each of the BIER Sub-TLVs it advertizes.

A router can discover the MTU of a BIER sub-domain by identifying all the prefixes that have a BIER Sub-TLV for the sub-domain. It then computes the minimum of the advertised MTU values for that sub-domain. This includes its own local sub-domain MTU. This allows all the routers in the sub-domain to discover the same sub-domain wide MTU.

Note that a router should announce a new local MTU for a sub-domain immediately if the value becomes smaller than what it currently announces. This would happen if the MTU of an interface is configured to a smaller value, or the first BIER neighbor for a sub-domain is detected on an interface, and the MTU of the interface is less than all the other local BIER interfaces in the sub-domain. However, if BIER neighbors go away, or if an interface goes down, so that the local MTU becomes larger, a router SHOULD NOT immediately announce the larger value. A router MAY after some delay announce the new larger MTU. The intention is that dynamic events such as a quick link flap should not cause the announced MTU to be increased.

4. IS-IS BIER Sub-Domain MTU Sub-sub-TLV

A router uses the BIER Sub-Domain MTU Sub-sub-TLV to announce the minimum BIER MTU of all its BIER enabled interfaces in a sub-domain. The BIER Sub-Domain MTU is the largest BIER encapsulated payload that can be sent out of the interfaces in a sub-domain. The Sub-sub-TLV MUST be ignored if it is included multiple times.

```

      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                       |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Type: TBD

Length: 2

MTU: MTU in octets

5. OSPF BIER Sub-Domain MTU Sub-TLV

A router uses the BIER Sub-Domain MTU Sub-TLV to announce the minimum BIER MTU of all its BIER enabled interfaces in a sub-domain. The BIER Sub-Domain MTU is the largest BIER encapsulated payload that can be sent out of the interfaces in a sub-domain. The Sub-TLV MUST be ignored if it is included multiple times.

```

      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                       |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      MTU                         |      Reserved                     |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Type: TBD2

Length: 4

MTU: MTU in octets

6. IANA considerations

An allocation from the "sub-sub-TLVs for BIER Info sub-TLV" registry as defined in [RFC8401] is requested for the IS-IS BIER Sub-Domain MTU Sub-sub-TLV. Please replace the string TBD in this document with the appropriate value.

An allocation from the "OSPF Extended Prefix sub-TLV" registry as defined in [RFC7684] is requested for the OSPF BIER Sub-Domain MTU Sub-TLV. Please replace the string TBD2 in this document with the appropriate value.

7. Acknowledgments

The authors would like to thank Greg Mirsky in particular for fruitful discussions and input. Valuable comments were also provided by Toerless Eckert, Tony Przygienda and Xie Jingrong.

8. References

8.1. Normative References

- [I-D.ietf-bier-ospf-bier-extensions]
Psenak, P., Kumar, N., Wijnands, I., Dolganow, A., Przygienda, T., Zhang, Z., and S. Aldrin, "OSPFv2 Extensions for BIER", draft-ietf-bier-ospf-bier-extensions-18 (work in progress), June 2018.
- [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>>.
- [RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <<https://www.rfc-editor.org/info/rfc7684>>.
- [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>>.
- [RFC8401] Ginsberg, L., Ed., Przygienda, T., Aldrin, S., and Z. Zhang, "Bit Index Explicit Replication (BIER) Support via IS-IS", RFC 8401, DOI 10.17487/RFC8401, June 2018, <<https://www.rfc-editor.org/info/rfc8401>>.

8.2. Informative References

[I-D.ietf-bier-path-mtu-discovery]

Mirsky, G., Przygienda, T., and A. Dolganow, "Path Maximum Transmission Unit Discovery (PMTUD) for Bit Index Explicit Replication (BIER) Layer", draft-ietf-bier-path-mtu-discovery-04 (work in progress), June 2018.

Authors' Addresses

Stig Venaas
Cisco Systems, Inc.
Tasman Drive
San Jose CA 95134
USA

Email: stig@cisco.com

IJsbrand Wijnands
Cisco Systems, Inc.
De kleetlaan 6a
Diegem 1831
Belgium

Email: ice@cisco.com

Les Ginsberg
Cisco Systems, Inc.
Tasman Drive
San Jose CA 95134
USA

Email: ginsberg@cisco.com

Mahesh Sivakumar
Juniper Networks
1133 Innovation Way
Sunnyvale CA 94089
USA

Email: sivakumar.mahesh@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

J. Xie
Huawei Technologies
L. Geng
L. Wang
China Mobile
G. Yan
M. McBride
Y. Xia
Huawei
July 2, 2018

Encapsulation for BIER in Non-MPLS IPv6 Networks
draft-xie-bier-6man-encapsulation-01

Abstract

Bit Index Explicit Replication (BIER) introduces a new multicast-specific BIER Header. Currently BIER has two types of encapsulation formats: one is MPLS encapsulation, the other is Ethernet encapsulation. This document proposes a BIER IPv6 encapsulation for Non-MPLS IPv6 Networks using an IPv6 Destination Option extension header.

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

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 <https://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 January 3, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem Statement and Requirements	3
3.1. Problem Statement	3
3.2. Requirements	4
4. IPv6 BIER Encapsulation	4
4.1. Considerations	4
4.2. IPv6 BIER Destination Option	4
4.3. The whole IPv6 header for BIER packets	5
5. BIER Forwarding in Non-MPLS IPv6 Networks	7
6. Security Considerations	8
7. IANA Considerations	8
8. Acknowledgements	8
9. References	8
9.1. Normative References	8
9.2. Informative References	9
Authors' Addresses	9

1. Introduction

Bit Index Explicit Replication (BIER) [RFC8279] is an architecture that provides optimal multicast forwarding without requiring intermediate routers to maintain any per-flow state by using a multicast-specific BIER header. [RFC8296] defines two types of BIER encapsulation formats: one is MPLS encapsulation, the other is non-MPLS encapsulation. The Non-MPLS encapsulation defined in [RFC8296] is in fact an Ethernet encapsulation with an ethertype 0xAB37, and an 'Ethernet encapsulation' will be used to refer to such an encapsulation in the following text. This document proposes a BIER IPv6 encapsulation for Non-MPLS IPv6 Networks using an IPv6 Destination Option extension header.

2. Terminology

Readers of this document are assumed to be familiar with the terminology and concepts of the documents listed as Normative References.

3. Problem Statement and Requirements

3.1. Problem Statement

MPLS is a very popular and successful encapsulation. One of the benefits of MPLS is its ability to easily stack a label onto another, thus forming a label stack. This same label stacking benefit is also available for BIER by using an MPLS encapsulation. For example, an MPLS-encapsulated BIER packet can easily run over an MPLS tunnel, either a legacy RSVP-TE/LDP LSP, or an MPLS Segment Routing tunnel. Such a mechanism is the key to obtain the capability of "fast reroute" or "bypass a Non-capable router". To quote [RFC8279]:

- o In the event that unicast traffic to the BFR-NBR is being sent via a "bypass tunnel" of some sort, the BIER-encapsulated multicast traffic sent to the BFR-NBR SHOULD also be sent via that tunnel. This allows any existing "fast reroute" schemes to be applied to multicast traffic as well as to unicast traffic.
- o Unicast tunnels are used to bypass non-BFRs.

Some other scenarios also need BIER to run on a tunnel, such as transferring a BIER packet through a whole Non-BIER network or domain.

The capability to run BIER on a tunnel, especially the widely deployed mpls tunnel, can be obtained by using a BIER MPLS encapsulation, but cannot be obtained by using a BIER Ethernet encapsulation. It is not possible either, to run BIER on other links such as POS, by using BIER Ethernet encapsulation.

The capability of running BIER on various kinds of links and tunnels, by using an MPLS encapsulation, is beneficial to BIER deployments. In an IPv6 network, however, there are considerations of using a non-MPLS encapsulation for unicast as the data-plane, such as SRH defined in [I-D.ietf-6man-segment-routing-header], where the function of a bypass tunnel uses an SRH header, with one or many Segments (or SIDs), instead of MPLS Labels.

3.2. Requirements

This chapter lists the BIER IPv6 encapsulation requirements needed to make the deployment of BIER on IPv6 network with SRH data-plane the same as on IPv4/IPv6 network with MPLS data-plane. These BIER IPv6 encapsulation requirements should provide similar benefits to MPLS encapsulation such as "fast reroute" or "run on any link or interface".

1. The listed requirements MUST be supported with any L1/L2 over which BIER layer can be realized.
2. It SHOULD support a hop-by-hop replication to multiple destinations in a BIER Domain.
3. It SHOULD support BIER on an "SRH tunnel".
4. It SHOULD align with the recommendations of the 6MAN working group.

4. IPv6 BIER Encapsulation

4.1. Considerations

BIER is generally a hop-by-hop and one-to-many architecture, while Segment Routing is a source-routing and one-to-one architecture. One of the challenges of an BIER IPv6 Encapsulation is how to allow BIER to run over a Segment Routing tunnel. A suitable method for such a combination is to use a Multicast Address as the Last Segment (or SID). After all the source-routing hops have been processed, the remaining Multicast Address becomes the IPv6 Destination Address. A hop-by-hop replicating diagram begins by using the Destination Multicast Address.

We then need to decide where to place the BIER header. According to [RFC8200], [RFC6564], and [RFC7045], a suitable place for a well-known BIER header is an IPv6 Destination Option extension header. Such a Destination Option carrying BIER header is only used for a hop-by-hop Multicast Address destination, but not for the transit router along the source-routing path.

4.2. IPv6 BIER Destination Option

The IPv6 BIER Destination Option is carried by the IPv6 Destination Option Header (indicated by a Next Header value 60). It is used in a packet sent by an IPv6 BFIR router to inform the routers in an IPv6 BIER domain to replicate to destination BFER routers.

The IPv6 BIER Destination Option is encoded in type-length-value (TLV) format as follows:

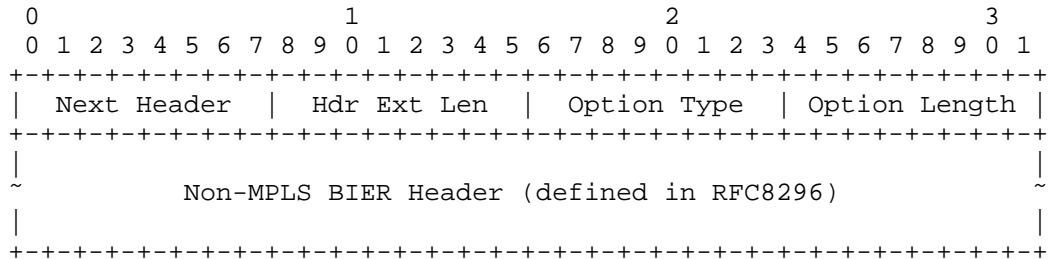


Figure 1: IPv6 BIER Destination Option

Next Header 8-bit selector. Identifies the type of header immediately following the Destination Options header.

Hdr Ext Len 8-bit unsigned integer. Length of the Destination Options header in 8-octet units, not including the first 8 octets.

Option Type TBD. Need to be allocated by IANA.

Option Length 8-bit unsigned integer. Length of the option, in octets, excluding the Option Type and Option Length fields.

Non-MPLS BIER Header The Non-MPLS BIER Header defined in RFC8296, including the BIFT-id.

4.3. The whole IPv6 header for BIER packets

[RFC8200] specifies that the Destination Option Header can be located either before the Routing Header or after the Routing Header. However, this document requires that the Destination Option Header with a BIER Destination Option TLV is always located after the Routing Header if the Routing Header is present.

This is because the BIER header is always handled after the tunnels (or bypass tunnels) have been handled. BIER MPLS encapsulation has the same behavior. To quote [RFC8296]:

- o It is crucial to understand that in an MPLS network the first four octets of the BIER encapsulation header are also the last four octets of the MPLS header. Therefore, any prior MPLS label stack entries MUST have the S bit (see [RFC3032]) clear (i.e., the S bit must be 0).

Other IPv6 extension headers are not commonly used in the current Internet. For Example, [RFC6744] says that "IPv6 Destination Options headers, and the options carried by such headers, are extremely uncommon in the deployed Internet". [RFC6564] says that "Extension headers, with the exception of the Hop-by-Hop Options header, are not usually processed on intermediate nodes", and that "Reports from the field indicate that some IP routers deployed within the global Internet are configured either to ignore the presence of headers with hop-by-hop behavior or to drop packets containing headers with hop-by-hop behavior."

Such IPv6 extension headers will even be more uncommon when a BIER encapsulation is used in data-plane forwarding. The entire IPv6 header, with BIER encapsulation and Routing Header, is expected to look like this:

IPv6 header

Hop-by-Hop Options header [Not Used]

Destination Options header [Not Used]

Routing header [SRH Header with Multicast Address as last SID]

Fragment header [Not Used]

Authentication header [Not Used]

Encapsulating Security Payload header [Not Used]

Destination Options header [BIER header in BIER Option TLV]

Upper-layer header [Data-plane Data]

Once a packet is encapsulated with a BIER Destination Option, it is basically assumed to be a data-plane multicast packet, so the 'OAM' or similar functions in the SRH Header Optional TLV Objects field should not exist.

The last Segment (SID) in the SRH header, or Segment List[0], should be a Multicast Address to indicate a hop-by-hop behavior. Such a Multicast Address can be reserved or unreserved as the Destination Option Header can inform the routers to do the address check. A reserved multicast address should be indicating a 'BIER specific' address.

BIER header has a 'proto' field to identify the type of BIER packet payload, and the IANA has created a registry called "BIER Next

Protocol Identifiers" to assign the value. That means the 'Upper-layer header' of a BIER packet have already been identified by the 'proto' field of the BIER header in the Destination Option Header. Thus the 'Next Header' in the Destination Option Header is not need to identify the 'Upper-layer header' any more, and is recommended to be set to 'No Next Header (value 59)'.

5. BIER Forwarding in Non-MPLS IPv6 Networks

In a Non-MPLS IPv6 Network, BIER may be deployed in a hop-by-hop manner, or possibly be deployed through an SRH tunnel either for "bypassing Non-capable BIER routers" or "fast rerouting". Here is an example where a packet is firstly forwarded through an SRH tunnel and then through a hop-by-hop BIER domain.

When a router along the Segment Routing path receives an IPv6 BIER packet with an SRH header, and if the IPv6 destination address is not one of the router's address, then the packet is forwarded by an IPv6 FIB lookup of the destination address and none of the IPv6 extension headers will be checked. If the IPv6 Destination Address is one of the router's address, and also one of the router's Segment (or SID) of some type, then the router will do a specific function indicated by the Segment, as defined in

[I-D.filsfils-spring-srv6-network-programming]. If the IPv6 Destination Address is a specific type of Segment, called BIER Segment or BIER SID, then the according function is called Endpoint BIER function or 'End.BF' function for short.

When router receives a packet destined to X and X is a local End.BF SID, the router does:

1. IF SL > 0
2. decrement SL
3. update IPv6 DA with SRH[SL]
4. IF SL = 0 & STATE(SRH[0]) = BIER
5. update IPv6 header NH with SRH NH
6. pop the SRH
7. forward the updated packet
8. ELSE
9. drop the packet
10. ELSE
11. drop the packet

Figure 2: End.BF Function

The End.BF function is used for the SRH tunnel destination router to terminate the source-routing SRH forwarding and begin the hop-by-hop BIER IPv6 forwarding. After the SRH header is popped, the multicast

address in the updated IPv6 Destination Address indicates the BIER information of this 'host', and the packet will be forwarded according to the BIER Header in the BIER Destination Option TLV in the IPv6 Destination Option extension header of this 'host'.

In the following hop-by-hop forwarding procedure, the IPv6 Destination Address in an incoming packet indicates the BIER information of this 'host', and the packet will be forwarded according to the BIER Header in the BIER Destination Option TLV in the IPv6 Destination Option extension header. A router is required to ignore the IPv6 BIER Destination Option if the IPv6 Destination Address of a packet is not a multicast address, or is a multicast address without indicating the BIER information of this 'host'.

6. Security Considerations

An IPv6 BIER Destination Option with Multicast Address Destination would be used only when an IPv6 BIER state with the specific Multicast Address Destination has been built by the control-plane. Otherwise the packet with an IPv6 BIER Destination Option will be discarded.

7. IANA Considerations

Allocation is expected from IANA for a Destination Option Type codepoint from the "Destination Options and Hop-by-Hop Options" sub-registry of the "Internet Protocol Version 6 (IPv6) Parameters" registry [RFC2780] at <<https://www.iana.org/assignments/ipv6-parameters/>>.

8. Acknowledgements

TBD.

9. References

9.1. Normative References

[I-D.filsfils-spring-srv6-network-programming]
 Filsfils, C., Li, Z., Leddy, J., daniel.voyer@bell.ca, d., daniel.bernier@bell.ca, d., Steinberg, D., Raszuk, R., Matsushima, S., Lebrun, D., Decraene, B., Peirens, B., Salsano, S., Naik, G., Elmalky, H., Jonnalagadda, P., and M. Sharif, "SRv6 Network Programming", draft-filsfils-spring-srv6-network-programming-04 (work in progress), March 2018.

- [I-D.ietf-6man-segment-routing-header]
Previdi, S., Filsfils, C., Leddy, J., Matsushima, S., and
d. daniel.voyer@bell.ca, "IPv6 Segment Routing Header
(SRH)", draft-ietf-6man-segment-routing-header-13 (work in
progress), May 2018.
- [RFC6564] Krishnan, S., Woodyatt, J., Kline, E., Hoagland, J., and
M. Bhatia, "A Uniform Format for IPv6 Extension Headers",
RFC 6564, DOI 10.17487/RFC6564, April 2012,
<<https://www.rfc-editor.org/info/rfc6564>>.
- [RFC7045] Carpenter, B. and S. Jiang, "Transmission and Processing
of IPv6 Extension Headers", RFC 7045,
DOI 10.17487/RFC7045, December 2013,
<<https://www.rfc-editor.org/info/rfc7045>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6
(IPv6) Specification", STD 86, RFC 8200,
DOI 10.17487/RFC8200, July 2017,
<<https://www.rfc-editor.org/info/rfc8200>>.
- [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>>.
- [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A.,
Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation
for Bit Index Explicit Replication (BIER) in MPLS and Non-
MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January
2018, <<https://www.rfc-editor.org/info/rfc8296>>.

9.2. Informative 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>>.

Authors' Addresses

Jingrong Xie
Huawei Technologies

Email: xiejingrong@huawei.com

Liang Geng
China Mobile
Beijing 10053

Email: gengliang@chinamobile.com

Lei Wang
China Mobile
Beijing 10053

Email: wangleiyjy@chinamobile.com

Gang Yan
Huawei

Email: yangang@huawei.com

Mike McBride
Huawei

Email: mmcbride7@gmail.com

Yang Xia
Huawei

Email: yolanda.xia@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

J. Xie
Huawei Technologies
X. Xu
Alibaba Inc.
G. Yan
M. McBride
Huawei Technologies
July 2, 2018

Use of BIER Entropy for Data Center CLOS Networks
draft-xie-bier-entropy-staged-dc-clos-00

Abstract

Bit Index Explicit Replication (BIER) introduces a new multicast-specific BIER Header. BIER can be applied to the Multi Protocol Label Switching (MPLS) data plane or Non-MPLS data plane. Entropy is a technique used in BIER to support load-balancing. This document examines and describes how BIER Entropy is to be applied to Data Center CLOS networks for path selection.

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

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 <https://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 January 3, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem Statement and Considerations	3
3.1. Problem Statement	3
3.2. Considerations	4
4. Use of BIER Entropy for DC CLOS Network	5
4.1. Use of BIER Entropy for DC CLOS Network	5
4.2. Steering for elephant flows	6
4.3. Path Division for Tenant flows to different SIs	6
4.4. Link Failure and Convergence	6
5. Data-Plane Processing	7
6. Security Considerations	7
7. IANA Considerations	7
8. Acknowledgements	7
9. References	7
9.1. Normative References	7
9.2. Informative References	8
Authors' Addresses	8

1. Introduction

Bit Index Explicit Replication (BIER) [RFC8279] is an architecture that provides optimal multicast forwarding without requiring intermediate routers to maintain any per-flow state by using a multicast-specific BIER header. [RFC8296] defines two types of BIER encapsulation formats: one is MPLS encapsulation, the other is non-MPLS encapsulation. Entropy is a technique used in BIER to support load-balancing. This document examines and describes how BIER Entropy is to be applied to Data Center CLOS networks for path selection.

2. Terminology

Readers of this document are assumed to be familiar with the terminology and concepts of the documents listed as Normative References.

3. Problem Statement and Considerations

3.1. Problem Statement

A common choice for a horizontally scalable topology used in Data Center is a Clos topology. This topology features an odd number of stages, for example, a 5-Stage Clos Topology as an example in [RFC7938].

ECMP is the fundamental load-sharing mechanism used by a Clos topology. Effectively, every lower-tier device will use all of its directly attached upper-tier devices to load-share traffic destined to the same IP prefix. The number of ECMP paths between any two Tier 3 devices in Clos topology is equal to the number of the devices in the middle stage (Tier 1). For example, Figure 1 illustrates a topology where Tier 3 device L1 has four paths to reach servers X and Y, via Tier 2 devices S1 and S2 and then Tier 1 devices S11, S12, S21, and S22, respectively.

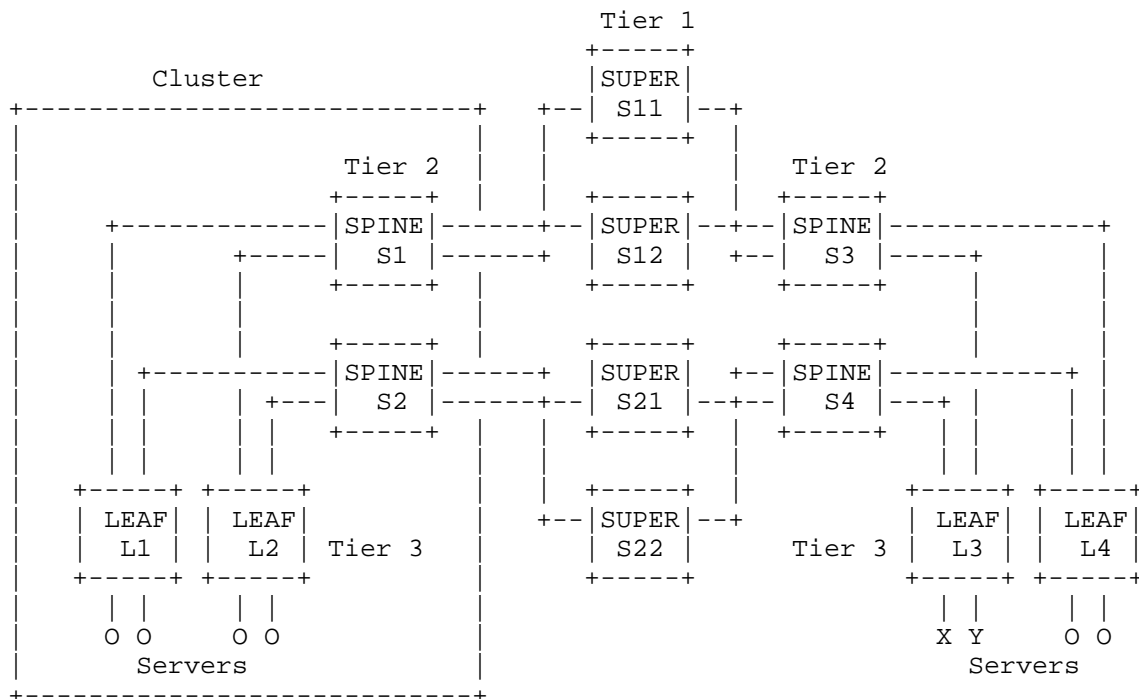


Figure 1: 5-Stage Clos Topology

When BIER is deployed in a multi-tenant data center network environment for efficient delivery of Broadcast, Unknown-unicast and Multicast (BUM) traffic, a network operator may want a deterministic path for every packet. For example, when L1 needs to send a BUM packet to L3 and L4, which are in different SIs, L1 has to send the packet twice, and expects the packet along two deterministic paths of L1->S1->S11->L3 and L1->S2->S21->L4 separately. Another example of using a deterministic path in a DC is for per-flow steering of "elephant" flows defined in [I-D.ietf-spring-segment-routing-msdc].

A deterministic path for a multicast path, with multiple staged equal cost paths, is comparable to a traffic-engineering path defined in [I-D.ietf-mpis-spring-entropy-label] for a unicast path with multiple hop equal cost paths.

3.2. Considerations

The idea behind entropy is that the ingress router computes a hash based on several fields from a given packet and places the result in an additional label, named "entropy label". Then this entropy label can be used as part of the hash keys used by a transit router. When

entropy label is used, the keys used in the hashing functions are still a local configuration matter. A router may solely use the entropy label or use a combination of multiple fields from the incoming packet. The hashing function is to randomly load balance the mass of flows between the small number of equal cost paths.

If one wants, however, to get a deterministic path from the equal cost paths, one can use part of the 20-bit entropy field. For example, bit 0 to bit 2 of entropy label can represent a value of 0 to 7, and thus can be used to select a deterministic path from 8 equal cost paths. And thus, a 20-bit entropy label can be used by routers in different tiers to select a deterministic path independently by using different parts of the 20-bit entropy label, and form an end-to-end deterministic path.

This is simple and applicable especially for DC CLOS networks, because data delivery in DC CLOS networks for tenants is always multi-staged, with the upstream direction stages having equal cost paths.

4. Use of BIER Entropy for DC CLOS Network

4.1. Use of BIER Entropy for DC CLOS Network

Take the 5-stage CLOS network in figure 1 as an example.

Tier 2 in every cluster has N nodes, and the Tier 1 has M nodes. M is equal to N multiplied by P.

Tier 3 switches, in upstream direction, act as stage 1 of data delivery and have N equal cost paths to every BFERs in other clusters. Tier 2 switches, in upstream direction, act as stage 2 of data delivery and have P equal cost paths to every BFERs in other clusters.

Example 1: One can configure, on each Tier 3 switch, the use of bit 0 for path selection when N is equal to 2, and configure, on each Tier 2 switch, to use bit 1 for path selection when P is equal to 2.

Example 2: One can configure, on each Tier 3 switch, the use of bit 0 to bit 1 for path selection when N is equal to 4, and configure on each Tier 2 switches the use of bit 2 to bit 7 for path selection when P is equal to 48.

Assume that, each Tier 3 and Tier 2 switch the the example have two parameters, X and Y, for using part of entropy label to do path selection, then in example 2:

- o Each of Tier 3 (Stage 1) switches has a pair of parameters ($X1=1$, $Y1=4$)
- o Each of Tier 2 (Stage 2) switches has a pair of parameters ($X2=X1*Y1=4$, $Y2=64$)
- o Each of Tier 3 (Stage 1) switches populates its BIFTs for ECMP, for example, BIFT-0 to BIFT-3.
- o Each of Tier 2 (Stage 2) switches populates its BIFTs for ECMP, for example, BIFT-0 to BIFT-47.

For each of Tier 3 (Stage 1) switches, each of the BIFT will have a preferred neighboring BFR. For example, LEAF L1 will have a preferred neighbor S1/S2 for BIFT-0/1 separately, and when forming the BIFT-0 table through the underlay routing to every BFER, the preferred neighboring BFR will have a highest priority among all the locally available ECMP path.

Then an end-to-end deterministic path for a BIER packet can be had by calculating an entropy label value like this:

$$\text{Entropy} = (P1-1)*X1 + (P2-1)*X2$$

Where P1 represents one of the Stage 1 equal cost paths with a value between 1 and N, and P2 represents one of the Stage 2 equal cost paths with a value between 1 and P.

4.2. Steering for elephant flows

One can steer an "elephant" flow to an end-to-end deterministic path, or some divided end-to-end deterministic paths across different SIs.

4.3. Path Division for Tenant flows to different SIs

When the VNEs for a tenant span multiple SIs, then it is useful to divide the BUM packets paths across different SIs.

One can configure a policy to use different paths for BIER SIs when using BIER as the BUM tunnel, on each VNE for each VNI.

4.4. Link Failure and Convergence

As stated above, each of the BIFT on a BFR will have a preferred neighboring BFR. But when the link to the preferred neighbor of some BIFT (say BIFT-X) fail, BIFT-X will converge normally, and will then probably not being the 'best' path. For example, the link between S1 and L2 fail, then the preferred neighbor of BIFT-0 of LEAF L1, S1, is

no longer the neighboring BFR for LEAF L2, and the flow using a Entropy using LEAF L1's BIFT-0 will have to replicate on L1, one packet to S1 for BFER L3 and L4, and one packet to S2 for BFER L2. If the flow changes to use a Entropy using LEAF L1's BIFT-1, it will then be the 'best' path, because the flow doesn't have to replicate on L1, only one to S1 for BFER L2 and L3 and L4. Such a change to a flow's entropy is the Ingress switch's responsibility, possibly with the assistance of a controller.

5. Data-Plane Processing

The use of BIER entropy label to select a path between some equal cost paths is a local configuration matter. This draft defines a method to use part of the 20-bit entropy label in each router, and this needs a data-plane to do some bit operation function. It is expected to be easier than hashing function.

6. Security Considerations

This document introduces no new security considerations beyond those already specified in [RFC8279] and [RFC8296].

7. IANA Considerations

This document contains no actions for IANA.

8. Acknowledgements

TBD.

9. References

9.1. Normative References

- [I-D.ietf-mpls-spring-entropy-label]
Kini, S., Kompella, K., Sivabalan, S., Litkowski, S., Shakir, R., and J. Tantsura, "Entropy label for SPRING tunnels", draft-ietf-mpls-spring-entropy-label-11 (work in progress), May 2018.
- [I-D.ietf-spring-segment-routing-msdc]
Filsfils, C., Previdi, S., Dawra, G., Aries, E., and P. Lapukhov, "BGP-Prefix Segment in large-scale data centers", draft-ietf-spring-segment-routing-msdc-09 (work in progress), May 2018.

- [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, <<https://www.rfc-editor.org/info/rfc7938>>.
- [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>>.
- [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <<https://www.rfc-editor.org/info/rfc8296>>.
- [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.

9.2. Informative 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>>.

Authors' Addresses

Jingrong Xie
Huawei Technologies

Email: xiejingrong@huawei.com

Xiaohu Xu
Alibaba Inc.

Email: xiaohu.xxh@alibaba-inc.com

Gang Yan
Huawei Technologies

Email: yangang@huawei.com

Mike McBride
Huawei Technologies

Email: mmcbride7@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

J. Xie
M. McBride
M. Chen
Huawei Technologies
L. Geng
China Mobile
July 2, 2018

Multicast VPN Using MPLS P2MP and BIER
draft-xie-bier-mvpn-mpls-p2mp-02

Abstract

MVPN is a widely deployed multicast service with mLDP or RSVP-TE P2MP as the P-tunnel. Bit Index Explicit Replication (BIER) is an architecture that provides optimal multicast forwarding without requiring intermediate routers to maintain any per-flow state by using a multicast-specific BIER header. This document introduces a seamless transition mechanism from legacy MVPN using mLDP/RSVP-TE P2MP to MVPN using BIER by combining P2MP and BIER to form a P2MP based BIER as the P-tunnel. This will leverage the widely supported P2MP capability in both data-plane and control-plane, and will help introducing BIER in existing multicast networks to shift multicast delivery from MVPN using mLDP/RSVP-TE P2MP by two means: It is easier and more efficient for legacy routers to support BIER forwarding on the basis of widely supported P2MP forwarding, and it is more seamless for existing multicast networks to deploy BIER when some routers do not support BIER forwarding.

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

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 <https://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 January 3, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Terminology	3
3. Applicability Statement	4
4. MVPN using P2MP based BIER	5
4.1. Overview	5
4.2. MVPN Transition from P2MP to P2MP based BIER	5
4.2.1. Use of the PTA in x-PMSI A-D Routes	6
4.3. Building P2MP based BIER forwarding state	8
5. P2MP based BIER Forwarding Procedures	8
5.1. Overview	8
5.2. P2MP based BIER forwarding	9
5.3. When Mid, Leaf or Bud nodes do not support P-CAPABILITY	11
5.4. When Leaf or Bud nodes do not support D-CAPABILITY	13
6. Provisioning Considerations	15
7. IANA Considerations	16
8. Security Considerations	16
9. Acknowledgements	16
10. References	17
10.1. Normative References	17
10.2. Informative References	18
Authors' Addresses	18

1. Introduction

[RFC6513] and [RFC6514] specify the protocols and procedures that a Service Provider (SP) can use to provide Multicast Virtual Private Network (MVPN) service to its customers. Multicast tunnels are created through an SP's backbone network; these are known as "P-tunnels". The P-tunnels are used for carrying multicast traffic across the backbone. The MVPN specifications allow the use of several different kinds of P-tunnel technology, such as mLDP P2MP and RSVP-TE P2MP. It is common for such a P-tunnel having a multicast-specific path.

Bit Index Explicit Replication (BIER) [RFC8279] is an architecture that provides optimal multicast forwarding through a "multicast domain", without requiring intermediate routers to maintain any per-flow state, by using a multicast-specific BIER header (per [RFC8296]).

[I-D.ietf-bier-mvpn] delivers a solution of MVPN using SPF based BIER defined in [RFC8279]. It can not, however, support a multicast-specific path well, something common in legacy MVPN deployment.

[RFC8279] provides a solution to support mid nodes without BIER-capability. It cannot, however, support deployment on a network that has edge nodes without BIER-capability, which may be common in some SP-networks, especially when most of the nodes in a network or part of a network are edge or service nodes.

This document introduces a seamless transition mechanism from legacy MVPN to MVPN using P2MP based BIER, by applying a BIER encapsulation in data-plane to eliminate per-flow states, while preserving existing features such as multicast-specific PATH.

It also introduces a seamless deployment solution on networks with Non-BIER-capability Edge nodes and/or Mid nodes, by exploring the P2MP/tree based BIER forwarding procedure in detail. Such a P2MP/tree based BIER is mentioned but not explored in detail in RFC8279.

2. Terminology

Readers of this document are assumed to be familiar with the terminology and concepts of the documents listed as Normative References. For convenience, some of the more frequently used terms and new terms list below.

- o LSP: Label Switch Path
- o LSR: Label Switching Router

- o P2MP: Point to Multi-point
- o P-tunnel: A multicast tunnel through the network of one or more SPs. P-tunnels are used to transport MVPN multicast data.
- o PMSI: Provider Multicast Service Interface
- o x-PMSI A-D route: a route that is either an I-PMSI A-D route or an S-PMSI A-D route.
- o PTA: PMSI Tunnel attribute. A type of BGP attribute known as the PMSI Tunnel attribute.
- o P2MP based BIER: BIER using P2MP LSP as topology
- o P-CAPABILITY: A capability to Process BitString in BIER Header of a packet.
- o D-CAPABILITY: A capability to Disposit BIER Header of a packet, including or excluding the BIER Label.
- o BSL: Bit String Length, that is 64, 128, 256, etc (per [RFC8279]).

3. Applicability Statement

The BIER architecture document [RFC8279] describes how each node forwards BIER packets hop by hop to neighboring nodes without generating duplicate packets. This forwarding is for the case where a form of underlay called "many to many " and built by IGP is used. Obviously, the case of underlay of "one to many" or P2MP is a simpler scenario, and the forwarding procedure naturally applies. However, as is well-known, such a forwarding procedure requires the support of hardware. The usage of the same forwarding method for both complex scenarios and simple scenarios will inevitably require complex hardware forwarding.

This document describes how BIER forwarding can be customized and simplified with an underlay of "one to many" or P2MP (see chapter 5). This customization and simplification eliminates some of the unnecessary data plane processing and so is easier to implement with existing hardware. Based on this customization of the forwarding method for P2MP-based BIER, a variety of Partial Deployment methods are given for the different capabilities of the hardware to support BIER forwarding. Compared with RFC8279, when there is no BIER forwarding capability on edge nodes, Partial Deployment can be carried out ; For the case where the intermediate node has no BIER forwarding capability, P2MP forwarding can be used without the need for unicast replication.

This document also describes a MVPN Transition solution that eliminates the per-flow state by introducing BIER MPLS encapsulation and forwarding in data-plane, while preserving the original control-plane protocol and its features, especially when some sort of path customizing being used. The said path customization include RSVP-TE P2MP using an explicit path, and MLDP P2MP where static route was used. These features can continue to retain, making the transition process seamless.

4. MVPN using P2MP based BIER

4.1. Overview

According to [RFC8279], the P2MP based BIER is a BIER which using a form of tree as the underlay. The P2MP LSP is not only a LSP, but also a topology as the BIER underlay. The P2MP based BIER is P-tunnel, which is used for bearing multicast flows. Every flow can be seen as binding to an independent tunnel, which is constructed by the BitString in the BIER header of every packet of the flow. Multicast flows are transported in SPMSI-only mode, on P2MP based BIER tunnels, and never directly on P2MP LSP tunnel.

Section 4.2 describes the overall principle of transitioning a Legacy MVPN using P2MP to a MVPN using BIER. It also describes the detail use of new types of PTA in BGP MVPN routes to indicate PEs to initialize the building of P2MP based BIER forwarding.

Section 4.3 describes the Underlay protocols to build P2MP based BIER forwarding briefly.

4.2. MVPN Transition from P2MP to P2MP based BIER

This section describes a MVPN transitioning solution that eliminates the per-flow state by introducing BIER MPLS encapsulation and forwarding procedure in data-plane, while preserving the originally deployed control-plane protocol and its features, especially when some sort of path customizing being used.

When transitioning a MVPN using mLDP P2MP P-tunnel, then continue using mLDP to build a P2MP based BIER forwarding, preserving the original mLDP features. For example, mLDP uses static route to specify a path other than the path of IGP.

When transitioning a MVPN using RSVP-TE P2MP P-tunnel, then continue using RSVP-TE to build a P2MP based BIER forwarding, preserving the original RSVP-TE features. For example, RSVP-TE use explicit path to specify a path other than the path of IGP.

4.2.1. Use of the PTA in x-PMSI A-D Routes

As defined in [RFC6514], the PMSI Tunnel attribute (PTA) carried by an x-PMSI A-D route identifies the P-tunnel that is used to instantiate a particular PMSI. If a PMSI is to be instantiated by P2MP LSP based BIER, the PTA is constructed by a BFIR, which is also an Ingress LSR. This document defines the following Tunnel Types:

+ TBD - RSVP-TE built P2MP BIER

+ TBD - mLDP built P2MP BIER

Allocation is expected from IANA for two new tunnel type codepoints from the "P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types" registry. These codepoints will be used to indicate that the PMSI is instantiated by MLDP or RSVP-TE extension with support of BIER.

When the Tunnel Type is set to RSVP-TE built P2MP BIER, the Tunnel Identifier includes two parts, as follows:

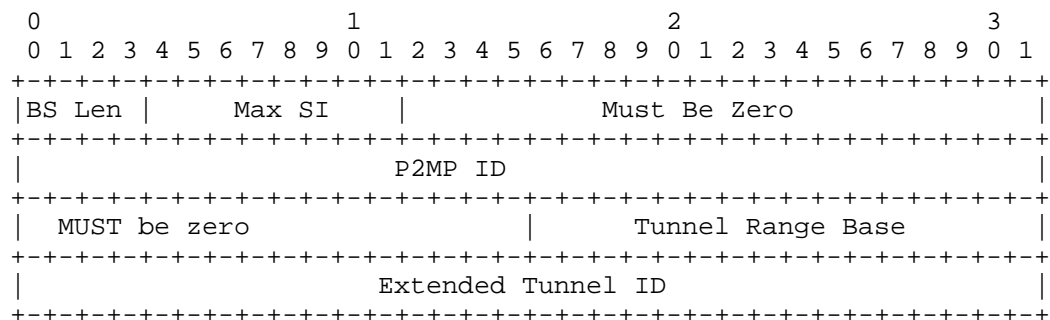


Figure 1: PTA of RSVP-TE built P2MP BIER

BS Len: A 4 bits field. The values allowed in this field are specified in section 2 of [RFC8296].

Max SI: A 1 octet field. Maximum Set Identifier (section 1 of [RFC8279]) used in the encapsulation for this BIER sub-domain.

<Extended Tunnel ID, Reserved, Tunnel Range Base, P2MP ID>: A ID as carried in the RSVP-TE P2MP LSP SESSION Object defined in [RFC4875].

The "Tunnel Range" is the set of P2MP LSPs beginning with the Tunnel Range base and ending with ((Tunnel Range base)+(Tunnel Number)- 1). A unique Tunnel Range is allocated for the BSL and a Sub-domain-ID implicated by the P2MP.

The size of the Tunnel Range is determined by the number of Set Identifiers (SI) (section 1 of [RFC8279]) that are used in the topology of the P2MP-LSP. Each SI maps to a single Tunnel in the Tunnel Range. The first Tunnel is for SI=0, the second Tunnel is for SI=1, etc.

When the Tunnel Type is set to mLDP built P2MP BIER, the Tunnel Identifier include two parts, as follows:

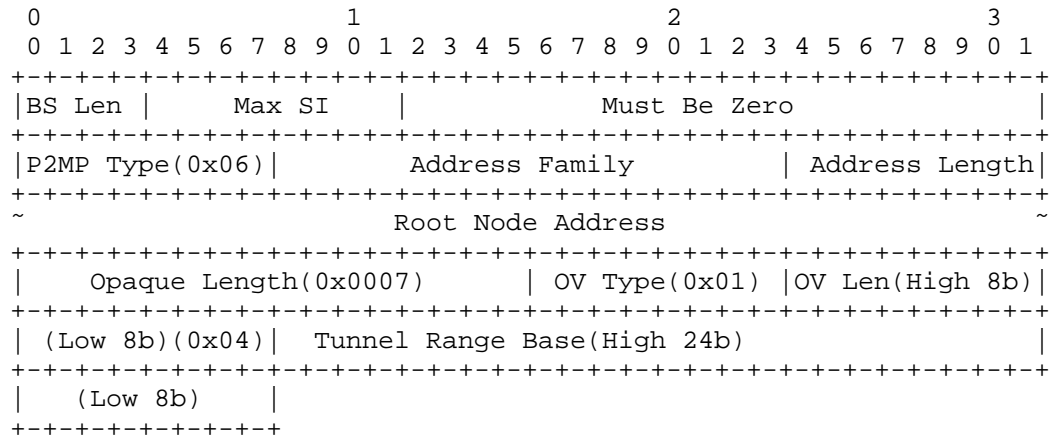


Figure 2: PTA of MLDP built P2MP BIER

BS Len: A 4 bits field. The values allowed in this field are specified in section 2 of [RFC8296].

Max SI: A 1 octet field. Maximum Set Identifier (section 1 of [RFC8279]) used in the encapsulation for this BIER sub-domain.

<Type=0x06, AF, AL, RootNodeAddr, Opqgue Length=0x0007, OV Type=0x01, OV Len=0x04, Tunnel Range Base>: A P2MP Forwarding Equivalence Class (FEC) Element, with a Generic LSP Identifier TLV as the opaque value element, defined in [RFC6388].

The "Tunnel Range" is the set of P2MP LSPs beginning with the Tunnel Range base and ending with ((Tunnel Range base)+(Tunnel Number)- 1). A unique Tunnel Range is allocated for the BSL and a Sub-domain-ID implicated by the P2MP.

The size of the Tunnel Range is determined by the number of Set Identifiers (SI) (section 1 of [RFC8279]) that are used in the topology of the P2MP-LSP. Each SI maps to a single Tunnel in the Tunnel Range. The first Tunnel is for SI=0, the second Tunnel is for SI=1, etc.

When the Tunnel Type is any of the above, The "MPLS label" field contain an upstream-assigned non-zero MPLS label. It is assigned by the router (a BFIR) that constructs the PTA. Absence of an MPLS Label is indicated by setting the MPLS Label field to zero.

When the Tunnel Type is any of the above, two of the flags, LIR and LIR-pF, in the PTA "Flags" field are meaningful. Details about the use of these flags can be found in [RFC6513], [I-D.ietf-bess-mvpn-expl-track] and [I-D.ietf-bier-mvpn]].

4.3. Building P2MP based BIER forwarding state

When P2MP based BIER are used, then it is not necessary to use IGP or BGP to build the BIER routing table and forwarding table. Instead, the BIER layer information is carried by MLDP or RSVP-TE, when they build the P2MP tree.

The detail procedure for building P2MP based BIER forwarding state using mLDLP or RSVP-TE is outside the scope of this document.

5. P2MP based BIER Forwarding Procedures

5.1. Overview

This document specifies one OPTIONAL Forwarding Procedure of BIER encapsulation packet, on the condition that the BIER underlay topology is P2MP LSP, as describes in the above sections. It is in fact a customized forwarding procedure, and a detail exploration of BIER forwarding along a multicast-specific tree. Comparing to the common Forwarding Procedure described in [RFC8279], there is some considerable simplification:

1. Not need to Edit the BitString when forwarding packet to Neighbor, for the underlay P2MP topology is already loop-free and duplicate-free. This can further lead to a method to by-pass the BIER encapsulation packet when a node does not support the BitString process.
2. Not need to do a disposition function by parsing the BitString, for a P2MP can identify a disposition function by a node's Label when the P2MP is built. This can further reduce the complex BitString processing for legacy hardware on edge, and lead to a method to deploy on exist network when an edge node does not support BitString process.

The main principle of the optional forwarding procedure of the P2MP based BIER is, on the basis of P2MP forwarding procedure according to the BIER-MPLS label, to use the BitString to prune/filter the

undesired P2MP downstream. This is a smooth enhancement to the widely deployed P2MP forwarding, and easier to deploy on existing routers comparing to the many-to-many BIER forwarding.

The enhancement to the P2MP forwarding is to add a Forwarding BitMask to existing NHLFE defined in [RFC3031], for checking with the BitString in a packet, to determine whether the packet is to be forwarded or pruned. If the checking result by AND'ing a packet's BitString with the F-BM of the NHLFE (i.e., Packet->BitString &= F-BM) is non-zero, then forward the packet to the next-hop indicated by the NHLFE entry, and the Label is switched to the proper one in the NHLFE. If the result is zero, then do not forward the packet to the next-hop indicated by the NHLFE entry.

5.2. P2MP based BIER forwarding

For a P2MP tree, every node has a role of Root, Branch, Leaf, or Bud, as specified in [RFC4611].

EXAMPLE 1: Take the following figure as an example.

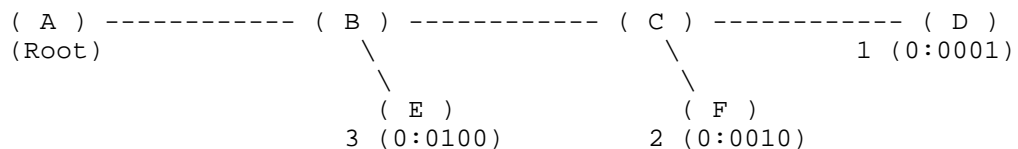


Figure 3: P2MP-based BIER Topology without BUD nodes

Forwarding Table on A:

- o NHLFE(TreeID, OutInterface<toB>, OutLabel<alloc by B>, F-BM<0111>)

Forwarding Table on C:

- o ILM(inLabel<alloc by C>, action<TreeID>, Flag=Branch|CheckBS, BSL)
- o NHLFE(TreeID, OutInterface<toD>, OutLabel<alloc by D>, F-BM<0001>)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>, F-BM<0010>)

For Node C, the ability to receive a MPLS-encapsulation BIER packet, match ILM and get a TreeID, replicate to NHLFE Entries of the TreeID according to the result of AND'ing the BitString of packet and the F-BM of a NHLFE Entry, is called a P-CAPABILITY, which means to Process BitString in each packet.

Forwarding Table on B is the same to C.

Forwarding Table on D:

- o ILM(inLabel<alloc by D>, action<TreeID>, Flag=Leaf|CheckBS, BSL)
- o LEAF(TreeID, F-BM<0001>, flag=PopBIERincluding)

When Node D receive a MPLS-encapsulation BIER packet, it get the Label and match ILM, then do a replication according to the LEAF and check whether to proceed by AND'ing the BitString in the replicated packet and the F-BM in the LEAF entry. When the AND'ing result is non-zero then do a POP to the packet to disposit the whole BIER header Including the BIER Label, which has a length of (12+BSL/8) octets.

Node D need to have a P-CAPABILITY, for it need to Process BitString in each packet to determin whether to replicate to a special LEAF, and then disposit the whole BIER header Including the BIER Label and forward the IP multicast packet further. Node D also need to do the disposition as well, which is called a D-CAPABILITY. D-CAPABILITY means to disposit the BIER header including or excluding the BIER Label in the begining. Here PopBIERincluding means pop the BIER header including the BIER Label, while PopBIERexcluding means pop the BIER header excluding the BIER Label.

Forwarding Tables on E and F are same to D.

Comparing to the forwarding procedure defined in [RFC8279], there are two benefits of using the customized P2MP based BIER forwarding:

1. Not need to walk every physical neighbor, but only need to walk downstream neighbors on a P2MP tree.
2. Not need to edit the BitString in every packet, but only need to swap the BIER Label.

EXAMPLE 2: Another example with P2MP BUD Nodes.

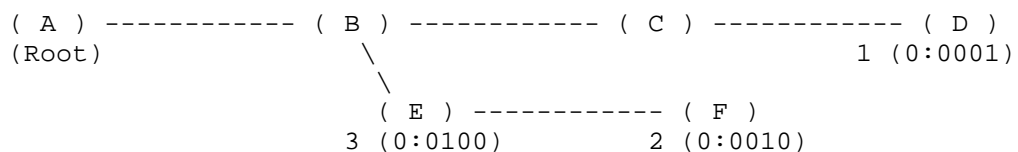


Figure 4: P2MP-based BIER Topology with BUD nodes

Forwarding Table on B (Branch Node):

- o ILM(inLabel<alloc by B>, action<TreeID>, Flag=Branch|CheckBS, BSL)

- o NHLFE(TreeID, OutInterface<toE>, OutLabel<alloc by E>, F-BM<0110>)
- o NHLFE(TreeID, OutInterface<toC>, OutLabel<alloc by C>, F-BM<0001>)

Node B, which is a Branch Node, only need to use its P-CAPABILITY.

Forwarding Table on E (BUD Node):

- o ILM(inLabel<alloc by E>, action<TreeID>, Flag=Bud|CheckBS, BSL)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>, F-BM<0010>)
- o LEAF(TreeID, F-BM<0100>, flag=PopBIERincluding)

When Node E receive a MPLS-encapsulation BIER packet, it get the Label and match ILM, then do a replication according to the NHLFEs and check whether to proceed by AND'ing the BitString in the replicated packet and the F-BM in the NHLFE/LEAF entry. When the AND'ing result is non-zero for the second LEAF then do a POP to the packet to disposit the whole BIER header, which has a length of (12+BSL/8) octets.

Node E, which is a BUD Node, has both the two capacities: P-CAPABILITY and D-CAPABILITY. P-CAPABILITY is need to be used for every NHLFE/LEAF, and D-CAPABILITY is need for the NHLFE that has a PopBIERincluding flag.

5.3. When Mid, Leaf or Bud nodes do not support P-CAPABILITY

The procedures of Section 5.2 presuppose that, within a given BIER domain, all the nodes adjacent to a given BFR in a given routing underlay are also BFRs. However, it is possible to use BIER even when this is not the case. In this section, we describe procedures that can be used if the routing underlay is a P2MP tree with BIER information in the domain.

For a P2MP tree, every node has a role of Root, Branch, Leaf, or Bud. The role is determined when the tree is built. The method is suitable for conditions when Mid, Leaf or Bud nodes do not support P-CAPABILITY.

EXAMPLE 1: Take Figure 4 as an example.

If D, F, E support BIER, and C don't support BIER, then we can configure on C to indicate it to use P2MP for BIER packets forwarding. Then C build a P2MP forwarding entry, while still pass the BIER information in control-plane. For example, D send a P2MP FEC Mapping message to C with a BitMask 0001, F send a P2MP FEC

Mapping message to C with a BitMask 0010, and C send a P2MP FEC Mapping message to B with a BitMask, but C build a P2MP forward entry like this:

- o ILM(inLabel<alloc by C>, action<TreeID>, Flag=Branch)
- o NHLFE(TreeID, OutInterface<toD>, OutLabel<alloc by D>)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>)

If D don't support BIER P-CAPABILITY, but it support BIER D-CAPABILITY, then the above method is still valid.

Forwarding Table on D when D don't have a P-CAPABILITY:

- o ILM(inLabel<alloc by D>, action<TreeID>, Flag=Leaf, BSL)
- o NHLFE(TreeID, flag=PopBIERincluding)

When Node D receive a MPLS-encapsulation BIER packet, it get the Label and match ILM, then do a replication according to the NHLFE but don't do the check by AND'ing the BitString in the replicated packet and the F-BM in the NHLFE entry. And then do a POP to the packet to disposit the whole BIER header, which has a length of (12+BSL/8) octets.

Another alternative form of Forwarding Table on D can also be the following when D don't have a P-CAPABILITY:

- o ILM(inLabel<alloc by D>, action<PopBIERincluding>, Flag=Leaf, BSL)

When Node D receive a MPLS-encapsulation BIER packet, it get the Label and match ILM, then do a POP action according to the ILM to pop the whole (12+BSL/8) octets from the Label position.

EXAMPLE 2: Take BUD Node E in Figure 5 as another example.

Forwarding Table on Bud Node E when E don't have a P-CAPABILITY:

Forwarding Table on E when E don't have a P-CAPABILITY:

- o ILM(inLabel<alloc by E>, action<TreeID>, Flag=Bud, BSL)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>)
- o LEAF(TreeID, flag=PopBIERincluding)

One can see that, this method can support widely Non-BIER Nodes in a network, no matter the node has a Mid, Leaf or Bud role, and would never result in any ingress-replication through unicast tunnel, which may cause a overload on a link.

One can also see that, [RFC8279] only support Non BIER-capability nodes being the Mid nodes, and never allow a BFER nodes to be Non BIER-capability.

5.4. When Leaf or Bud nodes do not support D-CAPABILITY

A more tolerant variant of the above, when Leaf or Bud nodes do not support D-CAPABILITY, would be the following:

EXAMPLE 1: Take Figure 4 as an example.

If D even don't support BIER P-CAPABILITY or D-CAPABILITY, then POP the whole BIER Header except the first four octets Label field of a packet before it come to D. This requires C to build a forwarding table like this:

Forwarding Table on C (Branch Node):

- o ILM(inLabel<alloc by E>, action<TreeID>, Flag=Branch|CheckBS, BSL)
- o NHLFE(TreeID, OutInterface<toD>, OutLabel<alloc by D>, F-BM<0001>, Flag=PopBIERexcluding)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>, F-BM<0010>)

The Flag PopBIERexcluding means POP the BIER Header excluding the first 4 octets BIER Label in a packet, that is a Length of (8+BSL/8)

If D don't support BIER P-CAPABILITY or D-CAPABILITY, and C don't support BIER P-CAPABILITY, then it requires B to build a forwarding table, to ensure the BIER Header except the first four octets Label field of a packet is popped before replicated to C, and requires C to build a forwarding table of a pure P2MP branch, and requires F to build a forwarding table of a pure P2MP Leaf. Their forwarding tables are like below:

Forwarding Table on B (Branch Node):

- o ILM(inLabel<alloc by B>, action<TreeID>, Flag=Branch|CheckBS, BSL)
- o NHLFE(TreeID, OutInterface<toC>, OutLabel<alloc by C>, F-MB<0011>, Flag=PopBIERexcluding)

- o NHLFE(TreeID, OutInterface<toE>, OutLabel<alloc by E>, F-BM<0100>)

Forwarding Table on C (Branch Node):

- o ILM(inLabel<alloc by C>, action<TreeID>, Flag=Branch)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>)

Forwarding Table on D (Branch Node):

- o ILM(inLabel<alloc by D>, action<PopLabel>, Flag=Leaf)

Here PopLabel mean to pop the Label, which is in fact a P2MP LSP Label. It is a basic capability of any LSR.

Forwarding Table on F (Branch Node):

- o ILM(inLabel<alloc by F>, action<PopLabel>, Flag=Leaf)

Here PopLabel mean to pop the Label, which is in fact a P2MP LSP Label. It is a basic capability of any LSR, and the Forwarding table on F is in fact a P2MP one.

Note that, although F support BIER, which means it can deal with a BIER packet, but it must downshift its forwarding table to a pure P2MP one, because the packet it received doesn't include a BIER Header but a P2MP Label packet due to the POP behaving of its upstream node.

EXAMPLE 2: Take Figure 5 as another example.

If E even don't support BIER P-CAPABILITY or D-CAPABILITY, then POP the whole BIER Header Except the first four octets Label field of a packet before it come to D. This requires B to build a forwarding table like this:

Forwarding Table on B (Branch Node):

- o ILM(inLabel<alloc by B>, action<TreeID>, Flag=Branch|CheckBS, BSL)
- o NHLFE(TreeID, OutInterface<toC>, OutLabel<alloc by C>, F-MB<0011>)
- o NHLFE(TreeID, OutInterface<toE>, OutLabel<alloc by E>, F-BM<0100>, Flag=PopBIERexcluding)

Forwarding Table on E (Bud Node):

- o ILM(inLabel<alloc by E>, action<TreeID>, Flag=Bud)
- o NHLFE(TreeID, OutInterface<toF>, OutLabel<alloc by F>)
- o LEAF(TreeID, flag=PopLabel)

Forwarding Table on F (Branch Node):

- o ILM(inLabel<alloc by F>, action<PopLabel>, Flag=Leaf)

Note that, although F support BIER, which means it can deal with a BIER packet, but it must downshift its forwarding table to a pure P2MP Leaf, because the packet it received doesn't include a BIER Header but a P2MP Label packet due to the POP behaving of its upstream node.

One can see that, when some Leaf or Bud nodes even don't have a D-CAPABILITY, we can do a POP action to disposing the BIER header excluding the BIER Label in the begining before the packet arrive the node. This is similar to a Penultimate Hop Popping in a P2P LSP.

6. Provisioning Considerations

P2MP based BIER use concepts of both P2MP and BIER. Some provisioning considerations list below:

Sub-domain:

In P2MP based BIER, every P2MP is a specific BIER underlay topology, and an implicit Sub-domain. RSVP-TE/MLDP build the BIER information of the implicit sub-domain when building the P2MP tree. MVPN get the implicit sub-domain by provisioning.

BFR-prefix:

In P2MP LSP based BIER, every BFR is also a LSR. So the BFR-prefix in the sub-domain is by default identified by LSR-id. Additionally, When BFR/LSR is also a MVPN PE, BFR-prefix is also the same as Originating Router's IP Address of x-PMSI A-D route or Leaf A-D route.

BFR-id:

When using protocols like RSVP-TE, which initializes P2MP LSP from a specific Ingress Node, BFR-id which is unique in P2MP LSP scope, can be auto-provisioned by Ingress Node, or conventionally configure on every Egress Nodes.

BSL and BIER-MPLS Label Block Size:

In P2MP LSP based BIER, Every P2MP LSP or implicit sub-domain requires a single BSL, and a specific BIER-MPLS Label block size for this BSL.

VPN-Label:

The P2MP based BIER 'P-tunnel' can be shared by multiple VPNs or a single VPN. When a P2MP based BIER being shared by multiple VPNs, an Upstream-assigned VPN-Label is required. It can be auto-provisioned or manual configured by the BFIR or Ingress LSR.

In fact, [RFC6513] has defined the method of "Aggregating Multiple MVPNs on a Single P-Tunnel". But unfortunately it is not widely deployed because of the serious trade-off between state saving and bandwidth waste. The BIER encapsulation and forwarding method give it a chance to eliminate the trade-off while gaining a completely state saving.

Even when such an aggregating is not used, it is still adequate to use BIER to save state by sharing one P2MP based BIER "P-tunnel" for multi flows in one specific VPN.

For seamless transitioning from legacy MVPN deployment and existing network, it is recommended not to use such an aggregating, as well as to use such an aggregating.

7. IANA Considerations

Allocation is expected from IANA for two new tunnel type codepoints for "RSVP-TE built P2MP based BIER" and "MLDP built P2MP based BIER" from the "P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types" registry.

8. Security Considerations

This document does not introduce any new security considerations other than already discussed in [RFC8279].

9. Acknowledgements

The authors would like to thank Eric Rosen, Tony Przygienda, IJsbrand Wijnands and Toerless Eckert for their thoughtful comments and kind suggestions.

10. References

10.1. Normative References

- [I-D.ietf-bess-mvpn-expl-track]
Dolganow, A., Kotalwar, J., Rosen, E., and Z. Zhang,
"Explicit Tracking with Wild Card Routes in Multicast
VPN", draft-ietf-bess-mvpn-expl-track-09 (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.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol
Label Switching Architecture", RFC 3031,
DOI 10.17487/RFC3031, January 2001,
<<https://www.rfc-editor.org/info/rfc3031>>.
- [RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S.
Yasukawa, Ed., "Extensions to Resource Reservation
Protocol - Traffic Engineering (RSVP-TE) for Point-to-
Multipoint TE Label Switched Paths (LSPs)", RFC 4875,
DOI 10.17487/RFC4875, May 2007,
<<https://www.rfc-editor.org/info/rfc4875>>.
- [RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B.
Thomas, "Label Distribution Protocol Extensions for Point-
to-Multipoint and Multipoint-to-Multipoint Label Switched
Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011,
<<https://www.rfc-editor.org/info/rfc6388>>.
- [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>>.
- [RFC6625] Rosen, E., Ed., Rekhter, Y., Ed., Hendrickx, W., and R.
Qiu, "Wildcard in Multicast VPN Auto-Discovery Routes",
RFC 6625, DOI 10.17487/RFC6625, May 2012,
<<https://www.rfc-editor.org/info/rfc6625>>.

- [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>>.
- [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <<https://www.rfc-editor.org/info/rfc8296>>.

10.2. Informative 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>>.

Authors' Addresses

Jingrong Xie
Huawei Technologies

Email: xiejingrong@huawei.com

Mike McBride
Huawei Technologies

Email: mmcbride7@gmail.com

Mach Chen
Huawei Technologies

Email: mach.chen@huawei.com

Liang Geng
China Mobile
Beijing 100053

Email: gengliang@chinamobile.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

J. Xie
Huawei Technologies
L. Geng
L. Wang
China Mobile
M. McBride
G. Yan
Huawei
July 2, 2018

Segmented MVPN Using IP Lookup for BIER
draft-xie-bier-mvpn-segmented-01

Abstract

This document specifies an alternative of the control plane and data plane procedures that allow segmented MVPN using BIER. This allows the use of a more efficient explicit-tracking as the BIER overlay, with a slight change in the forwarding procedure of a segmentation point BFR by a lookup of the IP header. This document updates [I-D.ietf-bier-mvpn].

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

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 <https://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 January 3, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem Statement and Considerations	3
3.1. Problem Statement	3
3.2. Considerations	4
4. Segmented MVPN using IP Lookup for BIER	4
4.1. Explicit-tracking using LIR-pF Flag	4
4.2. Forwarding Procedure of Segmentation Point	7
5. Security Considerations	7
6. IANA Considerations	7
7. Acknowledgements	7
8. References	7
8.1. Normative References	7
8.2. Informative References	8
Authors' Addresses	8

1. Introduction

When using BIER to transport an MVPN data packet through a BIER domain, an ingress PE functions as a BFIR (see [RFC8279]). The BFIR must determine the set of BFERs to which the packet needs to be delivered. This can be done through an explicit-tracking function using a LIR and/or LIR-pF flag in BGP MVPN routes, per the [RFC6513],[RFC6514],[RFC6625],[I-D.ietf-bess-mvpn-expl-track], and [I-D.ietf-bier-mvpn].

Using a LIR-pF Flag will bring some extra benefits, as [I-D.ietf-bier-mvpn] and [I-D.ietf-bess-mvpn-expl-track] have stated. But unfortunately, the LIR-pF explicit tracking for a segmented MVPN deployment is not allowed in the current draft [I-D.ietf-bier-mvpn],

because the draft requires a per-flow upstream-assigned label to do the data-plane per-flow lookup on the segmentation point BFR.

This document specifies an alternative of the control plane and data plane procedures that allow segmented MVPN using BIER in both segments. This allows the use of the more efficient LIR-pF explicit-tracking as the BIER overlay, with a slight change in the forwarding procedure of a segmentation point BFR by using IP lookup. This will bring some significant benefits to the segmented MVPN deployment, including:

- o Getting a much better multicast join latency by eliminating the round trip interaction of S-PMSI AD routes and Leaf AD routes. Especially, the S-PMSI A-D routes may need a data-driven procedure to trigger, and make the multicast join latency even worse.
- o Greatly reducing the number of S-PMSI A-D routes that BFIR and BFRs need to save.
- o Consolidated forwarding procedure of IP lookup for every BIER Overlay functioning routers, such as BFIR, BFER, segmentation point BFR, and segmentation point BFR with BFER function.

2. Terminology

Readers of this document are assumed to be familiar with the terminology and concepts of the documents listed as Normative References.

3. Problem Statement and Considerations

3.1. Problem Statement

BIER is a stateless multicast forwarding by introducing a multicast-specific BIER header in the data plane. The maximal number of BFRs a packet can reach is limited by the bit string length of a BIER header. For a network with many routers in multiple IGP areas (typically an Inter-Area network), it may be more expected to use a segmented MVPN when deploying BIER than traditional MVPN.

However, it is not allowed in the [I-D.ietf-bier-mvpn] to use a LIR-pF explicit-tracking when deploying a segmented MVPN. This will lead to a low efficiency of explicit-tracking, and cause a worse multicast join latency. Here we take a scenario of inter-area segmented MVPN with both segments using BIER as an example.

3.2. Considerations

A BFIR is always needed to know the BFERs interested in a specific flow. This is a function of a BIER overlay defined in [RFC8279]. A segmentation point BFR in a segmented MVPN deployment, saying ABR, will play similar roles of both BFIR and BFER. It needs to do a disposition of a BIER Header, and then do an imposition of a new BIER Header. It requires the ABR router to maintain per-flow states, and especially, such per-flow states always include a set of BFERs who are interested in a specific flow by using an explicit-tracking procedure.

This behavior is completely different from a traditional segmented MVPN deployment, e.g, with both of the two segments using P2MP label switch.

In a traditional segmented MVPN with both segments using P2MP label switch, it is expected to receive a MPLS packet and replicate to downstream routers after swap the MPLS Label. A lookup of IP packet is not expected. Also, in a traditional segmented MVPN deployment, an MPLS label represents a P-tunnel, which may carry one, many or even all multicast flow(s) of a VPN, so it is not always a per-flow state on the segmentation point router.

In conclusion, the pattern of forwarding packets on segmentation points only by lookup of MPLS label mapped from multicast flow(s) is significantly unnecessary when BIER is introduced. Instead, doing a per-flow lookup of IP header on segmentation points is more efficient and consolidated.

4. Segmented MVPN using IP Lookup for BIER

4.1. Explicit-tracking using LIR-pF Flag

In a scenario of Inter-area Segmented MVPN with both segments using BIER, the determination of the set of BFERs that need to receive the a specific multicast flow of (C-S1,C-G1) in each segment, can be obtained by using a LIR-pF flag. Suppose a topology of this:

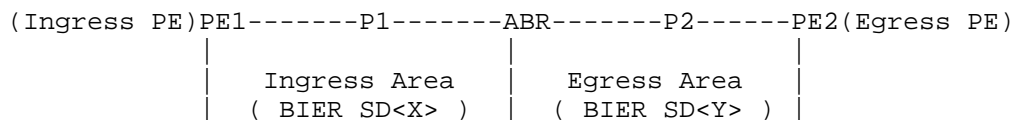


Figure 1: Example topology

PE1 is Ingress PE, and the area of { PE1 -- P1 -- ABR } is called an Ingress Area.

PE2 is Egress PE, and the area of { ABR -- P2 -- PE2 } is called an Egress Area.

The Ingress PE is configured to use a BIER tunnel type for a MVPN instance for the Ingress Area, and the ABR is configured to use a BIER tunnel type for the MVPN instance for the Egress Area.

The Ingress PE originates a wildcard S-PMSI A-D route (C-*,C-*) and the PTA of that route has the following settings:

- o The LIR-pF and LIR flags be set.
- o The tunnel type be set to "BIER".
- o A non-zero MPLS label be specified.

ABR receives the S-PMSI A-D route from the Ingress PE, and re-advertises the route to the Egress PE, with a PTA type "BIER", and PTA flags of LIR and LIR-pF, and a new non-zero upstream-assigned MPLS label allocated by ABR per-VPN.

Egress PE receives the S-PMSI A-D route from the ABR, and checks if it need to response with a Leaf A-D route to this S-PMSI A-D route using the process of the "match for reception" and "match for tracking" as defined in [I-D.bess-mvpn-expl-track]. In this example, for a C-flow of (C-S1, C-G1), the checking result of "matched for tracking" is the S-PMSI(C-*, C-*), and the checking result of "matched for reception" is also the S-PMSI(C-*, C-*). Egress PE will then send a Leaf A-D route (RD, C-S1, C-G1, Root=PE1, Leaf=PE2) to the ABR with a PTA flag LIR-pF, and a Leaf A-D route (RD, C-*, C-*, Root=PE1, Leaf=PE2) without a PTA flag LIR-pF.

ABR then has an explicit-tracking result of a new per-flow information of (RD, C-S1, C-G1, Root=PE1) with Egress PE as its leaf or BFER. ABR's "matched for tracking" result to this flow(RD, C-S1, C-G1, PE1) will then be updated with a new record, and ABR then sends a Leaf A-D route (RD, C-S1, C-G1, Root=PE1, Leaf=ABR) to Ingress PE.

Ingress PE then has an explicit-tracking result of a new per-flow information of (RD, C-S1, C-G1, Root=PE1) with ABR as its leaf or BFER.

From this procedure description one can see that:

1. The S-PMSI A-D(C-*, C-*) route is functioning as a per-VPN anchor of the upstream and the downstream(s), which can be called a BIER FEC in this document, saying BIER FEC(*,*).

2. The Leaf A-D(S,G) routes are functioning as a per-flow anchor of the downstream(s) and the upstream, which are also BIER FECs accordingly, saying BIER FEC(S,G).
3. The Tuple of (Root=PE1, RD) in S-PMSI (C-*, C-*) or Leaf AD(C-*, C-*) or Leaf AD(C-S, C-G) represents an VRF on the ABR implicitly.

ABR knows the per-vpn information of a (Root=PE1, RD) tuple when receiving and re-advertising the S-PMSI A-D(*,*) route bound with a PTA, where:

- o Inbound SD (InSD): in PTA of the received S-PMSI(*,*) route.
- o Inbound VpnLabel (InVpnLabel): in PTA of the received S-PMSI(*,*) route.
- o Inbound BfirId (InBfirId): in PTA of the received S-PMSI(*,*) route.
- o Outbound SD(OutSD): in PTA of the re-advertised S-PMSI(*,*) route.
- o Outbound VpnLabel (OutVpnLabel): in PTA of the re-advertised S-PMSI(*,*) route.
- o Outbound BfirId (OutBfirId): in PTA of the re-advertised S-PMSI(*,*) route.

ABR establishes a per-flow control-plane state accordingly like this:

- o Per-flow upstream state, according to the Leaf A-D (C-S, C-G) route send to the Ingress PE: (PE1, RD, C-S1, C-G1, InSD, InBfirId, InVpnLabel).
- o Per-flow downstream state(s), according to the Leaf A-D(C-S, C-G) route(s) received by the ABR from Egress PE(s): (PE1, RD, C-S1, C-G1, Leaf, OutSD, OutBfirId, OutVpnLabel).

ABR knows the BIER Label(s) it allocated for InSD and OutSD, saying InBierLabel for InSD<X> and OutBierLabel for OutSD<Y>, and thus it can establish the per-flow forwarding state:

- o Per-flow upstream forwarding state: (InBierLabel, InBfirId, InVpnLabel, C-S1, C-G1).
- o Per-flow downstream(s) forwarding state: (InBierLabel, InBfirId, InVpnLabel, C-S1, C-G1, Leaf, OutBfirId, OutVpnLabel, OutBitString)

4.2. Forwarding Procedure of Segmentation Point

The Forwarding procedure of a segmentation point BFR is a combination of a deposition and a re-imposition of the whole BIER header and the upstream-assigned Vpn Label. One can think it as swapping of a series of fields like below:

- o swapping the InBierLabel with an OutBierLabel.
- o swapping the InBfirId with an OutBfirId.
- o swapping the InVpnLabel with an OutVpnLabel.
- o swapping the InBitString with an OutBitString.

The key of a per-flow lookup on ABR is a tuple of (InBierLabel, InBfirId, InVpnLabel) and a tuple of (C-S1, C-G1), representing a VRF and a flow respectively. All the elements are from a BIER packet, and such an IP lookup can be seen the same as an MFIB lookup, if the (InBierLabel, InBfirId, InVpnLabel) tuple is mapped to a VRF locally on the ABR.

5. Security Considerations

The procedures of this document do not, in themselves, provide privacy, integrity, or authentication for the control plane or the data plane.

6. IANA Considerations

No IANA allocation is required.

7. Acknowledgements

TBD.

8. References

8.1. Normative References

- [I-D.ietf-bess-mvpn-expl-track]
Dolganow, A., Kotalwar, J., Rosen, E., and Z. Zhang,
"Explicit Tracking with Wild Card Routes in Multicast
VPN", draft-ietf-bess-mvpn-expl-track-09 (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.
- [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>>.
- [RFC6625] Rosen, E., Ed., Rekhter, Y., Ed., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, DOI 10.17487/RFC6625, May 2012, <<https://www.rfc-editor.org/info/rfc6625>>.
- [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>>.

8.2. Informative 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>>.

Authors' Addresses

Jingrong Xie
Huawei Technologies

Email: xiejingrong@huawei.com

Liang Geng
China Mobile
Beijing 10053

Email: gengliang@chinamobile.com

Lei Wang
China Mobile
Beijing 10053

Email: wangleiyjy@chinamobile.com

Mike McBride
Huawei

Email: mmcbride7@gmail.com

Gang Yan
Huawei

Email: yangang@huawei.com

BIER
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

Z. Zhang
Juniper Networks
July 2, 2018

BIER Penultimate Hop Popping
draft-zzhang-bier-php-00

Abstract

Bit Index Explicit Replication (BIER) can be used as provider tunnel for MVPN/GTM [RFC6514] [RFC7716] or EVPN BUM [RFC7432]. It is possible that not all routers in the provider network support BIER and there are various methods to handle BIER incapable transit routers. However the MVPN/EVPN PEs are assumed to be BIER capable - they are BFIRs/BFERs. This document specifies a method to allow BIER incapable routers to act as MVPN/EVPN PEs with BIER as the transport, by having the upstream BFR (connected directly or indirectly via a tunnel) of a PE remove the BIER header and send the payload to the PE.

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

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 <https://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 January 3, 2019.

Copyright Notice

Copyright (c) 2018 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 (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Terminologies	2
2. Introduction	2
3. Specifications	3
4. Security Considerations	4
5. IANA Considerations	4
6. Acknowledgements	5
7. References	5
7.1. Normative References	5
7.2. Informative References	6
Author's Address	6

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 BIER architecture includes three layers: the "routing underlay", the "BIER layer", and the "multicast flow overlay". The multicast flow overlay is responsible for the BFERs to signal to BFIRs that they are interested in receiving certain multicast flows so that BFIRs can encode the correct bitstring for BIER forwarding by the BIER layer.

MVPN and EVPN are two similar overlays where BGP Auto-Discovery routes for MVPN/EVPN are exchanged among all PEs to signal which PEs need to receive multicast traffic for all or certain flows.

Typically the same provider tunnel type is used for traffic to reach all receiving PEs.

Consider an MVPN/EVPN deployment where enough P/PE routers are BIER capable for BIER to become the preferred the choice of provider tunnel. However, some PEs cannot be upgraded to support BIER forwarding. While there are ways to allow an ingress PE to send traffic to some PEs with one type of tunnel and send traffic to some other PEs with a different type of tunnel, the procedure becomes complicated and forwarding is not optimized.

One way to solve this problem is to use Penultimate Hop Popping (PHP) so that the upstream BFR can pop the BIER header and send the payload "natively" (note that the upstream BFR can be connected directly or indirectly via a tunnel to the PE). This is similar to MPLS PHP though it is the BIER header that is popped. In case of MPLS encapsulation, even the signaling is similar - a BIER incapable router signals as if it supported BIER, but to request PHP at the penultimate hop, it signals an Implicit Null label instead of a regular BIER label as the Label Range Base in its BIER MPLS Encapsulation sub-TLV.

In order for the PE to be able to correctly forward the packets resulting from the PHP, certain conditions must be met, as specified in Section 3.

While the above text uses MVPN/EVPN as example, BIER PHP is applicable to any scenario where the multicast flow overlay edge router does not support BIER.

This works well if a BIER incapable PE only needs to receive multicast traffic. If it needs to send multicast traffic as well, then it must Ingress Replicate to a BIER capable helper PE, who will in turn relay the packet to other PEs. The helper PE is either a Virtual Hub as specified in [RFC7024] for MVPN and [I-D.keyupate-bess-evpn-virtual-hub] for EVPN, or an AR-Replicator as specified in [I-D.ietf-bess-evpn-optimized-ir] for EVPN.

3. Specifications

The procedures in this section can be applied only if, by means outside the scope of this document, it is known that one of the following conditions is met.

- o The payload after BIER header is IPv4 or IPv6 (i.e., the Proto field in the BIER header is 4 or 6).

Notice that in this case the Destination Address in the IPv4/IPv6 header must be in the address space for the BIER layer.

- o The payload after BIER header is MPLS packet with downstream-assigned label at top of stack (i.e., the Proto field in the BIER header is 2), For example, labels from a Domain-wide Common Block (DCB) are used as specified in [I-D.zhang-bess-mvpn-evpn-aggregation-label].

For MPLS encapsulation, a BIER incapable router, if acting as a multicast flow overlay router, MUST signal its BIER information as specified in [RFC8401] or [I-D.ietf-bier-ospf-bier-extensions] or [I-D.ietf-bier-idr-extensions], with the Label Range Base in the BIER MPLS Encapsulation sub-TLV set to Implicit Null Label [RFC3032].

For non-MPLS encapsulation, a PHP sub-sub-TLV is included in the BIER sub-TLV attached to the BIER incapable router's BIER prefix to request BIER PHP from other BFRs. The sub-sub-TLV's type is TBD, and the length is 0. The PHP sub-sub-TLV MAY be used for MPLS encapsulation as well.

If a BFR follows section 6.9 of [RFC8279] to handle BIER incapable routers, it must treat a router as BIER incapable if the Label Range Base advertised by the router is Implicit Null, or if the router advertises a PHP sub-sub-TLV, so that the router is not used as a transit BFR.

If the downstream neighbor for a BIER prefix is the one advertising the prefix with a PHP sub-sub-TLV or with an Implicit Null Label as the Label Range Base in its BIER MPLS Encapsulation sub-sub-TLV, then when the corresponding BIRT or BIFT entry is created/updated, the forwarding behavior MUST be that the BIER header is removed and the payload be sent to the downstream router without the BIER header, either directly or over a tunnel.

4. Security Considerations

To be added.

5. IANA Considerations

This document requests a new sub-sub-TLV type value from the "Sub-sub-TLVs for BIER Info Sub-TLV" registry in the "IS-IS TLV Codepoints" registry:

Type	Name
----	----
TBD	BIER PHP Request

This document also requests a new sub-TLV type value from the OSPFv2 Extended Prefix TLV Sub-TLV registry:

Type	Name
----	----
TBD	BIER PHP Request

6. Acknowledgements

The authors want to thank Eric Rosen and Antonie Przygienda for their review, comments and suggestions.

7. References

7.1. Normative References

- [I-D.ietf-bess-evpn-optimized-ir]
Rabadan, J., Sathappan, S., Henderickx, W., Sajassi, A., Isaac, A., and M. Katiyar, "Optimized Ingress Replication solution for EVPN", draft-ietf-bess-evpn-optimized-ir-03 (work in progress), February 2018.
- [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-05 (work in progress), March 2018.
- [I-D.ietf-bier-ospf-bier-extensions]
Psenak, P., Kumar, N., Wijnands, I., Dolganow, A., Przygienda, T., Zhang, Z., and S. Aldrin, "OSPFv2 Extensions for BIER", draft-ietf-bier-ospf-bier-extensions-18 (work in progress), June 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-00 (work in progress), March 2017.
- [I-D.zzhang-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-zzhang-bess-mvpn-evpn-aggregation-label-01 (work in progress), April 2018.

- [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>>.
- [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <<https://www.rfc-editor.org/info/rfc8296>>.
- [RFC8401] Ginsberg, L., Ed., Przygienda, T., Aldrin, S., and Z. Zhang, "Bit Index Explicit Replication (BIER) Support via IS-IS", RFC 8401, DOI 10.17487/RFC8401, June 2018, <<https://www.rfc-editor.org/info/rfc8401>>.

7.2. Informative References

- [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>>.

Author's Address

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net