

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 25, 2021

H. Chen  
M. McBride  
Futurewei  
A. Wang  
China Telecom  
G. Mishra  
Verizon Inc.  
Y. Liu  
China Mobile  
Y. Fan  
Casa Systems  
L. Liu  
Fujitsu  
X. Liu  
Volta Networks  
February 21, 2021

BIER Egress Protection  
draft-chen-bier-egress-protect-01

Abstract

This document describes a mechanism for fast protection against the failure of an egress node of a "Bit Index Explicit Replication" (BIER) domain. It does not have any per-flow state in the core of the domain. For a multicast packet to an egress node of the domain, when the egress node fails, its upstream hop as a PLR sends the packet to the egress' backup node once the PLR detects the failure.

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] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 August 25, 2021.

## 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
1.1. Terminology . . . . .	3
2. Overview of BIER Egress Protection . . . . .	4
3. Protocol Extensions . . . . .	5
3.1. Extensions to OSPF . . . . .	5
3.2. Extensions to IS-IS . . . . .	6
4. BIER Extensions . . . . .	7
4.1. Egress Protection Bit Index Routing Tables . . . . .	7
4.2. Egress Protection Bit Index Forwarding Tables . . . . .	9
4.3. Updated Forwarding Procedure . . . . .	9
4.4. Switching between EP and Normal Forwarding . . . . .	10
5. Example Application of BIER Egress Protection . . . . .	11
5.1. Example BIER Topology . . . . .	11
5.2. BIRT and BIFT on a BFR . . . . .	12
5.3. EP-BIRTs and EP-BIFTs on a BFR . . . . .	13
5.4. Forwarding using EP-BIFT . . . . .	15
6. Security Considerations . . . . .	17
7. IANA Considerations . . . . .	17
8. Acknowledgements . . . . .	17
9. References . . . . .	17
9.1. Normative References . . . . .	17
9.2. Informative References . . . . .	18
Authors' Addresses . . . . .	19



## 1. Introduction

[RFC8279] specifies "Bit Index Explicit Replication" (BIER). It provides optimal forwarding of multicast data packets through a "multicast/BIER domain". It does not require the use of a protocol for explicitly building multicast distribution trees, and it does not require intermediate nodes to maintain any per-flow state.

This document describes a mechanism for fast protection against the failure of an egress node of a "Bit Index Explicit Replication" (BIER) domain, which is called BIER Egress Protection.

This BIER Egress Protection does not require intermediate nodes to maintain any per-flow state for fast protection against the failure of an egress node of the flow.

### 1.1. Terminology

BFR: Bit-Forwarding Router.

BFIR: Bit-Forwarding Ingress Router.

BFER: Bit-Forwarding Egress Router.

BFR-id: BFR Identifier. It is a number in the range [1,65535].

BFR-NBR: BFR Neighbor.

F-BM: Forwarding Bit Mask.

BFR-prefix: An IP address (either IPv4 or IPv6) of a BFR.

BIRT: Bit Index Routing Table. It is a table that maps from the BFR-id (in a particular sub-domain) of a BFER to the BFR-prefix of that BFER, and to the BFR-NBR on the path to that BFER.

BIFT: Bit Index Forwarding Table.

FRR: Fast Re-Route.

PLR: Point of Local Repair.

LFA: Loop-Free Alternate.

RLFA: Remote LFA.

DLFA: Remote LFA with Directed forwarding.



IGP: Interior Gateway Protocol.

LSDB: Link State DataBase.

SPF: Shortest Path First.

SPT: Shortest Path Tree.

OSPF: Open Shortest Path First.

IS-IS: Intermediate System to Intermediate System.

LSA: Link State Advertisement in OSPF.

LSP: Link State Protocol Data Unit (PDU) in IS-IS.

SPT-old(R): The SPT rooted at node R using LSDB before X fails (i.e., old LSDB).

SPT-new(R, X): The SPT rooted at node R using LSDB without X after X fails (i.e., new LSDB).

P-Space  $P(R, X)$ : The set of nodes that are reachable from R without going through X. In other words, it is the set of nodes that are not downstream of X in SPT-old(R).

Extended P-Space  $P'(R, X)$ : The set of nodes that are reachable from R or a neighbor of R, without going through X.

Q-Space  $Q(D, X)$ : The set of nodes that do not use X to reach destination D using the old LSDB.

PQ node(R, X): A member of both the P-Space  $P(R, X)$  (or the extended P-Space  $P'(R, X)$ ) and the Q-Space  $(D, X)$ .

## 2. Overview of BIER Egress Protection

For fast protecting an egress node of a BIER domain, a backup egress node is configured on the egress node. After the configuration, the egress node distributes the information about the backup egress to its direct neighbors.

For clearly distinguishing between an egress node and a backup egress node, an egress node is called a primary egress node sometimes.

For a multicast packet to a primary egress node of the domain, when the primary egress node fails, its upstream hop as a point of local



repair (PLR) sends the packet to the backup egress node configured to protect the primary egress node once the PLR detects the failure.

A Bit-Forwarding Router (BFR) in a BIER sub-domain builds and maintains an "Egress Protection Bit Index Routing Table" (EP-BIRT) for each of its BFR Neighbors (BFR-NBRs) that are egress nodes of the domain to provide fast protection against the failure of an egress node. The BFR builds each EP-BIRT based on a BIRT defined in [RFC8279]. An "Egress Protection Bit Index Forwarding Table" (EP-BIFT) is derived from an EP-BIRT in a way that is similar to the way in which a BIFT is derived from a BIRT, which is defined in [RFC8279].

Once the BFR as a PLR detects the failure of its BFR-NBR X that is a primary egress node of the domain, for a multicast packet targeting to the primary egress node, the PLR uses the EP-BIFT for X to send the packet to the backup egress node configured to protect the primary egress node.

### 3. Protocol Extensions

This section defines extensions to OSPF and IS-IS for advertising the backup information (including the backup egress node for protecting a primary egress node) to its direct neighbors.

#### 3.1. Extensions to OSPF

When a node P (as a primary egress node) has a backup egress node configured to protect against its failure, node P advertises the information about the backup egress node to its neighbors in its router information opaque LSA of LS type 9 or 10. The information is included in a backup egress node TLV. The format of the TLV is shown in Figure 1.

After each of the neighbors receives the backup egress node TLV, it knows that node P as a primary egress node will be protected by the backup egress node in the TLV. Once detecting the failure of node P, it sends the packet targeting to node P towards the backup egress node.



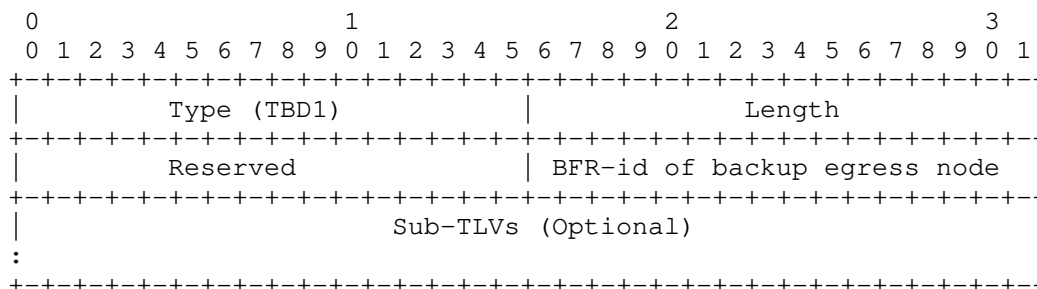


Figure 1: OSPF Backup Egress TLV

Type: 2 octets, its value (TBD1) is to be assigned by IANA.

Length: 2 octets, its value is 4 plus the length of the Sub-TLVs included. If no Sub-TLV is included, its value is 4.

Reserved: 2 octets, it MUST be set to zero when sending and be ignored while receiving.

BFR-id of backup egress node: 2 octets, its value is the BFR-id of the backup egress node configured to protect against the failure of the primary egress node.

Sub-TLVs (Optional): No Sub-TLV is defined now.

### 3.2. Extensions to IS-IS

For supporting fast protection against the failure of a primary egress node in a BIER domain, a new IS-IS TLV, called IS-IS backup egress node TLV, is defined. It contains the BFR-id of a backup egress node.

When a node P (as a primary egress node) has a backup egress node configured to protect against its failure, node P advertises the information about the backup egress node to its neighbors using a IS-IS backup egress node TLV.

This TLV may be advertised in IS-IS Hello (IIH) PDUs, LSPs, or in Circuit Scoped Link State PDUs (CS-LSP) [RFC7356]. The format of the TLV is shown in Figure 2.



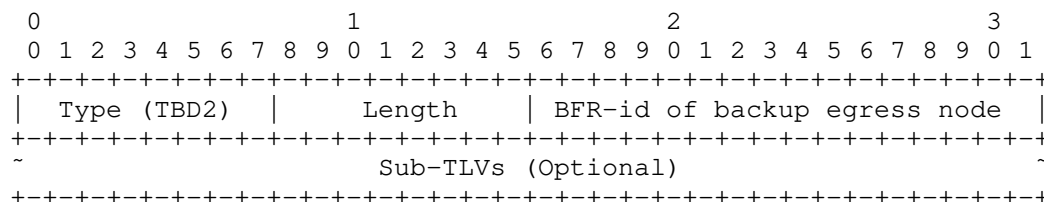


Figure 2: IS-IS Backup Egress TLV

Type: 1 octet, its value (TBD2) is to be assigned by IANA.

Length: 1 octet, its value is 2 plus the length of the Sub-TLVs included. If no Sub-TLV is included, its value is 2.

BFR-id of backup egress node: 2 octets, its value is the BFR-id of the backup egress node configured to protect against the failure of the primary egress node.

Sub-TLVs (Optional): No Sub-TLV is defined now.

## 4. BIER Extensions

#### 4.1. Egress Protection Bit Index Routing Tables

If a BFR is a direct neighbor of an egress node in a BIER sub-domain, it builds and maintains a number of "Egress Protection Bit Index Routing Tables" (EP-BIRTs). There is an EP-BIRT for each of the BFR's neighbors that are egress nodes of the domain. The BFR builds each EP-BIRT based on its BIRT. Comparing to the BIRT, an EP-BIRT has a piece of new backup information for each BFER.

The new backup information for a BFER indicates if the BFER as an egress node is protected by the BFR. If so, the information further includes the backup egress node configured to protect the BFER.

In one implementation, the new backup information is represented by {EP, BE-BFER}. EP (short for Egress Protection) is a flag, indicating whether the BFER as an egress node is protected. EP = 1 means that the BFER is protected. EP = 0 means that the BFER is not protected. BE-BFER (short for Backup Egress BFER) is the BFER (i.e., BFER-id) of the backup egress node when EP = 1. BE-BFER is NULL (0) when EP = 0.

In the EP-BIRT for BFR-NBR X that is an egress node, the row having X as BFER and as its next hop BFR-NBR contains the new backup information {EP = 1, BE-BFER}, where BE-BFER is the BFER (i.e., BFER-id) of backup egress node for protecting the egress node. Each of



the other rows in the EP-BIRT contains the new backup information {EP = 0, BE-BFER = NULL}.

When the egress node fails, for a multicast packet targeting to the primary egress node BFER (PE-BFER), the BFR sends the packet to the BE-BFER through using the route to the backup egress node. The BFR clears the bit for PE-BFER and adds the bit for BE-BFER in the packet's BitString first, and then forwards the packet according to the forwarding entry for BE-BFER.

The EP-BIRT for BFR-NBR X that is an egress node considers the failure of X. It has a route or say a next hop (i.e., BFR-NBR N on the path, where N is not X) to every BFER except for X.

The BFR may build the EP-BIRT for BFR-NBR X by copying its BIRT to the EP-BIRT and sets the new information for each BFER to empty such as {EP = 0, BE-BFER = NULL} first. And then it updates each of the rows in the EP-BIRT that has X as BFER or next hop BFR-NBR X.

For the BFR-id of a BFER in the EP-BIRT for egress node X, when the next hop BFR-NBR on the path to the BFER is X, the BFR checks whether the BFER is X. If the BFER is not X, the BFR changes next hop BFR-NBR X to a backup next hop (BNH) when there is a BNH on a backup path to the BFER without going through X and the link from the BFR to X. If the BFER is X, the BFR adds the new backup information {EP = 1, BE-BFER} for the BFER as PE-BFER.

If there is not any BNH to a BFER to protect against the failure of X, the next hop BFR-NBR X to the BFER in the EP-BIRT for BFR-NBR X is changed to NULL. For a multicast packet having the BFER as one of its destinations, if the next hop BFR-NBR to the BFER is NULL, the BFR does not send the packet to the next hop BFR-NBR NULL but drops it when X fails.

Note: In another option, the next hop BFR-NBR X to the BFER in the EP-BIRT for BFR-NBR X keeps unchanged when there is not any BNH to the BFER to protect against the failure of X. In this case, for a multicast packet having the BFER as one of its destinations, the BFR sends the packet to X when X fails.

In one implementation, the BNH is the Loop-Free Node-Protecting Alternate defined in [RFC5286] to protect against the failure of X and link from the BFR to X. In another implementation, the BNH is the virtual Loop-Free Alternate (LFA), i.e., PQ node, defined in [RFC7490]. In a special case, a PQ node is a Loop-Free Node-Protecting Alternate defined in [RFC5286].



#### 4.2. Egress Protection Bit Index Forwarding Tables

From each EP-BIRT on the BFR, an "Egress Protection Bit Index Forwarding Table" (EP-BIFT) is derived. In addition to having a route to a BFER in each row of the EP-BIFT which is the same as the EP-BIRT, it has a "Forwarding Bit Mask" (F-BM) in its each row. For the rows in the EP-BIRT that have the same SI and the same BFR-NBR and the same new backup information {EP, BE-BFER}, the F-BM for each of these rows in the EP-BIFT is the logical OR of the BitStrings of these rows.

This EP-BIFT is programmed into the data plane and is not used to forward any packet in normal operations. It is activated to forward a packet with a BIER header once the BFR detects the failure of BFR-NBR. The header contains SI, BitString, BitStringLength, and sub-domain.

#### 4.3. Updated Forwarding Procedure

The forwarding procedure defined in [RFC8279] is updated/enhanced for an EP-BIFT to consider the egress protection (i.e., the new information {EP, BE-BFER} in the EP-BIFT). For a multicast packet with the BitString indicating a BFER as one of its destinations, the updated forwarding procedure sends the packet towards the backup egress node of the BFER if the BFER is protected. It checks whether EP = 1 in the forwarding entry for the BFER. If EP = 1, the procedure clears the bit for the BFER as PE-BFER and adds the bit for BE-BFER in the packet's BitString first, and then forwards the packet using the row (i.e., forwarding entry) for BE-BFER.

The updated procedure is described in Figure 3. It is used with an EP-BIFT for BFR-NBR X as egress node on a BFR to forward multicast packets when X fails. It can also be used with a BIFT on the BFR to forward multicast packets in normal operations if the new backup information in each row of the BIFT is empty such as {EP = 0, BE-BFER = NULL}.



```

Packet = the packet received by BFR;
FOR each BFER k (from the rightmost in Packet's BitString) {
  IF BFER k is the BFR itself {
    copies Packet, sends the copy to the multicast
    flow overlay and clears bit k in Packet's BitString
  } ELSE {
    finds the row in the EP-BIFT for the sub-domain using
    Packet's SI and BitString as the key/index
    IF EP == 1 {
      clears bit k in Packet's BitString; //BFER k is PE-BFER
      adds bit j in Packet's BitString; //BFER j is BE-BFER
    } ELSE {
      IF BFR-NBR in the row is not NULL {
        Copies Packet, updates the copy's BitString by ANDing
        it with F-BM in the row, sends updated copy to BFR-NBR
      } // BFR-NBR == NULL, not sent Packet to BFR-NBR
      updates Packet's BitString by ANDing it with the INVERSE
      of the F-BM in the row
    }
  }
}

```

Figure 3: Updated Forwarding Procedure

#### 4.4. Switching between EP and Normal Forwarding

The EP-BIFTs will be pre-computed and installed ready for activation when an egress node failure is detected. Once the BFR detects the failure of its BFR-NBR X as an egress, it activates the EP-BIFT for X to forward packets with BIER headers and de-activates its BIFT. After activation of the EP-BIFT, it remains in effect until it is no longer required.

In general, when the routing protocol has re-converged on the new topology taking into account the failure of X, the BIRT is re-computed using the updated LSDB and the BIFT is re-derived from the BIRT. Once the BIFT is installed ready for activation, it is activated to forward packets with BIER headers and the EP-BIFT for X is de-activated.

From the new topology, the BFR computes/re-computes the EP-BIRT for each BFR-NBR Y as an egress of the BFR and the EP-BIFT for Y is derived/re-derived from the EP-BIRT for Y. The EP-BIFT is installed/re-installed ready for activation when Y fails.



## 5. Example Application of BIER Egress Protection

This section illustrates an example application of BIER Egress Protection on a BFR in a BIER topology in Figure 4.

### 5.1. Example BIER Topology

An example BIER topology for a BIER sub-domain is shown in Figure 4. It has 8 nodes/BFRs A, B, C, D, E, F, G and H. Each of the links connecting these nodes/BFRs has a cost. The link cost of 1 is default and is not indicated in the figure. The link costs of other values such as 2 and 3 are indicated in the figure.

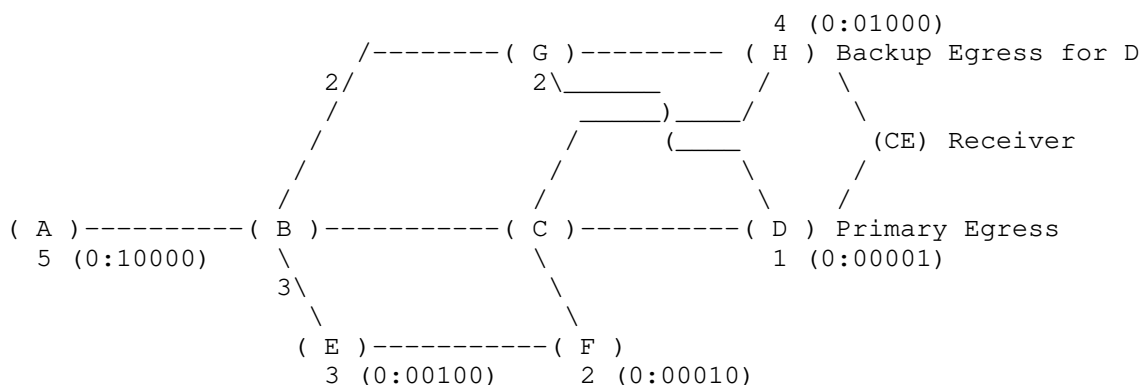


Figure 4: Example BIER Topology

Nodes/BFRs D, F, E, H and A are BFRs and have BFR-ids 1, 2, 3, 4, and 5 respectively. For simplicity, these BFR-ids are represented by (SI:BitString), where SI = 0 and BitString is of 5 bits. BFR-ids 1, 2, 3, 4, and 5 are represented by (0:00001), (0:00010), (0:00100), (0:01000) and (0:10000) respectively.

BFR H is configured to protect BFR D on BFR D. Suppose that this information is distributed to BFR D's neighbors BFR C and BFR G by IGP. BFR C and BFR G know that H is the backup egress to protect the primary egress D.

CE is a multicast traffic Receiver, which is dual homed to primary egress node D and backup egress node H for protecting primary egress D. During normal operations, there is no multicast traffic to CE from backup egress node H and CE receives the multicast traffic only from primary egress node D. There is no duplicated traffic to receiver CE. This is different from MoFRR in [RFC7431], where the same traffic is sent through two separated paths/trees to both primary egress node D and backup egress node H, to which the receiver



CE is dual homed. When primary egress node D fails, the multicast traffic is sent to CE from backup egress node H.

The fast egress protection mechanism in this document will use less network resources such as link bandwidth than MoFRR in [RFC7431].

## 5.2. BIRT and BIFT on a BFR

Every BFR in a BIER sub-domain/topology builds and maintains a Bit Index Routing Table (BIRT). For the BIER topology in Figure 4, each of 8 nodes/BFRs A, B, C, D, E, F, G and H builds and maintains a BIRT using the LSDB for the topology.

The BIRT built on BFR C (i.e. node C) is shown in Figure 5.

BFR-id (SI:BitString)	BFR-Prefix of Dest BFER	BFR-NBR (Next Hop)
1 (0:00001)	D	D
2 (0:00010)	F	F
3 (0:00100)	E	F
4 (0:01000)	H	H
5 (0:10000)	A	B

Figure 5: Bit Index Routing Table on BFR C

The 1st row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER D with BFR-id 1 is BFR D.

The 2nd row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER F with BFR-id 2 is BFR F.

The 3rd row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER E with BFR-id 3 is BFR F.

The 4-th row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER H with BFR-id 4 is BFR H.

The 5-th row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER A with BFR-id 5 is BFR B.



From this BIRT on BFR C, a Bit Index Forwarding Table (BIFT) is derived. This BIFT is shown in Figure 6.

The 2nd and 3-th rows in the BIRT have the same SI = 0 and next hop BFR-NBR = F. The F-BM for each of these two rows in the BIFT is the logical OR of the BitStrings of these rows, which is 00110 (00010 OR 00100 = 00110).

The F-BM for 1st row in the BIFT is 00001.

The F-BM for 4-th row in the BIFT is 01000.

The F-BM for 5-th row in the BIFT is 10000.

BFR-id (SI:BitString)	F-BM	BFR-NBR (Next Hop)
1 (0:00001)	00001	D
2 (0:00010)	00110	F
3 (0:00100)	00110	F
4 (0:01000)	01000	H
5 (0:10000)	10000	B

Figure 6: Bit Index Forwarding Table on BFR C

### 5.3. EP-BIRTs and EP-BIFTs on a BFR

Each of the BFRs that are neighbors of egress nodes (i.e., BFERs) in a BIER sub-domain/topology builds and maintains a number of Egress Protection Bit Index Routing Tables (EP-BIRTs).

For the BIER topology in Figure 4,

BFR B is the neighbor of BFERs A and E;  
 BFR C is the neighbor of BFERs D, F and H;  
 BFR E is the neighbor of BFER F;  
 BFR F is the neighbor of BFER E;  
 BFR G is the neighbor of BFERs D and H.

Each of 5 nodes/BFRs B, C, E, F and G builds and maintains a number of EP-BIRTs using the LSDB for the topology for its every BFR-NBR as an egress node.



For example, BFR C (i.e., node C) in the BIER topology builds and maintains three EP-BIRTs for its three BFR-NBRs (BFRs D, F and H) that are egress nodes respectively.

The EP-BIRT for BFER D built by BFR C based on the BIRT on BFR C (refer to Figure 5) is shown in Figure 7.

The BIRT is copied to the EP-BIRT for BFER D (i.e., the first three columns of the EP-BIRT). The new backup information (i.e., the 4-th column) for every row in the EP-BIRT is initialized to {EP = 0, BE-BFER = 0/NULL}.

BFR-id (SI:BitString)	BFR-Prefix of Dest BFER	BFR-NBR (Next Hop)	{EP,BE-BFER} (Backup Info)
1 (0:00001)	D	D/NULL	EP=1, BE-BFER=H
2 (0:00010)	F	F	EP=0, BE-BFER=0
3 (0:00100)	E	F	EP=0, BE-BFER=0
4 (0:01000)	H	H	EP=0, BE-BFER=0
5 (0:10000)	A	B	EP=0, BE-BFER=0

Figure 7: EP-BIRT for BFER D on BFR C

In the EP-BIRT for BFER D, the row that has BFR-NBR == D is the 1st row. This row has the new backup information {EP = 1, BE-BFER = H}, which indicates that BFER D (i.e., primary egress node D) is protected by BFER H (i.e., backup egress node H). Each of the other rows has the new backup information {EP = 0, BE-BFER = 0/NULL}.

The 1st row in the EP-BIRT indicates that the next hop BFR-NBR on the path to BFER D with BFR-id 1 is NULL (changed to NULL from D). There is no backup next hop (BNH) to D when D fails.

The 2nd row in the EP-BIRT indicates that the next hop BFR-NBR on the path to BFER F with BFR-id 2 is BFR F.

The 3rd row in the EP-BIRT indicates that the next hop BFR-NBR on the path to BFER E with BFR-id 3 is BFR F.

The 4-th row in the EP-BIRT indicates that the next hop BFR-NBR on the path to BFER H with BFR-id 4 is BFR H.



The 5-th row in the EP-BIRT indicates that the next hop BFR-NBR on the path to BFER A with BFR-id 5 is BFR B.

From this EP-BIRT for BFER D on BFR C, an Egress Protection Bit Index Forwarding Table (EP-BIFT) is derived. This EP-BIFT for BFER D is shown in Figure 8.

The 2nd and 3rd rows in the EP-BIRT have the same SI = 0, the same next hop BFR-NBR = E and the same backup information {EP=0, BE-BFER=0}. The F-BM for each of these two rows in the EP-BIFT is the logical OR of the BitStrings of these rows, which is 00110 (00010 OR 00100 = 00110).

BFR-id (SI:BitString)	F-BM	BFR-NBR (Next Hop)	{EP,BE-BFER} (Backup Info)
1 (0:00001)	00001	NULL	EP=1, BE-BFER=H
2 (0:00010)	00110	F	EP=0, BE-BFER=0
3 (0:00100)	00110	F	EP=0, BE-BFER=0
4 (0:01000)	01000	H	EP=0, BE-BFER=0
5 (0:10000)	10000	B	EP=0, BE-BFER=0

Figure 8: EP-BIFT for BFER D on BFR C

The F-BM for 1st row in the EP-BIFT is 00001.

The F-BM for 4-th row in the EP-BIFT is 01000.

The F-BM for 5-th row in the EP-BIFT is 10000.

#### 5.4. Forwarding using EP-BIFT

Suppose that there is a multicast traffic from BFR A as ingress/BFIR to egresses/BFERs D, F and E. For every packet of the traffic, after receiving it, BFR A adds a BIER header into the packet and sends the packet with the BIER header to BFR B, which sends the packet BFR C. The BIER header contains (SI:BitString) = (0:00111) for egresses/BFERs D, F and E.

In normal operations, after receiving the packet from BFR B, BFR C copies, updates and sends the packet to BFR D and BFR F using the BIFT on BFR C according to the forwarding procedure defined in [RFC8279].



Once BFR C detects the failure of its BFR-NBR D, which is a BFER, after receiving the packet from BFR B, BFR C copies, updates and sends the packet using the EP-BIFT for BFER D on BFR C according to the updated forwarding procedure.

For the packet targeting to BFER D (i.e., primary egress node), BFR C sends it towards BFER H (i.e., backup egress node), which is configured to protect BFER D.

For example, once BFR C detects the failure of its BFR-NBR D, after receiving the packet from BFR B, BFR C copies, updates and sends the packet to BFR H and BFR F using the EP-BIFT for BFER D on BFR C.

The packet received by BFR C from BFR B contains (SI:BitString) = (0:00111). The rightmost one bit in BitString is bit 1. For BFER 1 (0:00001) (i.e., BFR D as BFER), BFR C gets the 1st row (i.e., forwarding entry) in the EP-BIFT for BFER D. EP = 1 in the row indicates that BFER D is protected against the failure of D. BFR C clears bit 1 in Packet's BitString and sets bit 4 (i.e., the bit for BE-BFER = H) in Packet's BitString to one. The BitString in Packet is 01110 now. This lets BFR C send Packet to BE-BFER H.

For the packet containing BitString = 01110, the rightmost one bit in BitString is bit 2. For BFER 2 (0:00010) (i.e., BFR F as BFER), BFR C gets the 2nd row (i.e., forwarding entry) in the EP-BIFT for BFER D. EP = 0 and the next hop BFR-NBR is F in the row. BFR C copies, updates and sends the packet to F.

The packet sent to F contains the updated BitString = 00110, which is 01110 & F-BM in the 2nd row = 01110 & 00110 = 00110.

After sending the packet to F, BFR C updates the original packet by ANDing its BitString with the INVERSE of the F-BM in the 2nd row. The updated BitString = 01000, which is 01110 & ~F-BM in the row = 01110 & 11001 = 01000.

For the packet containing BitString = 01000, the rightmost one bit in BitString is bit 4. For BFER 4 (0:01000) (i.e., BFR H as BFER), BFR C gets the 4-th row (i.e., forwarding entry) in the EP-BIFT for BFER D. EP = 0 and the next hop BFR-NBR is H in the row. BFR C copies, updates and sends the packet to H. The packet sent to H contains BitString = 01000.

After sending the packet to H, BFR C updates the original packet by ANDing its BitString with the INVERSE of the F-BM in the 4-th row. The updated BitString = 00000, which is 01000 & ~F-BM in the row = 01000 & 10111 = 00000.



The updated packet has BitString without any one bit. BFR C finishes forwarding the packet to F and H (backup for D). BFR F will send the packet to E.

## 6. Security Considerations

TBD.

## 7. IANA Considerations

No requirements for IANA.

## 8. Acknowledgements

The authors would like to thank people for their comments to this work.

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <<https://www.rfc-editor.org/info/rfc5226>>.
- [RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, DOI 10.17487/RFC5250, July 2008, <<https://www.rfc-editor.org/info/rfc5250>>.
- [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>>.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <<https://www.rfc-editor.org/info/rfc5714>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.



- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, DOI 10.17487/RFC7490, April 2015, <<https://www.rfc-editor.org/info/rfc7490>>.
- [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>>.
- [RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <<https://www.rfc-editor.org/info/rfc7770>>.
- [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>>.
- [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>>.

## 9.2. Informative References

- [I-D.ietf-rtgwg-segment-routing-ti-lfa]  
Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-05 (work in progress), November 2020.
- [I-D.ietf-spring-segment-protection-sr-te-paths]  
Hegde, S., Bowers, C., Litkowski, S., Xu, X., and F. Xu, "Segment Protection for SR-TE Paths", draft-ietf-spring-segment-protection-sr-te-paths-00 (work in progress), September 2020.



- [RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <<https://www.rfc-editor.org/info/rfc7431>>.
- [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>>.
- [RFC8444] Psenak, P., Ed., Kumar, N., Wijnands, IJ., Dolganow, A., Przygienda, T., Zhang, J., and S. Aldrin, "OSPFv2 Extensions for Bit Index Explicit Replication (BIER)", RFC 8444, DOI 10.17487/RFC8444, November 2018, <<https://www.rfc-editor.org/info/rfc8444>>.

## Authors' Addresses

Huaimo Chen  
Futurewei  
Boston, MA  
USA

Email: [Huaimo.chen@futurewei.com](mailto:Huaimo.chen@futurewei.com)

Mike McBride  
Futurewei

Email: [michael.mcbride@futurewei.com](mailto:michael.mcbride@futurewei.com)

Aijun Wang  
China Telecom  
Beiqijia Town, Changping District  
Beijing, 102209  
China

Email: [wangaj3@chinatelecom.cn](mailto:wangaj3@chinatelecom.cn)



Gyan S. Mishra  
Verizon Inc.  
13101 Columbia Pike  
Silver Spring MD 20904  
USA

Phone: 301 502-1347  
Email: gyan.s.mishra@verizon.com

Yisong Liu  
China Mobile

Email: liuyisong@chinamobile.com

Yanhe Fan  
Casa Systems  
USA

Email: yfan@casa-systems.com

Lei Liu  
Fujitsu

USA

Email: liulei.kddi@gmail.com

Xufeng Liu  
Volta Networks

McLean, VA  
USA

Email: xufeng.liu.ietf@gmail.com



Network Working Group  
Internet-Draft  
Intended status: Experimental  
Expires: August 25, 2021

H. Chen  
M. McBride  
Futurewei  
A. Wang  
China Telecom  
G. Mishra  
Verizon Inc.  
Y. Liu  
China Mobile  
Y. Fan  
Casa Systems  
L. Liu  
Fujitsu  
X. Liu  
Volta Networks  
February 21, 2021

BIER Fast ReRoute  
draft-chen-bier-frr-02

Abstract

This document describes a mechanism for fast re-route (FRR) protection against the failure of a node or link in the core of a "Bit Index Explicit Replication" (BIER) domain. It does not have any per-flow state in the core. For a multicast packet to traverse a node in the domain, when the node fails, its upstream hop as a PLR reroutes the packet around the failed node once it detects the failure.

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] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 August 25, 2021.

#### 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
1.1. Terminology . . . . .	4
2. BIER FRR Solution . . . . .	5
2.1. Overview of BIER forwarding . . . . .	5
2.2. FRR Bit Index Routing Tables . . . . .	6
2.3. FRR Bit Index Forwarding Tables . . . . .	7
2.4. Updated Forwarding Procedure . . . . .	7
2.5. Switching between FRR and Normal Forwarding . . . . .	8
3. Example Application of BIER FRR . . . . .	8
3.1. Example BIER Topology . . . . .	9
3.2. BIRT and BIFT on a BFR . . . . .	9
3.3. FRR-BIRTs and FRR-BIFTs on a BFR . . . . .	11
3.4. Forwarding using FRR-BIFT . . . . .	13
4. Security Considerations . . . . .	14
5. IANA Considerations . . . . .	14
6. Acknowledgements . . . . .	15
7. References . . . . .	15
7.1. Normative References . . . . .	15
7.2. Informative References . . . . .	16
Authors' Addresses . . . . .	17



## 1. Introduction

[RFC8279] specifies "Bit Index Explicit Replication" (BIER). It provides optimal forwarding of multicast data packets through a "multicast/BIER domain". It does not require the use of a protocol for explicitly building multicast distribution trees, and it does not require intermediate nodes to maintain any per-flow state.

[I-D.merling-bier-frr] proposes a tunnel-based fast re-route (FRR) method for protecting a node or link in the core of a BIER domain, which is called tunnel-based BIER-FRR. It tunnels BIER packets around the failure to BIER nodes downstream in multicast distribution trees. For a (next hop) node failure, it tunnels BIER packets to the next next hop nodes (NNHs). The BIFT in every BFR is enhanced to have two forwarding entries for every BFER. One is the primary forwarding entry with primary NH such as BFR neighbor and primary bit mask, and the other is the backup forwarding entry with backup NH such as NNH and backup bit mask. Using one BIFT in a BFR for both normal and backup forwarding will save memory.

In normal operations, the primary forwarding entries are used to forward BIER packets. When a failure such as a node failure happens, the backup forwarding entry corresponding to the failure and the other primary forwarding entries are used to forward BIER packets. In the BIFT, the primary bit mask in every primary forwarding entry is computed before the failure. After the failure, the primary bit mask needs to be recomputed from the changed topology. Before the primary bit mask is recomputed and updated, some of BIER packets may be forwarded incorrectly.

This document describes a mechanism for fast re-route (FRR) protection against the failure of a node or link in the core of a BIER domain, which resolves the above issue. It is based on LFA, which is called LFA-based BIER-FRR. On a BFR, there is a FRR BIFT for each of its neighbors, which has considered the neighbor failure. There is one forwarding entry for every BFER in any BIFT, including normal BIFT and FRR BIFT. This may use more memory.

In normal operations, the normal BIFT is used to forward BIER packets. When a neighbor fails, the BFR as PLR uses the FRR BIFT for the neighbor to forward BIER packets. For a BIER packet to traverse the BFR and the failed neighbor, the BFR reroutes the packet around the failed neighbor using the FRR BIFT for the neighbor. For a BIER packet to traverse the BFR and any other neighbors, the BFR forwards the packet to its expected next hop neighbors using the forwarding entries with these BFR neighbors in the FRR BIFT.



### 1.1. Terminology

BFR: Bit-Forwarding Router.

BFIR: Bit-Forwarding Ingress Router.

BFER: Bit-Forwarding Egress Router.

BFR-id: BFR Identifier. It is a number in the range [1,65535].

BFR-NBR: BFR Neighbor.

F-BM: Forwarding Bit Mask.

BFR-prefix: An IP address (either IPv4 or IPv6) of a BFR.

BIRT: Bit Index Routing Table. It is a table that maps from the BFR-id (in a particular sub-domain) of a BFER to the BFR-prefix of that BFER, and to the BFR-NBR on the path to that BFER.

BIFT: Bit Index Forwarding Table.

FRR: Fast Re-Route.

PLR: Point of Local Repair.

LFA: Loop-Free Alternate.

RLFA: Remote LFA.

DLFA: Remote LFA with Directed forwarding.

IGP: Interior Gateway Protocol.

LSDB: Link State DataBase.

SPF: Shortest Path First.

SPT: Shortest Path Tree.

SPT-old(R): The SPT rooted at node R using LSDB before X fails (i.e., old LSDB).

SPT-new(R, X): The SPT rooted at node R using LSDB without X after X fails (i.e., new LSDB).



P-Space  $P(R,X)$ : The set of nodes that are reachable from R without going through X. In other words, it is the set of nodes that are not downstream of X in SPT-old(R).

Extended P-Space  $P'(R,X)$ : The set of nodes that are reachable from R or a neighbor of R, without going through X.

Q-Space  $Q(D,X)$ : The set of nodes that do not use X to reach destination D using the old LSDB.

PQ node(R,X): A member of both the P-Space  $P(R, X)$  (or the extended P-Space  $P'(R, X)$ ) and the Q-Space  $(D, X)$ .

## 2. BIER FRR Solution

A Bit-Forwarding Router (BFR) in a BIER sub-domain builds and maintains a "FRR Bit Index Routing Table" (FRR-BIRT) for each of its BFR Neighbors (BFR-NBRs) to provide BIER-FRR. The BFR builds each FRR-BIRT based on a BIRT defined in [RFC8279]. A "FRR Bit Index Forwarding Table" (FRR-BIFT) is derived from a FRR-BIRT in the same way as a BIFT is derived from a BIRT, which is defined in [RFC8279].

The forwarding procedure defined in [RFC8279] is enhanced/updated for FRR-BIFTs. Once the BFR as a PLR detects the failure of its BFR-NBR X, it uses the FRR-BIFT for X to forward packets with BIER headers to get around failed X according to the updated/enhanced forwarding procedure.

### 2.1. Overview of BIER forwarding

This section briefs the BIRT, BIFT and forwarding procedure defined in [RFC8279].

There is a "Bit Index Routing Table" (BIRT) for a BIER sub-domain on a BFR. The BIRT maps the BFR Identifier (BFR-id) (in the sub-domain) of a Bit-Forwarding Egress Router (BFER) to the BFR-prefix of that BFER, and to the BFR-NBR on the shortest path to that BFER. In other words, the BIRT has a route or say a next hop (i.e., BFR-NBR on the path) to every BFER.

From the BIRT on the BFR, a "Bit Index Forwarding Table" (BIFT) is derived. In addition to having a route to a BFER in each row of the BIFT which is the same as the BIRT, it has a "Forwarding Bit Mask" (F-BM) in its each row. For the rows in the BIRT that have the same SI and the same BFR-NBR, the F-BM for each of these rows in the BIFT is the logical OR of the BitStrings of these rows.



This BIFT is programmed into the data plane and used to forward a packet with a BIER header. The header contains SI, BitString, BitStringLength, and sub-domain.

When a BFR receives a packet, for each BFER *k* (from the rightmost to the leftmost) represented in the SI and BitString of the packet, if BFER *k* is the BFR itself, the BFR copies the packet, sends the copy to the multicast flow overlay and clears bit *k* in the original packet; otherwise the BFR finds the row (i.e., forwarding entry) in the BIFT for the sub-domain using the SI and BitString as the key or say index, and then copies, updates and forwards the packet to the BFR-NBR (i.e., the next hop) indicated by the row (i.e., forwarding entry).

After copying the packet and before forwarding it to the BFR-NBR, the packet's BitString is updated by ANDing it with the F-BM in the forwarding entry (i.e., `PacketCopy->BitString &= F-BM`).

After forwarding the updated packet to a BFR-NBR and before forwarding the original packet to another BFR-NBR, the original packet's BitString is changed by ANDing it with the INVERSE of the F-BM (i.e., `Packet->BitString &= ~F-BM`).

## 2.2. FRR Bit Index Routing Tables

Each BFR in a BIER sub-domain builds and maintains a number of "FRR Bit Index Routing Tables" (FRR-BIRTs). There is a FRR-BIRT for each BFR-NBR of the BFR. The BFR builds each FRR-BIRT based on its BIRT. It has the same format as the BIRT.

The FRR-BIRT for BFR-NBR *X* of the BFR considers the failure of *X* and maps the BFR-id (in the sub-domain) of a BFER to the BFR-prefix of that BFER, and to BFR-NBR *N* on the path to that BFER. In other words, the FRR-BIRT has a route or say a next hop (i.e., BFR-NBR *N* on the path, where *N* is not *X*) to every BFER when BFR-NBR *X* fails.

The BFR may build the FRR-BIRT for BFR-NBR *X* by copying its BIRT to the FRR-BIRT first, and then change the next hop with value BFR-NBR *X* in the FRR-BIRT to a backup next hop (BNH) to protect against the failure of *X*. In other words, for the BFR-id of a BFER in the FRR-BIRT for BFR-NBR *X*, if the next hop BFR-NBR on the path to the BFER is *X*, it is changed to a BNH when there is a BNH on a backup path to the BFER without going through *X* and the link from the BFR to *X*.

If there is not any BNH to a BFER to protect against the failure of *X*, the next hop BFR-NBR *X* to the BFER in the FRR-BIRT for BFR-NBR *X* is changed to NULL. For a multicast packet having the BFER as one of its destinations, if the next hop BFR-NBR to the BFER is NULL, the



BFR does not send the packet to the next hop BFR-NBR NULL but drops it when X fails.

Note: In another option, the next hop BFR-NBR X to the BFER in the FRR-BIRT for BFR-NBR X keeps unchanged when there is not any BNH to the BFER to protect against the failure of X. In this case, for a multicast packet having the BFER as one of its destinations, the BFR sends the packet to X when X fails.

In one implementation, the BNH is the Loop-Free Node-Protecting Alternate defined in [RFC5286] to protect against the failure of X and link from the BFR to X. In another implementation, the BNH is the virtual Loop-Free Alternate (LFA), i.e., PQ node, defined in [RFC7490]. In a special case, a PQ node is a Loop-Free Node-Protecting Alternate defined in [RFC5286].

### 2.3. FRR Bit Index Forwarding Tables

From each FRR-BIRT on the BFR, a "FRR Bit Index Forwarding Table" (FRR-BIFT) is derived. In addition to having a route to a BFER in each row of the FRR-BIFT which is the same as the FRR-BIRT, it has a "Forwarding Bit Mask" (F-BM) in its each row. For the rows in the FRR-BIRT that have the same SI and the same BFR-NBR, the F-BM for each of these rows in the FRR-BIFT is the logical OR of the BitStrings of these rows.

This FRR-BIFT is programmed into the data plane and is not used to forward any packet in normal operations. It is activated to forward a packet with a BIER header once the BFR detects the failure of BFR-NBR. The header contains SI, BitString, BitStringLength, and sub-domain.

### 2.4. Updated Forwarding Procedure

The forwarding procedure defined in [RFC8279] is updated/enhanced for a FRR-BIFT to consider the case where the next hop BFR-NBR to a BFER is NULL. For a multicast packet with the BitString indicating a BFER as one of its destinations, the updated forwarding procedure checks whether the next hop BFR-NBR to the BFER in the FRR-BIFT is NULL. If it is NULL, the procedure will not send the packet to this next hop BFR-NBR NULL but drop the packet.

The updated procedure is described in Figure 1. It is used with a FRR-BIFT for BFR-NBR X on a BFR to forward multicast packets when X fails. It can also be used with a BIFT on the BFR to forward multicast packets in normal operations.



```
Packet = the packet received by BFR;
FOR each BFER k (from the rightmost in Packet's BitString) {
  IF BFER k is the BFR itself {
    copies Packet, sends the copy to the multicast
    flow overlay and clears bit k in Packet's BitString
  } else {
    finds the row in the FRR-BIFT for the sub-domain using
    Packet's SI and BitString as the key/index
    IF BFR-NBR in the row is not NULL {
      Copies Packet, updates the copy's BitString by ANDing
      it with F-BM in the row, sends updated copy to BFR-NBR
    } // BFR-NBR == NULL, not sent Packet to BFR-NBR
    updates Packet's BitString by ANDing it with the INVERSE
    of the F-BM in the row
  }
}
```

Figure 1: Updated Forwarding Procedure

### 2.5. Switching between FRR and Normal Forwarding

The FRR-BIFTs will be pre-computed and installed ready for activation when a failure is detected. Once the BFR detects the failure of its BFR-NBR X, it activates the FRR-BIFT for X to forward packets with BIER headers and de-activates its BIFT. After activation of the FRR-BIFT, it remains in effect until it is no longer required.

In general, when the routing protocol has re-converged on the new topology taking into account the failure of X, the BIRT is re-computed using the updated LSDB and the BIFT is re-derived from the BIRT. Once the BIFT is installed ready for activation, it is activated to forward packets with BIER headers and the FRR-BIFT for X is de-activated.

From the new topology, the BFR computes/re-computes the FRR-BIRT for each BFR-NBR Y of the BFR and the FRR-BIFT for Y is derived/re-derived from the FRR-BIRT for Y. The FRR-BIFT is installed/re-installed ready for activation when Y fails.

### 3. Example Application of BIER FRR

This section illustrates an example application of BIER FRR on a BFR in a BIER topology in Figure 2.



### 3.1. Example BIER Topology

An example BIER topology for a BIER sub-domain is shown in Figure 2. It has 8 nodes/BFRs A, B, C, D, E, F, G and H. Each of the links connecting these nodes/BFRs has a cost. The link cost of 1 is default and is not indicated in the figure. The link cost of other value such as 2 is indicated in the figure.

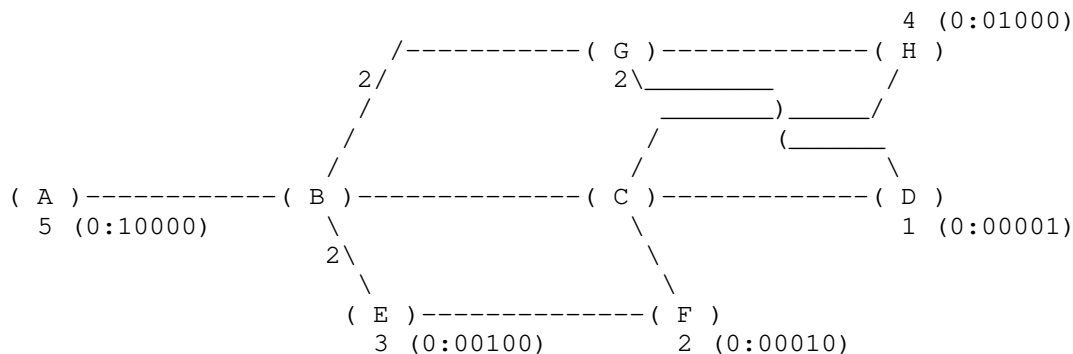


Figure 2: Example BIER Topology

Nodes/BFRs D, F, E, H and A are BFRs and have BFR-ids 1, 2, 3, 4, and 5 respectively. For simplicity, these BFR-ids are represented by (SI:BitString), where SI = 0 and BitString is of 5 bits. BFR-ids 1, 2, 3, 4, and 5 are represented by (0:00001), (0:00010), (0:00100), (0:01000) and (0:10000) respectively.

### 3.2. BIRT and BIFT on a BFR

Every BFR in a BIER sub-domain/topology builds and maintains a Bit Index Routing Table (BIRT). For the BIER topology in Figure 2, each of 8 nodes/BFRs A, B, C, D, E, F, G and H builds and maintains a BIRT using the LSDB for the topology.

The BIRT built on BFR B (i.e. node B) is shown in Figure 3.



BFR-id (SI:BitString)	BFR-Prefix of Dest BFER	BFR-NBR (Next Hop)
1 (0:00001)	D	C
2 (0:00010)	F	C
3 (0:00100)	E	E
4 (0:01000)	H	C
5 (0:10000)	A	A

Figure 3: BIRT on BFR B

The 1st row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER D with BFR-id 1 is BFR C.

The 2nd row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER F with BFR-id 2 is BFR C.

The 3rd row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER E with BFR-id 3 is BFR E.

The 4-th row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER H with BFR-id 4 is BFR C.

The 5-th row in the BIRT indicates that the next hop BFR-NBR on the shortest path to BFER A with BFR-id 5 is BFR A.

From this BIRT on BFR B, a Bit Index Forwarding Table (BIFT) is derived. This BIFT is shown in Figure 4.

The 1st, 2nd and 4-th rows in the BIRT have the same SI = 0 and next hop BFR-NBR = C. The F-BM for each of these three rows in the BIFT is the logical OR of the BitStrings of these rows, which is 01011 (00001 OR 00010 OR 01000 = 01011).

The F-BM for 3rd row in the BIFT is 00100. The F-BM for 5-th row in the BIFT is 10000.



BFR-id (SI:BitString)	F-BM	BFR-NBR (Next Hop)
1 (0:00001)	01011	C
2 (0:00010)	01011	C
3 (0:00100)	00100	E
4 (0:01000)	01011	C
5 (0:10000)	10000	A

Figure 4: BIFT on BFR B

### 3.3. FRR-BIRTs and FRR-BIFTs on a BFR

Every BFR in a BIER sub-domain/topology builds and maintains a number of FRR Bit Index Routing Tables (FRR-BIRTs). For the BIER topology in Figure 2, each of 8 nodes/BFRs A, B, C, D, E, F, G and H builds and maintains a number of FRR-BIRTs using the LSDB for the topology for its every BFR-NBR.

For example, BFR B (i.e., node B) in the BIER topology builds and maintains four FRR-BIRTs for its four BFR-NBRs (BFR C, BFR E, BFR A and BFR G) respectively. The FRR-BIRT for BFR C built by BFR B is shown in Figure 5.

BFR-id (SI:BitString)	BFR-Prefix of Dest BFER	BFR-NBR (Next Hop)
1 (0:00001)	D	G
2 (0:00010)	F	E
3 (0:00100)	E	E
4 (0:01000)	H	G
5 (0:10000)	A	A

Figure 5: FRR BIRT for BFR C on BFR B



The 1st row in the FRR-BIRT indicates that the next hop BFR-NBR on the path to BFER D with BFR-id 1 is BFR G. G is the Loop-Free Node-Protecting Alternate defined in [RFC5286] to protect against the failure of C and link from B to C.

The 2nd row in the FRR-BIRT indicates that the next hop BFR-NBR on the path to BFER F with BFR-id 2 is BFR E. E is the Loop-Free Node-Protecting Alternate defined in [RFC5286] to protect against the failure of C and link from B to C.

The 3rd row in the FRR-BIRT indicates that the next hop BFR-NBR on the path to BFER E with BFR-id 3 is BFR E.

The 4-th row in the FRR-BIRT indicates that the next hop BFR-NBR on the path to BFER H with BFR-id 4 is BFR G. G is the Loop-Free Node-Protecting Alternate defined in [RFC5286] to protect against the failure of C and link from B to C.

The 5-th row in the FRR-BIRT indicates that the next hop BFR-NBR on the path to BFER A with BFR-id 5 is BFR A.

From this FRR-BIRT for BFR C on BFR B, a FRR Bit Index Forwarding Table (FRR-BIFT) is derived. This FRR-BIFT for BFR C is shown in Figure 6.

The 1st and 4-th rows in the FRR-BIRT have the same SI = 0 and next hop BFR-NBR = G. The F-BM for each of these two rows in the FRR-BIFT is the logical OR of the BitStrings of these rows, which is 01001 (00001 OR 01000 = 01001).

BFR-id (SI:BitString)	F-BM	BFR-NBR (Next Hop)
1 (0:00001)	01001	G
2 (0:00010)	00110	E
3 (0:00100)	00110	E
4 (0:01000)	01001	G
5 (0:10000)	10000	A

Figure 6: FRR BIFT for BFR C on BFR B



The 2nd and 3rd rows in the FRR-BIRT have the same SI = 0 and next hop BFR-NBR = E. The F-BM for each of these two rows in the FRR-BIFT is the logical OR of the BitStrings of these rows, which is 00110 (00010 OR 00100 = 00110).

The F-BM for 5-th row in the FRR-BIFT is 10000.

The number of entries in a FRR BIFT is the number of BFERs. Each FRR BIFT on a BFR can be compressed through combining all the entries with the same BFR-BNR and F-BM into one entry. The number of entries in a compressed FRR BIFT is the number of neighbors of the BFR minus one.

For example, the compressed FRR-BIFT for BFR C on BFR B is shown in Figure 7. The number of entries in it is three, which equals the number (four) of neighbors of BFR B minus one.

BFR-id (SI:BitString)	F-BM	BFR-NBR (Next Hop)
1, 4 (0:01001)	01001	G
2, 3 (0:00110)	00110	E
5 (0:10000)	10000	A

Figure 7: Compressed FRR BIFT for BFR C on BFR B

For a BIER packet with a BFR-ID as a destination, the entry containing the BFR-ID is used to forward the packet.

### 3.4. Forwarding using FRR-BIFT

Suppose that there is a multicast traffic from BFR A as ingress/BFIR to egresses/BFERs D, F, E and H. For every packet of the traffic, after receiving it, BFR A adds a BIER header into the packet and sends the packet with the BIER header to BFR B. The BIER header contains (SI:BitString) = (0:01111) for egresses/BFERs D, F, E and H.

In normal operations, after receiving the packet from BFR A, BFR B copies, updates and sends the packet to BFR C and BFR E using the BIFT on BFR B according to the forwarding procedure defined in [RFC8279].

Once BFR B detects the failure of its BFR-NBR X, after receiving the packet from BFR A, BFR B copies, updates and sends the packet using



the FRR-BIFT for X on BFR B to avoid X and link from B to X according to the forwarding procedure defined in [RFC8279].

For example, once BFR B detects the failure of its BFR-NBR C, after receiving the packet from BFR A, BFR B copies, updates and sends the packet to BFR G and BFR E using the FRR-BIFT for BFR C on BFR B to avoid C and link from B to C.

The packet received by BFR B from BFR A contains (SI:BitString) = (0:01111). The rightmost one bit in BitString is bit 1. For BFER 1 (0:00001) (i.e., BFR D as BFER), BFR B gets the 1st row (i.e., forwarding entry) in the FRR-BIFT for BFR C. The next hop BFR-NBR is G in the row. BFR B copies, updates and forwards the packet to G.

The packet sent to G contains the updated BitString = 01001, which is 01111 & F-BM in the row = 01111 & 01001.

After sending the packet to G, BFR B updates the original packet by ANDing its BitString with the INVERSE of the F-BM in the row. The updated BitString = 00110, which is 01111 & ~F-BM in the row = 01111 & 00110.

For the packet containing BitString = 00110, the rightmost one bit in BitString is bit 2. For BFER 2 (0:00010) (i.e., BFR F as BFER), BFR B gets the 2nd row (i.e., forwarding entry) in the FRR-BIFT for BFR C. The next hop BFR-NBR is E in the row. BFR B copies, updates and forwards the packet to E.

The packet sent to E contains the updated BitString = 00110, which is 00110 & F-BM in the 2nd row = 00110 & 00110.

After sending the packet to E, BFR B updates the original packet by ANDing its BitString with the INVERSE of the F-BM in the 2nd row. The updated BitString = 00000, which is 00110 & ~F-BM in the row = 00110 & 11001.

The updated packet has BitString without any one bit. BFR B finishes forwarding the packet from A to D, F, E and H.

#### 4. Security Considerations

TBD.

#### 5. IANA Considerations

No requirements for IANA.



## 6. Acknowledgements

The authors would like to thank Jeffrey Zhang, Daniel Merling and Geng Xuesong for their comments to this work.

## 7. References

### 7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <<https://www.rfc-editor.org/info/rfc5226>>.
- [RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, DOI 10.17487/RFC5250, July 2008, <<https://www.rfc-editor.org/info/rfc5250>>.
- [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>>.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <<https://www.rfc-editor.org/info/rfc5714>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, DOI 10.17487/RFC7490, April 2015, <<https://www.rfc-editor.org/info/rfc7490>>.



- [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>>.
- [RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <<https://www.rfc-editor.org/info/rfc7770>>.
- [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>>.
- [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>>.

## 7.2. Informative References

- [I-D.ietf-rtgwg-segment-routing-ti-lfa]  
Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-05 (work in progress), November 2020.
- [I-D.ietf-spring-segment-protection-sr-te-paths]  
Hegde, S., Bowers, C., Litkowski, S., Xu, X., and F. Xu, "Segment Protection for SR-TE Paths", draft-ietf-spring-segment-protection-sr-te-paths-00 (work in progress), September 2020.
- [I-D.merling-bier-frr]  
Merling, D. and M. Menth, "BIER Fast Reroute", draft-merling-bier-frr-00 (work in progress), March 2019.
- [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>>.
- [RFC8444] Psenak, P., Ed., Kumar, N., Wijnands, IJ., Dolganow, A., Przygienda, T., Zhang, J., and S. Aldrin, "OSPFv2 Extensions for Bit Index Explicit Replication (BIER)", RFC 8444, DOI 10.17487/RFC8444, November 2018, <<https://www.rfc-editor.org/info/rfc8444>>.

## Authors' Addresses

Huaimo Chen  
Futurewei  
Boston, MA  
USA

Email: [Huaimo.chen@futurewei.com](mailto:Huaimo.chen@futurewei.com)

Mike McBride  
Futurewei

Email: [michael.mcbride@futurewei.com](mailto:michael.mcbride@futurewei.com)

Aijun Wang  
China Telecom  
Beiqijia Town, Changping District  
Beijing, 102209  
China

Email: [wangaj3@chinatelecom.cn](mailto:wangaj3@chinatelecom.cn)

Gyan S. Mishra  
Verizon Inc.  
13101 Columbia Pike  
Silver Spring MD 20904  
USA

Phone: 301 502-1347  
Email: [gyan.s.mishra@verizon.com](mailto:gyan.s.mishra@verizon.com)



Yisong Liu  
China Mobile

Email: liuyisong@chinamobile.com

Yanhe Fan  
Casa Systems  
USA

Email: yfan@casa-systems.com

Lei Liu  
Fujitsu

USA

Email: liulei.kddi@gmail.com

Xufeng Liu  
Volta Networks

McLean, VA  
USA

Email: xufeng.liu.ietf@gmail.com



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 25, 2021

H. Chen  
M. McBride  
Futurewei  
A. Wang  
China Telecom  
G. Mishra  
Verizon Inc.  
Y. Liu  
China Mobile  
Y. Fan  
Casa Systems  
L. Liu  
Fujitsu  
X. Liu  
Volta Networks  
February 21, 2021

BIER-TE Egress Protection  
draft-chen-bier-te-egress-protect-00

Abstract

This document describes a mechanism for fast protection against the failure of an egress node of an explicit point to multipoint (P2MP) multicast path/tree in a "Bit Index Explicit Replication" (BIER) Traffic Engineering (TE) domain. It does not have any per-flow state in the core of the domain. For a multicast packet to the egress node, when the egress node fails, its upstream hop as a PLR sends the packet to the egress' backup node once the PLR detects the failure.

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] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 August 25, 2021.

#### 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
1.1. Terminology . . . . .	3
2. Overview of BIER-TE Egress Protection . . . . .	4
3. Protocol Extensions . . . . .	5
3.1. Extensions to OSPF . . . . .	5
3.2. Extensions to IS-IS . . . . .	6
4. BIER-TE Extensions . . . . .	7
4.1. Extensions to BIER-TE BIFT for Egress Protection . . . . .	7
4.2. Updated Forwarding Procedure . . . . .	7
5. Example Application of BIER-TE Egress Protection . . . . .	9
5.1. Example BIER-TE Topology . . . . .	9
5.2. BIER-TE BIFT on a BFR . . . . .	10
5.3. Extended BIER-TE BIFT on a BFR . . . . .	11
5.4. Forwarding using Extended BIER-TE BIFT . . . . .	12
6. Security Considerations . . . . .	13
7. IANA Considerations . . . . .	13
8. Acknowledgements . . . . .	14
9. References . . . . .	14
9.1. Normative References . . . . .	14
9.2. Informative References . . . . .	15
Authors' Addresses . . . . .	16



## 1. Introduction

[I-D.ietf-bier-te-arch] introduces Bit Index Explicit Replication (BIER) Traffic/Tree Engineering (BIER-TE). It is an architecture for per-packet stateless explicit point to multipoint (P2MP) multicast path/tree and based on the BIER architecture defined in [RFC8279]. A multicast packet is replicated and forwarded along the P2MP path/tree encoded in the packet. It does not require intermediate nodes to maintain any per-path/tree state.

This document describes a mechanism for fast protection against the failure of an egress node of an explicit P2MP multicast path/tree in a BIER-TE domain. It is called BIER-TE Egress Protection. For a multicast packet to the egress node, when the egress node fails, its upstream hop as a PLR sends the packet to the egress' backup node (called backup egress node) once the PLR detects the failure.

This BIER-TE Egress Protection does not require intermediate nodes to maintain any per-path state for fast protection against the failure of an egress node of the path.

### 1.1. Terminology

BIER: Bit Index Explicit Replication.

BIER-TE: BIER Traffic/Tree Engineering.

BFR: Bit-Forwarding Router.

BFIR: Bit-Forwarding Ingress Router.

BFER: Bit-Forwarding Egress Router.

BFR-id: BFR Identifier. It is a number in the range [1,65535].

BFR-NBR: BFR Neighbor.

F-BM: Forwarding Bit Mask.

BFR-prefix: An IP address (either IPv4 or IPv6) of a BFR.

BIRT: Bit Index Routing Table. It is a table that maps from the BFR-id (in a particular sub-domain) of a BFER to the BFR-prefix of that BFER, and to the BFR-NBR on the path to that BFER.

BIFT: Bit Index Forwarding Table.

FRR: Fast Re-Route.



PLR: Point of Local Repair.

IGP: Interior Gateway Protocol.

LSDB: Link State DataBase.

SPF: Shortest Path First.

SPT: Shortest Path Tree.

OSPF: Open Shortest Path First.

IS-IS: Intermediate System to Intermediate System.

LSA: Link State Advertisement in OSPF.

LSP: Link State Protocol Data Unit (PDU) in IS-IS.

## 2. Overview of BIER-TE Egress Protection

For fast protecting an egress node of a BIER-TE domain, a backup egress node is configured on the egress node. After the configuration, the egress node distributes the information about the backup egress to its direct neighbors.

For clearly distinguishing between an egress node and a backup egress node, an egress node is called a primary egress node sometimes.

For a multicast packet to a primary egress node of the domain, when the primary egress node fails, its upstream hop as a point of local repair (PLR) sends the packet to the backup egress node configured to protect the primary egress node once the PLR detects the failure. The upstream hop of the primary egress node is its direct neighbor.

A Bit-Forwarding Router (BFR) in a BIER-TE sub-domain has a BIER-TE Bit Index Forwarding Tables (BIFT) [I-D.ietf-bier-te-arch]. A BIER-TE BIFT on a BFR comprises a forwarding entry for a BitPosition (BP) assigned to each of the adjacencies of the BFR. If the BP represents a forward connected adjacency, the forwarding entry for the BP forwards the multicast packet with the BP to the directly connected BFR neighbor of the adjacency. If the BP represents a BFER (i.e., egress node) or say a local decap adjacency, the forwarding entry for the BP decapsulates the multicast packet with the BP and passes a copy of the payload of the packet to the packet's NextProto within the BFR.

The BIER-TE BIFT on a BFR is extended to support protection against the failure of an egress node. For each forwarding entry of the



BIER-TE BIFT on the BFR, if it is for the BP representing a forward connected adjacency and its BFR-NBR is a BFER (i.e., primary egress node), the forwarding entry is extended to include a new forwarding entry, which is called backup forwarding entry or backup entry for short. This backup entry forwards the multicast packet with the BP to the backup egress node, which is configured to protect the primary egress node.

Once the BFR as a PLR detects the failure of its BFR-NBR X that is a primary egress node of the domain, for a multicast packet with the BP for the primary egress node, the PLR uses the backup forwarding entry in the extended BIER-TE BIFT to send the packet to the backup egress node configured to protect the primary egress node.

### 3. Protocol Extensions

This section defines extensions to OSPF and IS-IS for advertising the backup information (including the information about the backup egress node for protecting a primary egress node).

#### 3.1. Extensions to OSPF

When a node P (as a primary egress node) has a backup egress node configured to protect against its failure, node P advertises the information about the backup egress node to its neighbors in its router information opaque LSA of LS type 9 or 10. The information is included in a backup egress BP TLV. The format of the TLV is shown in Figure 1.

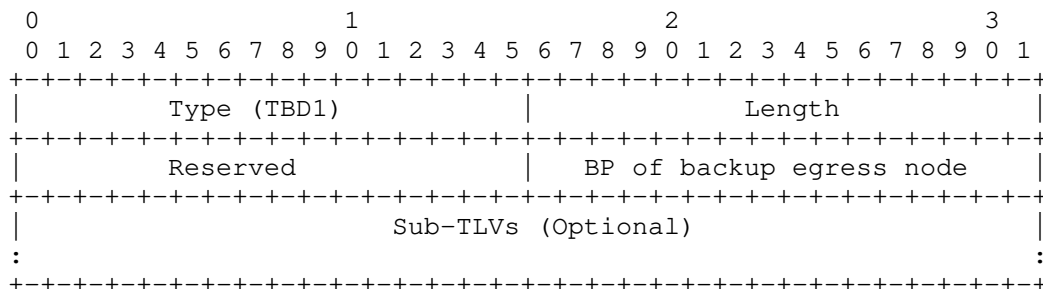


Figure 1: OSPF Backup Egress BP TLV

Type: 2 octets, its value (TBD1) is to be assigned by IANA.

Length: 2 octets, its value is 4 plus the length of the Sub-TLVs included. If no Sub-TLV is included, its value is 4.



Reserved: 2 octets, it MUST be set to zero when sending and be ignored while receiving.

BP of backup egress node: 2 octets, its value is the local decap BitPosition of the backup egress node, which is configured to protect against the failure of the primary egress node.

Sub-TLVs (Optional): No Sub-TLV is defined now.

After each of the neighbors receives the backup egress BP TLV from node P, it knows that node P as a primary egress node will be protected by the backup egress node in the TLV. Once detecting the failure of node P, it sends the packet with the BP for node P towards the backup egress node.

### 3.2. Extensions to IS-IS

For supporting fast protection against the failure of a primary egress node in a BIER-TE domain, a new IS-IS TLV, called IS-IS backup egress BP TLV, is defined. It contains the local decap BitPosition of the backup egress node configured to protect the primary egress node.

When a node P (as a primary egress node) has a backup egress node configured to protect against its failure, node P advertises the information about the backup egress node to its neighbors using a IS-IS backup egress BP TLV.

This TLV may be advertised in IS-IS Hello (IIH) PDUs, LSPs, or in Circuit Scoped Link State PDUs (CS-LSP) [RFC7356]. The format of the TLV is shown in Figure 2.

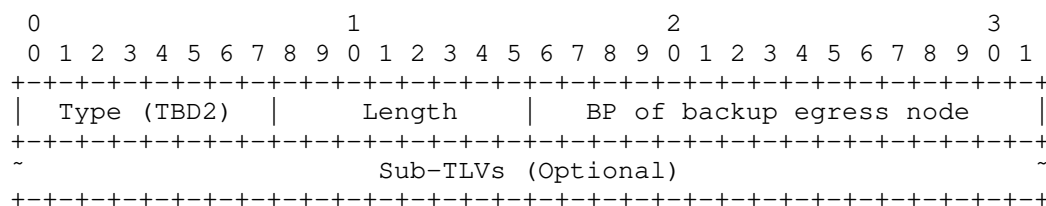


Figure 2: IS-IS Backup Egress BP TLV

Type: 1 octet, its value (TBD2) is to be assigned by IANA.

Length: 1 octet, its value is 2 plus the length of the Sub-TLVs included. If no Sub-TLV is included, its value is 2.



BP of backup egress node: 2 octets, its value is the local decap BitPosition of the backup egress node configured to protect against the failure of the primary egress node.

Sub-TLVs (Optional): No Sub-TLV is defined now.

#### 4. BIER-TE Extensions

This section describes extensions to a BIER-TE BIFT of a BFR for supporting fast protection against the failure of a primary egress node and enhancements on a forwarding procedure to use the extended BIER-TE BIFT for egress protection.

##### 4.1. Extensions to BIER-TE BIFT for Egress Protection

If a BFR is a neighbor of an egress node in a BIER-TE sub-domain, it has an extended BIER-TE BIFT to support protection against the failure of its neighbor egress node. The forwarding entry with the egress node (say X) as its BFR-NBR in the BIFT comprises a backup entry. The backup entry contains a flag EPA (which is short for Egress Protection is Active) and a backup path to a backup egress node (say Y) which is configured to protect the egress node.

In normal operations, the flag EPA in the backup entry for neighbor egress node X is set to 0 (zero). The flag EPA is set to 1 (one) when egress node X fails. EPA == 1 means that the egress protection for primary egress node X is active and the backup entry will be used to forward the packet with BP for egress node X to backup egress node Y along the backup path.

The backup path from the BFR to backup egress node Y is a path that satisfies a set of constraints and does not traverse primary egress node X or any link connected to X. In one implementation, the backup path is represented by the BitPositions for the adjacencies along the backup path.

##### 4.2. Updated Forwarding Procedure

The forwarding procedure defined in [I-D.ietf-bier-te-arch] is updated/enhanced for using an extended BIER-TE BIFT to consider the egress protection (i.e., the new backup entry containing EPA and backup path in the BIFT).

For a multicast packet with the BP in the BitString indicating a BFR-NBR as a primary egress node, the updated forwarding procedure on a BFR sends the packet towards the backup egress node of the primary egress node if the primary egress node fails.



It checks whether EPA equals to 1 (one) in the forwarding entry with the BFR-NBR that is the primary egress node. If EPA is 1 (i.e., the primary egress node fails and the egress protection for primary egress is active), then the procedure clears two BPs in the packet's BitString and checks whether the backup egress node is not one of the packet's destinations.

One BP is the BP for the primary egress node and the other is the BP for the forward connected adjacency from the BFR to the primary egress node. After these two BPs are cleared in the packet's BitString, the packet will not be sent to the failed primary egress node.

When the BP for the backup egress node in the packet's BitString is 0, the backup egress node is not one of the packet's destinations. In this case, the procedure adds the backup path to the backup egress node into the packet through adding the BPs for the backup path in the packet's BitString. Thus the packet will be sent to the backup egress node along the backup path.

The updated procedure is described in Figure 3. It can also be used by the BFR to forward multicast packets in normal operations.

```

Packet = the packet received by BFR;
FOR each BP k (from the rightmost in Packet's BitString) {
  IF BP k is local decap adjacency (or say BP of the BFR) {
    copies Packet, sends copy's payload to the multicast
    flow overlay and clears bit k in Packet's BitString
  } ELSE IF BP k is forward connect adjacency of the BFR {
    finds the forwarding entry in the BIER-TE BIFT for the domain
    using BP k;
    Clears BP k;
    IF EPA == 1 { //Egress Protection for BFR-NBR/egress is Active
      Clears BP for BFR-NBR in Packet's BitString;
      IF BP for backup egress is 0 in Packet's BitString {
        Adds BPs for backup path into Packet's BitString
      }
    } //egress removed, backup path to backup egress added
  } ELSE {
    Copies Packet, updates the copy's BitString by
    clearing all the BPs for the adjacencies of the BFR,
    and sends the updated copy to BFR-NBR
  }
}
}

```

Figure 3: Updated Forwarding Procedure







For example, for the link between nodes B and C in the figure, two forward connected adjacency BitPositions 3' and 4' are assigned to two ends of the link. BitPosition 3' is assigned on node B to B's end of the link. It is the forward connected adjacency of node C. BitPosition 4' is assigned on node C to C's end of the link. It is the forward connected adjacency of node B.

BFER H is configured to protect BFER D on BFR D. Suppose that this information is distributed to BFR D's neighbors BFR C and BFR G by IGP. BFR C and BFR G know that H is the backup egress to protect the primary egress D.

Similarly, BFER D is configured to protect BFER H on BFR H; BFER F is configured to protect BFER E on BFR E; and BFER E is configured to protect BFER F on BFR F. These are not shown in the figure.

CE is a multicast traffic Receiver, which is dual homed to primary egress node D and backup egress node H for protecting primary egress D. During normal operations, there is no multicast traffic to CE from backup egress node H and CE receives the multicast traffic only from primary egress node D. There is no duplicated traffic to receiver CE. This is different from MoFRR in [RFC7431], where duplicated traffic is sent to both the primary egress D and backup egress H, to which the receiver CE is dual homed. When primary egress node D fails, the multicast traffic is sent to CE from backup egress node H.

## 5.2. BIER-TE BIFT on a BFR

Every BFR in a BIER-TE sub-domain/topology has a BIER-TE BIFT. For the BIER-TE topology in Figure 4, each of 8 nodes/BFRs A, B, C, D, E, F, G and H has its BIER-TE BIFT for the topology.

The BIER-TE BIFT on BFR C (i.e. node C) is shown in Figure 5.

The 1st forwarding entry in the BIFT will forward a multicast packet with BitPosition 18' to D.

The 2nd forwarding entry in the BIFT will forward a multicast packet with BitPosition 12' to F.

The 3rd forwarding entry in the BIFT will forward a multicast packet with BitPosition 10' to H.

The 4-th forwarding entry in the BIFT will forward a multicast packet with BitPosition 3' to B.



Adjacency BP (SI:BitString)	Action	BFR-NBR (Next Hop)
18' (9:00100)	fw-connected	D
12' (8:00010)	fw-connected	F
10' (7:10000)	fw-connected	H
3' (6:00100)	fw-connected	B

Figure 5: BIER-TE BIFT on BFR C

The BIER-TE BIFT on BFR D (i.e. node D) is shown in Figure 6.

The 1st forwarding entry in the BIFT will forward a multicast packet with BitPosition 17' to C.

The 2nd forwarding entry in the BIFT will forward a multicast packet with BitPosition 15' to G.

The 3rd forwarding entry in the BIFT will locally decapsulate a multicast packet with BitPosition 1 and pass a copy of the payload of the packet to the packet's NextProto.

Adjacency BP (SI:BitString)	Action	BFR-NBR (Next Hop)
17' (9:00010)	fw-connected	C
15' (8:10000)	fw-connected	G
1 (0:00001)	local-decap	

Figure 6: BIER-TE BIFT on BFR D

### 5.3. Extended BIER-TE BIFT on a BFR

Each of the BFRs that are neighbors of egress nodes (i.e., BFRs) in a BIER-TE sub-domain/topology has an extended BIER-TE BIFT to support protection against the failure of every its neighbor egress node.

For example, the extended BIER-TE BIFT on BFR C is illustrated in Figure 7. It comprises a backup entry for each of its neighbor



egress nodes D, F and H. Each of these backup entries contains a flag EPA and a backup path to a backup egress node. EPA is set to zero in normal operations. EPA in the backup entry for neighbor egress node X is set to one when egress node X fails.

The backup entry of the 1st forwarding entry in the BIFT contains EPA = 0 and backup path {10', 4}. When egress node D fails, the EPA is set to one and the backup entry is used to forward a multicast packet with BitPosition 1 for D to D's backup egress node H with BitPosition 4 along the backup path.

The backup entry of the 2nd forwarding entry in the BIFT contains EPA = 0 and backup path {3', 2', 3}. When egress node F fails, EPA is set to one and the backup entry is used to forward a multicast packet with BitPosition 2 for F to F's backup egress node E with BitPosition 3 along the backup path.

Adjacency BP (SI:BitString)	Action	BFR-NBR (Next Hop)	Backup Entry (EP, BackupPath)
18' (9:00100)	fw-connected	D	EPA=0, {10', 4}
12' (8:00010)	fw-connected	F	EPA=0, {3', 2', 3}
10' (7:10000)	fw-connected	H	EPA=0, {18', 1}
3' (6:00100)	fw-connected	B	EPA=0, { }

Figure 7: Extended BIER-TE BIFT on BFR C

The backup entry of the 3rd forwarding entry in the BIFT contains EPA = 0 and backup path {18', 1}. When egress node H fails, EPA is set to one and the backup entry is used to forward a multicast packet with BitPosition 4 for H to H's backup egress node D with BitPosition 1 along the backup path.

#### 5.4. Forwarding using Extended BIER-TE BIFT

Suppose that there is a multicast packet with explicit path {7', 4', 18', 12', 2, 1} on BFR A. The path is encoded in the BitPositions of the packet. BFR A forwards the packet to BFR B according to the forwarding entry for 7' in its extended BIER-TE BIFT. BFR B forwards the packet to BFR C according to the forwarding entry for 4' in B's extended BIER-TE BIFT.



In normal operations, after receiving the packet from BFR B, BFR C copies and sends the packet to BFR D and BFR F using the forwarding entries for 18' and 12' in the extended BIER-TE BIFT on BFR C respectively.

Once BFR C detects the failure of egress node D, it sets EPA of the backup entry in the 1st forwarding entry to one. After receiving the packet from BFR B, BFR C copies and sends the packet to D's backup egress node H using the backup entry in the forwarding entry for 18' with BFR-NBR D in C's extended BIER-TE BIFT. It copies and sends the packet to F using the forwarding entry for 12' in C's extended BIER-TE BIFT.

The packet received by BFR C from BFR B contains (SI:BitString) = (0:00011)(8:00010)(9:00100), which represents {18', 12', 2, 1}. Since EPA in the backup entry in the forwarding entry with BFR-NBR == D is 1, BFR C copies and sends the packet to D's backup egress node H in the following steps.

At first, it obtains the backup entry from the forwarding entry for 18' with BFR-NBR D. EPA == 1 in the backup entry indicates that egress protection for egress node D is active. BFR C clears BitPositions 18' and 1 in Packet's BitString and adds the backup path {10', 4} into Packet's BitString. The updated BitString in Packet is (0:01010)(7:10000)(8:00010), which represents {12', 10', 4, 2}. This lets BFR C copy and send Packet to F and H using the forwarding entries for 12' and 10' in C's extended BIER-TE BIFT respectively.

When node H receives the packet with BitPosition 4 for H, it decapsulates the packet and passes a copy of the payload of the packet to the packet's NextProto, which sends the payload to the same CE as egress node D sends.

When node F receives the packet with BitPosition 2 for F, it decapsulates the packet and passes a copy of the payload of the packet to the packet's NextProto, which sends the payload to another CE (not shown in the figure).

## 6. Security Considerations

TBD.

## 7. IANA Considerations

No requirements for IANA.



## 8. Acknowledgements

The authors would like to thank people for their comments to this work.

## 9. References

### 9.1. Normative References

- [I-D.ietf-bier-te-arch]  
Eckert, T., Cauchie, G., and M. Menth, "Tree Engineering for Bit Index Explicit Replication (BIER-TE)", draft-ietf-bier-te-arch-09 (work in progress), October 2020.
- [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>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <<https://www.rfc-editor.org/info/rfc5226>>.
- [RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, DOI 10.17487/RFC5250, July 2008, <<https://www.rfc-editor.org/info/rfc5250>>.
- [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>>.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <<https://www.rfc-editor.org/info/rfc5714>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.



- [RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, DOI 10.17487/RFC7490, April 2015, <<https://www.rfc-editor.org/info/rfc7490>>.
- [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>>.
- [RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <<https://www.rfc-editor.org/info/rfc7770>>.
- [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>>.
- [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>>.

## 9.2. Informative References

- [I-D.eckert-bier-te-frr]  
Eckert, T., Cauchie, G., Braun, W., and M. Menth,  
"Protection Methods for BIER-TE", draft-eckert-bier-te-frr-03 (work in progress), March 2018.
- [I-D.ietf-rtgwg-segment-routing-ti-lfa]  
Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B.,  
and D. Voyer, "Topology Independent Fast Reroute using  
Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-05 (work in progress), November 2020.
- [I-D.ietf-spring-segment-protection-sr-te-paths]  
Hegde, S., Bowers, C., Litkowski, S., Xu, X., and F. Xu,  
"Segment Protection for SR-TE Paths", draft-ietf-spring-segment-protection-sr-te-paths-00 (work in progress),  
September 2020.



- [RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <<https://www.rfc-editor.org/info/rfc7431>>.
- [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>>.
- [RFC8444] Psenak, P., Ed., Kumar, N., Wijnands, IJ., Dolganow, A., Przygienda, T., Zhang, J., and S. Aldrin, "OSPFv2 Extensions for Bit Index Explicit Replication (BIER)", RFC 8444, DOI 10.17487/RFC8444, November 2018, <<https://www.rfc-editor.org/info/rfc8444>>.

## Authors' Addresses

Huaimo Chen  
Futurewei  
Boston, MA  
USA

Email: [Huaimo.chen@futurewei.com](mailto:Huaimo.chen@futurewei.com)

Mike McBride  
Futurewei

Email: [michael.mcbride@futurewei.com](mailto:michael.mcbride@futurewei.com)

Aijun Wang  
China Telecom  
Beiqijia Town, Changping District  
Beijing, 102209  
China

Email: [wangaj3@chinatelecom.cn](mailto:wangaj3@chinatelecom.cn)



Gyan S. Mishra  
Verizon Inc.  
13101 Columbia Pike  
Silver Spring MD 20904  
USA

Phone: 301 502-1347  
Email: gyan.s.mishra@verizon.com

Yisong Liu  
China Mobile

Email: liuyisong@chinamobile.com

Yanhe Fan  
Casa Systems  
USA

Email: yfan@casa-systems.com

Lei Liu  
Fujitsu

USA

Email: liulei.kddi@gmail.com

Xufeng Liu  
Volta Networks

McLean, VA  
USA

Email: xufeng.liu.ietf@gmail.com



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 25, 2021

H. Chen  
M. McBride  
Futurewei  
Y. Liu  
China Mobile  
A. Wang  
China Telecom  
G. Mishra  
Verizon Inc.  
Y. Fan  
Casa Systems  
L. Liu  
Fujitsu  
X. Liu  
Volta Networks  
February 21, 2021

BIER-TE Fast ReRoute  
draft-chen-bier-te-frr-00

Abstract

This document describes a mechanism for fast re-route (FRR) protection against the failure of a transit node or link on an explicit point to multipoint (P2MP) multicast path/tree in a "Bit Index Explicit Replication" (BIER) Traffic Engineering (TE) domain. It does not have any per-path state in the core. For a multicast packet to traverse a transit node along an explicit P2MP path, when the node fails, its upstream hop node as a PLR reroutes the packet around the failed node along the P2MP path once it detects the failure.

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] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 August 25, 2021.

#### 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
1.1. Terminology . . . . .	3
2. Overview of BIER-TE FRR . . . . .	4
3. BIER-TE Extensions for BIER-TE FRR . . . . .	5
3.1. Extensions to BIER-TE BIFT . . . . .	5
3.2. Updated Forwarding Procedure . . . . .	6
4. Example Application of BIER-TE FRR . . . . .	7
4.1. Example BIER-TE Topology . . . . .	7
4.2. BIER-TE BIFT on a BFR . . . . .	8
4.3. Extended BIER-TE BIFT on a BFR . . . . .	10
4.4. Forwarding using Extended BIER-TE BIFT . . . . .	13
4.4.1. Forwarding in Normal Operations . . . . .	13
4.4.2. Forwarding in Failure . . . . .	13
5. Security Considerations . . . . .	14
6. IANA Considerations . . . . .	14
7. Acknowledgements . . . . .	14
8. References . . . . .	14
8.1. Normative References . . . . .	14
8.2. Informative References . . . . .	16
Authors' Addresses . . . . .	16



## 1. Introduction

[I-D.ietf-bier-te-arch] introduces Bit Index Explicit Replication (BIER) Traffic/Tree Engineering (BIER-TE). It is an architecture for per-packet stateless explicit point to multipoint (P2MP) multicast path/tree and based on Bit Index Explicit Replication (BIER) architecture defined in [RFC8279]. It does not require intermediate nodes to maintain any per-path/tree state.

[I-D.eckert-bier-te-frr] describes three BIER-TE FRR methods for providing fast protections against the failure of an intermediate node or link on an explicit P2MP BIER-TE path. The first method is Point-to-Point Tunneling (PPT), where a BIER-TE packet is rerouted by the PLR around the failure to its NHs and NNHs through unicast tunnels. This method depends on the tunnels, whose configurations may increase the Opex. The second is BIER-in-BIER Encapsulation (BBE), where a BIER-TE packet is rerouted by the PLR to its NHs and NNHs through encapsulating the packet in another BIER-TE header. This additional header reroutes the packet around the failure towards its NHs and NNHs and may increase the overhead. The third is Header Modification (HM), where the backup path is added into the existing BIER-TE header through using an AddBitmask and a ResetBitmask. The issue of this method is that it may cause duplicated packets for some destinations.

This document describes a BIER-TE FRR mechanism without the above issues. For a multicast packet with a BIER-TE header to traverse a transit node along an explicit P2MP path, when the node fails, its upstream hop node as a point of local repair (PLR) reroutes the packet around the failed node to the next hop nodes of the failed node on the P2MP path once it detects the failure.

This BIER-TE FRR does not require intermediate nodes to maintain any per-path state for FRR protection against the failure of a transit node or link on any explicit P2MP multicast path.

### 1.1. Terminology

BIER: Bit Index Explicit Replication.

BIER-TE: BIER Traffic/Tree Engineering.

BFR: Bit-Forwarding Router.

BFIR: Bit-Forwarding Ingress Router.

BFER: Bit-Forwarding Egress Router.



BFR-id: BFR Identifier. It is a number in the range [1,65535].

BFR-NBR: BFR Neighbor.

F-BM: Forwarding Bit Mask.

BFR-prefix: An IP address (either IPv4 or IPv6) of a BFR.

BIRT: Bit Index Routing Table. It is a table that maps from the BFR-id (in a particular sub-domain) of a BFER to the BFR-prefix of that BFER, and to the BFR-NBR on the path to that BFER.

BIFT: Bit Index Forwarding Table.

FRR: Fast Re-Route.

PLR: Point of Local Repair.

IGP: Interior Gateway Protocol.

LSDB: Link State DataBase.

SPF: Shortest Path First.

SPT: Shortest Path Tree.

## 2. Overview of BIER-TE FRR

A Bit-Forwarding Router (BFR) in a BIER-TE domain has a BIER-TE Bit Index Forwarding Tables (BIFT) [I-D.ietf-bier-te-arch]. A BIER-TE BIFT on a BFR comprises a forwarding entry for a BitPosition (BP) assigned to each of the adjacencies of the BFR. If the BP represents a forward connected adjacency, the forwarding entry for the BP forwards the multicast packet with the BP to the directly connected BFR neighbor of the adjacency. If the BP represents a BFER (i.e., egress node) or say a local decap adjacency, the forwarding entry for the BP decapsulates the multicast packet with the BP and passes a copy of the payload of the packet to the packet's NextProto within the BFR.

To support BIER-TE FRR (i.e., fast re-route (FRR) protection against the failure of a transit node or link on an explicit P2MP multicast path in a BIER-TE domain), the BIER-TE BIFT on a BFR is extended. For each forwarding entry of the BIER-TE BIFT on the BFR, if it is for the BP representing a forward connected adjacency, the forwarding entry is extended to include a new forwarding entry, which is called FRR forwarding entry or FRR entry for short.



Suppose that the BFR-NBR in the forwarding entry for the BP is N. The FRR entry forwards the multicast packet with the BP to the N's next hops that are on the P2MP path encoded in the multicast packet.

Once the BFR as a PLR detects the failure of its BFR-NBR N that is a transit node, for a multicast packet with the BP attached to N, the PLR uses the FRR forwarding entry in the extended BIER-TE BIFT to send the packet to the N's next hop nodes that are on the P2MP path encoded in the multicast packet. These next hop nodes forward the packet along the P2MP path towards the egress nodes of the path.

Before sending the packet to the N's next hops, for any local decap BP for a destination/BFER in the header, the PLR removes/clears it if it is on the backup path and it is not reachable through the forward connected adjacency BPs in the header (i.e., it is not on any branch from the PLR).

### 3. BIER-TE Extensions for BIER-TE FRR

This section describes extensions to a BIER-TE BIFT of a BFR for supporting BIER-TE FRR and enhancements on a forwarding procedure to use the extended BIER-TE BIFT for BIER-TE FRR.

#### 3.1. Extensions to BIER-TE BIFT

Every BFR has an extended BIER-TE BIFT to support BIER-TE FRR protection against the failure of its neighbor transit node. The forwarding entry with transit node (say N) as its BFR-NBR in the BIFT comprises a FRR forwarding entry (or FRR entry for short). The FRR entry contains a flag FPA (which is short for FRR Protection is Active) and a backup path from the BFR to each of N's next hop nodes.

In normal operations, the flag FPA in the FRR entry for neighbor transit node N is set to 0 (zero). The flag FPA is set to 1 (one) when transit node N fails. FPA == 1 means that the FRR protection for transit node N is active and the FRR entry will be used to forward the packet with the BP for the adjacency from the BFR to node N towards N's next hop nodes on the P2MP path encoded in the packet's BitString along the backup paths.

The backup path from the BFR to a N's next hop node X is a path that satisfies a set of constraints and does not traverse transit node N or any link connected to N. In one implementation, the backup path is represented by the BitPositions for the adjacencies along the backup path.



### 3.2. Updated Forwarding Procedure

The forwarding procedure defined in [I-D.ietf-bier-te-arch] is updated/enhanced for using an extended BIER-TE BIFT to support BIER-TE FRR.

For a multicast packet with the BP in the BitString indicating a BFR-NBR as a transit node of the P2MP path encoded in the packet, the updated forwarding procedure on a BFR sends the packet towards the transit node's next hop nodes on the P2MP path if the transit node fails.

It checks whether FPA equals to 1 (one) in the forwarding entry with the BFR-NBR that is a transit node of the P2MP path. If FPA is 1 (i.e., the transit node fails and the FRR protection for the transit node is active), the procedure clears the BP for the adjacency to the transit node in the packet's BitString first. Secondly, for any local decap BP for a destination/BFER in the BitString, it removes/clears the BP if the BP is on the backup path and is not reachable by the forward connected adjacency BPs in the BitString (i.e., is not on any branch from the BFR as PLR). And then, for each next hop node of the failed transit node that is on the P2MP path encoded in the packet's BitString, it copies and sends the packet to the next hop node along the backup path from the BFR to the next hop node.

For each next hop node of transit node BFR-NBR (which is named as N for simplicity), when N's next hop node is on the P2MP path, the forwarding procedure clears the BP for the adjacency from N to the N's next hop node in the packet's BitString and adds the BPs for the backup path from the BFR to the N's next hop node. This lets the packet be copied and sent to the N's next hop nodes along the backup paths when transit node N fails and then towards the destinations along the P2MP path.

The updated procedure is described in Figure 1. It can also be used by the BFR to forward multicast packets in normal operations.



```

Packet = the packet received by BFR;
FOR each BP k (from the rightmost in Packet's BitString) {
  IF (BP k is local decap adjacency) {
    copies Packet, sends the copy to the multicast
    flow overlay and clears bit k in Packet's BitString
  } ELSE IF (BP k is forward connect adjacency of the BFR) {
    finds the forwarding entry in the BIER-TE BIFT for the domain
    using BP k;
    Clears BP k in Packet's BitString;
    IF (FPA == 1) { //FRR for BFR-NBR/transit N is Active
      FOR each BP j for a BFER in Packet's BitString {
        IF (BP j is on a backup path and
            is not reachable by BPs in BitString) {
          Clears BP j in Packet's BitString
        }
      }
      FOR each N's next hop on P2MP path in Packet's BitString {
        Clears the BP for the adjacency from N to the next hop;
        Adds the BPs for the backup path to N's next hop
        into Packet's BitString
      }
    } //Adjacency to N removed, backup path to N's next hop added
  } ELSE {
    Copies Packet, updates the copy's BitString by
    clearing all the BPs for the adjacencies of the BFR,
    and sends the updated copy to BFR-NBR
  }
}

```

Figure 1: Updated Forwarding Procedure

#### 4. Example Application of BIER-TE FRR

This section illustrates an example application of BIER-TE FRR on a BFR in a BIER-TE topology in Figure 2.

##### 4.1. Example BIER-TE Topology

An example BIER-TE topology for a BIER-TE domain is shown in Figure 2. It has 9 nodes/BFRs A, B, C, D, E, F, G, H and I. Nodes/BFRs D, F, E, H and A are BFERs and have local decap adjacency BitPositions 1, 2, 3, 4, and 5 respectively. For simplicity, these BPs are represented by (SI:BitString), where SI = 0 and BitString is of 8 bits. BPs 1, 2, 3, 4, and 5 are represented by 1 (0:00000001), 2 (0:00000010), 3 (0:00000100), 4 (0:00001000) and 5 (0:00010000) respectively.



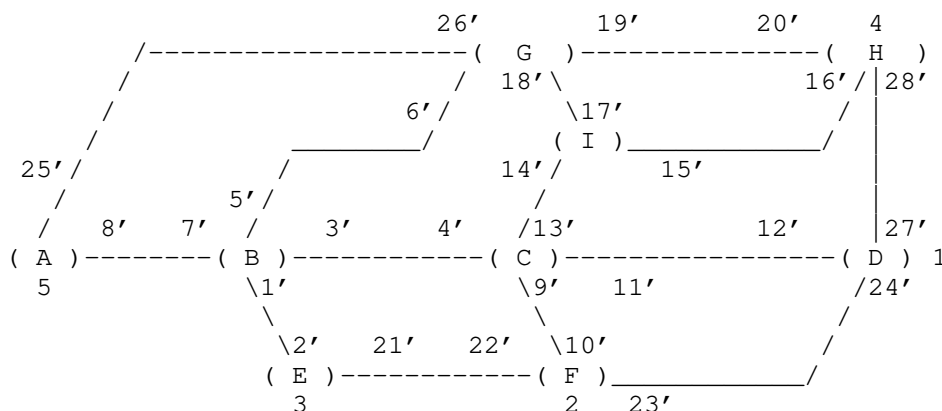


Figure 2: Example BIER-TE Topology

The BitPositions for the forward connected adjacencies are represented by  $i'$ , where  $i$  is from 1 to 28. In one option, they are encoded as  $(n+i)$ , where  $n$  is a power of 2 such as 32768. For simplicity, these BitPositions are represented by  $(SI:BitString)$ , where  $SI = (6 + (i-1)/8)$  and  $BitString$  is of 8 bits. BitPositions  $i'$  ( $i$  from 1 to 28) are represented by  $1'$  (6:00000001),  $2'$  (6:00000010),  $3'$  (6:00000100),  $4'$  (6:00001000),  $5'$  (6:00010000),  $6'$  (6:00100000),  $7'$  (6:01000000),  $8'$  (6:10000000),  $9'$  (7:00000001),  $10'$  (7:00000010), . . . ,  $24'$  (8:10000000),  $25'$  (9:00000001),  $26'$  (9:00000010), . . . ,  $28'$  (9:00001000).

For a link between two nodes X and Y, there are two BitPositions for two forward connected adjacencies. These two forward connected adjacency BitPositions are assigned on nodes X and Y respectively. The BitPosition assigned on X is the forward connected adjacency of Y. The BitPosition assigned on Y is the forward connected adjacency of X.

For example, for the link between nodes B and C in the figure, two forward connected adjacency BitPositions  $3'$  and  $4'$  are assigned to two ends of the link. BitPosition  $3'$  is assigned on node B to B's end of the link. It is the forward connected adjacency of node C. BitPosition  $4'$  is assigned on node C to C's end of the link. It is the forward connected adjacency of node B.

#### 4.2. BIER-TE BIFT on a BFR

Every BFR in a BIER-TE domain/topology has a BIER-TE BIFT. For the BIER-TE topology in Figure 2, each of 9 nodes/BFRs A, B, C, D, E, F, G, H and I has its BIER-TE BIFT for the topology.



The BIER-TE BIFT on BFR B (i.e. node B) is shown in Figure 3.

The 1st forwarding entry in the BIFT is for BitPosition 2', which is the forward connected adjacency from B to E. For a multicast packet with BitPosition 2', which indicates that the P2MP path in the packet traverses the adjacency from B to E, the forwarding entry forwards the packet to E along the link from B to E.

The 2nd forwarding entry in the BIFT is for BitPosition 4', which is the forward connected adjacency from B to C. For a multicast packet with BitPosition 4', which indicates that the P2MP path in the packet traverses the adjacency from B to C, the forwarding entry forwards the packet to C along the link from B to C.

The 3rd forwarding entry in the BIFT is for BitPosition 6', which is the forward connected adjacency from B to G. For a multicast packet with BitPosition 6', which indicates that the P2MP path in the packet traverses the adjacency from B to G, the forwarding entry forwards the packet to G along the link from B to G.

The 4-th forwarding entry in the BIFT is for BitPosition 8', which is the forward connected adjacency from B to A. For a multicast packet with BitPosition 8', which indicates that the P2MP path in the packet traverses the adjacency from B to A, the forwarding entry forwards the packet to A along the link from B to A.

Adjacency BP (SI:BitString)	Action	BFR-NBR (Next Hop)
2' (6:00000010)	fw-connected	E
4' (6:00001000)	fw-connected	C
6' (6:00100000)	fw-connected	G
8' (6:10000000)	fw-connected	A

Figure 3: BIER-TE BIFT on BFR B

The BIER-TE BIFT on BFR E (i.e. node E) is shown in Figure 4.

The 1st forwarding entry in the BIFT forwards a multicast packet with BitPosition 1' to B. It is for BitPosition 1', which is the forward connected adjacency from E to B. For a multicast packet with BitPosition 1', which indicates that the P2MP path in the packet



traverses the adjacency from E to B, the forwarding entry forwards the packet to B along the link from E to B.

The 2nd forwarding entry in the BIFT forwards a multicast packet with BitPosition 22' to F. It is for BitPosition 22', which is the forward connected adjacency from E to F. For a multicast packet with BitPosition 22', which indicates that the P2MP path in the packet traverses the adjacency from E to F, the forwarding entry forwards the packet to F along the link from E to F.

The 3rd forwarding entry in the BIFT locally decapsulates a multicast packet with BitPosition 3 and passes a copy of the payload of the packet to the packet's NextProto. It is for BitPosition 3, which is the local decap adjacency for BFER (i.e., egress) E. For a multicast packet with BitPosition 3, which indicates that the P2MP path in the packet has node E as one of its destinations (i.e., egress nodes), the forwarding entry decapsulates the packet and passes a copy of the payload of the packet to the packet's NextProto within node E.

Adjacency BP (SI:BitString)	Action	BFR-NBR (Next Hop)
1' (6:00000001)	fw-connected	B
22' (8:00001000)	fw-connected	F
3 (0:00000100)	local-decap	

Figure 4: BIER-TE BIFT on BFR D

#### 4.3. Extended BIER-TE BIFT on a BFR

Every BFR has an extended BIER-TE BIFT to support BIER-TE FRR protection against the failure of its neighbor transit node.

For example, the extended BIER-TE BIFT on BFR B is illustrated in Figure 5. Each forwarding entry with transit node (such as E, C and G) as its BFR-NBR in the BIFT comprises a FRR entry. Each of these FRR entries contains a flag FPA and a number of backup paths. For a forwarding entry with transit node X, its FRR entry has a backup path to each of node X's next hop nodes except for BFR B itself. FPA is set to zero in normal operations. FPA in the FRR entry for neighbor transit node X is set to one when node X fails.

On BFR B, the 1st forwarding entry in the BIFT has BFR-NBR E as transit node. Nodes F and B are the next hop nodes of node E in



Figure 2. The backup path from B to F without E or links attached to E goes through the link from B to C and then the link from C to F. This backup path from B to F is represented by B-->F: {4', 10'} in the FRR entry of the forwarding entry.

FPA in the FRR entry is set to 0 (zero) in normal operations. When transit node E fails, the FPA is set to 1 (one) and the FRR entry is used to forward a multicast packet with BitPosition 2' for adjacency from B to E towards E's next hop node F along the backup path from B to F if the P2MP path in the packet traverses node F.

Adjacency BP (SI:BitString)	Action	BFR-NBR (Next Hop)	FRR Entry (FPA, BackupPaths)
2' (6:00000010)	fw-connected	E	FPA=0, B-->F: {4', 10'}
4' (6:00000010)	fw-connected	C	FPA=0, B-->F: {2', 22'}, B-->D: {6', 20', 27'}, B-->I: {6', 17'}
6' (6:00100000)	fw-connected	G	FPA=0, B-->I: {4', 14'}, B-->H: {4', 14', 16'}, B-->A: {8'}
8' (6:10000000)	fw-connected	A	FPA=0, B-->G: {6'}

Figure 5: Extended BIER-TE BIFT on BFR B

On BFR B, the 2nd forwarding entry in the BIFT has BFR-NBR C as transit node. Nodes F, D, I and B are the next hop nodes of node C in Figure 2. The backup path from B to F without C or links attached to C goes through the link from B to E and then the link from E to F. This backup path from B to F is represented by B-->F: {2', 22'} in the FRR entry of the forwarding entry.

The backup path from B to D without C or links attached to C goes through the link from B to G, the link from G to H and then the link from H to D. This backup path from B to D is represented by B-->D: {6', 20', 27'} in the FRR entry of the forwarding entry.

The backup path from B to I without C or links attached to C goes through the link from B to G and then the link from G to I. This



backup path from B to I is represented by B-->I: {6', 17'} in the FRR entry of the forwarding entry.

FPA in the FRR entry is set to 0 (zero) in normal operations. When transit node C fails, the FPA is set to 1 (one) and the FRR entry is used to forward a multicast packet with BitPosition 4' for adjacency from B to C towards C's next hop nodes F, D or I along the backup paths from B to F, D or I respectively if the P2MP path in the packet traverses nodes F, D or I.

On BFR B, the 3rd forwarding entry in the BIFT has BFR-NBR G as transit node. Nodes I, H, A and B are the next hop nodes of node G in Figure 2. The backup path from B to I without G or links attached to G goes through the link from B to C and then the link from C to I. This backup path from B to I is represented by B-->I: {4', 14'} in the FRR entry of the forwarding entry.

The backup path from B to H without G or links attached to G goes through the link from B to C, the link from C to I and then the link from I to H. This backup path from B to H is represented by B-->H: {4', 14', 16'} in the FRR entry of the forwarding entry.

The backup path from B to A without G or links attached to G goes through the link from B to A. This backup path from B to A is represented by B-->A: {8'} in the FRR entry of the forwarding entry.

FPA in the FRR entry is set to 0 (zero) in normal operations. When transit node G fails, the FPA is set to 1 (one) and the FRR entry is used to forward a multicast packet with BitPosition 6' for adjacency from B to G towards G's next hop nodes I, H or A along the backup paths from B to I, H or A respectively if the P2MP path in the packet traverses nodes I, H or A.

On BFR B, the 4-th forwarding entry in the BIFT has BFR-NBR A as transit node. Nodes G and B are the next hop nodes of node A. The backup path from B to G without A or links attached to A goes through the link from B to G. This backup path from B to G is represented by B-->G: {6'} in the FRR entry of the forwarding entry.

FPA in the FRR entry is set to 0 (zero) in normal operations. When node A fails, the FPA is set to 1 (one) and the FRR entry is used to forward a multicast packet with BitPosition 8' for adjacency from B to A towards A's next hop node G along the backup path from B to G if the P2MP path in the packet traverses node A as a transit node.



#### 4.4. Forwarding using Extended BIER-TE BIFT

Suppose that there is an explicit multicast P2MP path from ingress A to egresses H and D, traversing from A to G to H and from A to B to C to D. This path is represented by BPs as {26', 20', 7', 4', 12', 4, 1}. The forwarding behaviors in normal operations and in case of BFR C failure are described below.

##### 4.4.1. Forwarding in Normal Operations

For a multicast packet with the path on BFR A, A sends the packet to G and B according to the forwarding entries for 26' and 7' in A's extended BIER-TE BIFT respectively. The packet received by G and B contains path {20', 4', 12', 4, 1}.

After receiving the packet from A, G forwards the packet to H according to forwarding entry for 20' in its extended BIER-TE BIFT. The packet received by H contains path {4', 12', 4, 1}. After receiving the packet from A, B forwards the packet to C according to forwarding entry for 4' in its extended BIER-TE BIFT. The packet received by C contains path {20', 12', 4, 1}.

After receiving the packet from G, H decapsulates the packet and passes a copy of the payload of the packet to the packet's NextProto according to forwarding entry for 4 in its extended BIER-TE BIFT. After receiving the packet from B, C forwards the packet to D according to forwarding entry for 12' in its extended BIER-TE BIFT. The packet received by D contains path {20', 4, 1}.

After receiving the packet from C, D decapsulates the packet and passes a copy of the payload of the packet to the packet's NextProto according to forwarding entry for 1 in its extended BIER-TE BIFT.

##### 4.4.2. Forwarding in Failure

Once BFR B detects the failure of node C, it sets FPA of the FRR entry in the 2nd forwarding entry with BFR-NBR C to one. After receiving the packet from BFR A, which contains path {20', 4', 12', 4, 1}, BFR B clears BP 4'; BFR B clears the BP 4 for BFER H since H is on the backup path from B to G to H to D, but not on any branch of the path from B.

For C's next hop node D on the P2MP path, BFR B clears adjacency BP 12' from C to D and adds the BPs for backup path {6', 20', 27'} from B to G to H to D. The packet has path {6', 20', 27', 1}. BFR B sends the packet to G according to forwarding entry for 6' in its extended BIER-TE BIFT. The packet received by G contains path {20', 27', 1}.



After receiving the packet from B, G forwards the packet to H according to forwarding entry for 20' in its extended BIER-TE BIFT. The packet received by H contains path {27', 1}.

After receiving the packet from G, H forwards the packet to D according to forwarding entry for 27' in its extended BIER-TE BIFT. The packet received by D contains path {1}.

After receiving the packet from H, D decapsulates the packet and passes a copy of the payload of the packet to the packet's NextProto according to forwarding entry for 1 in its extended BIER-TE BIFT.

## 5. Security Considerations

TBD.

## 6. IANA Considerations

No requirements for IANA.

## 7. Acknowledgements

The authors would like to thank Daniel Merling for his comments to this work.

## 8. References

### 8.1. Normative References

- [I-D.ietf-bier-te-arch] Eckert, T., Cauchie, G., and M. Menth, "Tree Engineering for Bit Index Explicit Replication (BIER-TE)", draft-ietf-bier-te-arch-09 (work in progress), October 2020.
- [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>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <<https://www.rfc-editor.org/info/rfc5226>>.
- [RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, DOI 10.17487/RFC5250, July 2008, <<https://www.rfc-editor.org/info/rfc5250>>.



- [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>>.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <<https://www.rfc-editor.org/info/rfc5714>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, DOI 10.17487/RFC7490, April 2015, <<https://www.rfc-editor.org/info/rfc7490>>.
- [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>>.
- [RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <<https://www.rfc-editor.org/info/rfc7770>>.
- [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>>.
- [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>>.



## 8.2. Informative References

- [I-D.eckert-bier-te-frr]  
Eckert, T., Cauchie, G., Braun, W., and M. Menth,  
"Protection Methods for BIER-TE", draft-eckert-bier-te-  
frr-03 (work in progress), March 2018.
- [I-D.ietf-rtgwg-segment-routing-ti-lfa]  
Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B.,  
and D. Voyer, "Topology Independent Fast Reroute using  
Segment Routing", draft-ietf-rtgwg-segment-routing-ti-  
lfa-05 (work in progress), November 2020.
- [I-D.ietf-spring-segment-protection-sr-te-paths]  
Hegde, S., Bowers, C., Litkowski, S., Xu, X., and F. Xu,  
"Segment Protection for SR-TE Paths", draft-ietf-spring-  
segment-protection-sr-te-paths-00 (work in progress),  
September 2020.
- [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>>.
- [RFC8444] Psenak, P., Ed., Kumar, N., Wijnands, IJ., Dolganow, A.,  
Przygienda, T., Zhang, J., and S. Aldrin, "OSPFv2  
Extensions for Bit Index Explicit Replication (BIER)",  
RFC 8444, DOI 10.17487/RFC8444, November 2018,  
<<https://www.rfc-editor.org/info/rfc8444>>.

## Authors' Addresses

Huaimo Chen  
Futurewei  
Boston, MA  
USA

Email: [Huaimo.chen@futurewei.com](mailto:Huaimo.chen@futurewei.com)



Mike McBride  
Futurewei

Email: michael.mcbride@futurewei.com

Yisong Liu  
China Mobile

Email: liuyisong@chinamobile.com

Aijun Wang  
China Telecom  
Beiqijia Town, Changping District  
Beijing, 102209  
China

Email: wangaj3@chinatelecom.cn

Gyan S. Mishra  
Verizon Inc.  
13101 Columbia Pike  
Silver Spring MD 20904  
USA

Phone: 301 502-1347  
Email: gyan.s.mishra@verizon.com

Yanhe Fan  
Casa Systems  
USA

Email: yfan@casa-systems.com

Lei Liu  
Fujitsu

USA

Email: liulei.kddi@gmail.com



Xufeng Liu  
Volta Networks

McLean, VA  
USA

Email: xufeng.liu.ietf@gmail.com



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 26, 2021

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

Z. Zhang  
ZTE Corporation  
M. Stenberg  
February 22, 2021

BIER Ingress Multicast Flow Overlay using Multicast Listener Discovery  
Protocols  
draft-ietf-bier-mld-05

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 August 26, 2021.

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 . . . . .	2
2. Terminology . . . . .	3
3. Overview . . . . .	4
4. Applicability Statement . . . . .	4
5. Querier and Listener Specifications . . . . .	4
5.1. Configuration Parameters . . . . .	5
5.2. MLDv2 instances. . . . .	6
5.2.1. Sending Queries . . . . .	6
5.2.2. Sending Reports . . . . .	6
5.2.3. Receiving Queries . . . . .	7
5.2.4. Receiving Reports . . . . .	8
5.3. Packet Forwarding . . . . .	8
6. BIER MLD/IGMP Extension Type . . . . .	8
7. Security Considerations . . . . .	9
8. IANA Considerations . . . . .	10
9. Acknowledgements . . . . .	10
10. References . . . . .	10
10.1. Normative References . . . . .	10
10.2. Informative References . . . . .	11
Appendix A. BIER Use Case in Data Centers . . . . .	12
A.1. Convention and Terminology . . . . .	14
A.2. BIER in data centers . . . . .	14
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.

### 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:

- o Toward all BMLD Nodes (i.e., providing to the BIER Layer the BFR-ids of all BMLD Nodes).
- o Without the IPv6 router alert option [RFC2711] in the hop-by-hop extension header [RFC8200] (or the IPv4 router alert option [RFC2113] for IPv4).
- o With the IP destination address set to the 'all BMLD Nodes' group address.
- o 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.
- o 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.
- o 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:

- o Toward all BMLD Queriers (i.e., providing to the BIER layer the BFR-ids of all BMLD Queriers).



- o Without the IPv6 router alert option [RFC2711] in the hop-by-hop extension header [RFC8200] (or the IPv4 router alert option [RFC2113] for IPv4).
- o With the IP destination address set to the 'all BMLD Queriers' group address.
- o With a deterministic IP source address. It is RECOMMENDED that the address is a BFR-Prefix of the sender.
- o 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.
- o 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.



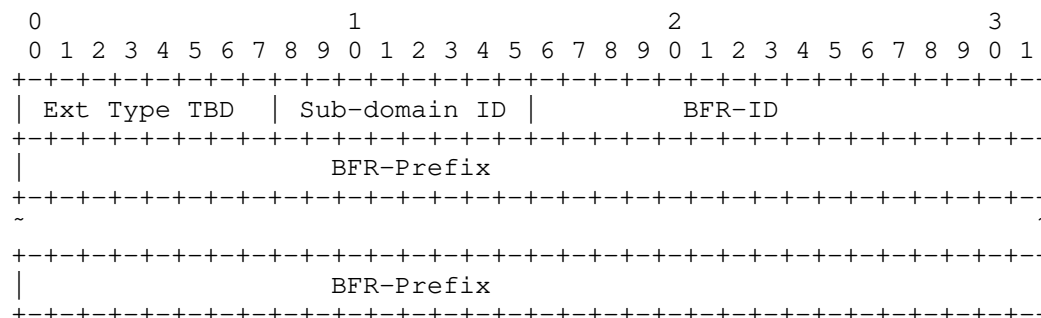


Figure 1: MLD/IGMP Extension Type for BIER

- o Ext Type: Assigned by IANA, identifying this BIER extension.
- o 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.
- o BFR-id: A two-octet field containing the BFR-id, in the specified sub-domain, of the router originating the message.
- o 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:

- o Redirect undesired traffic toward the spoofed router by subscribing to undesired multicast traffic.
- o 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

- [I-D.ietf-pim-igmp-mld-extension]  
Sivakumar, M., Venaas, S., Zhang, Z., and H. Asaeda,  
"IGMPv3/MLDv2 Message Extension", draft-ietf-pim-igmp-mld-  
extension-04 (work in progress), January 2021.
- [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>>.

## 10.2. Informative References

- [I-D.ietf-bier-mtud] Venaas, S., Wijnands, I., Ginsberg, L., and M. Sivakumar, "BIER MTU Discovery", draft-ietf-bier-mtud-00 (work in progress), February 2019.
- [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-09 (work in progress), November 2020.
- [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>>.

#### 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.



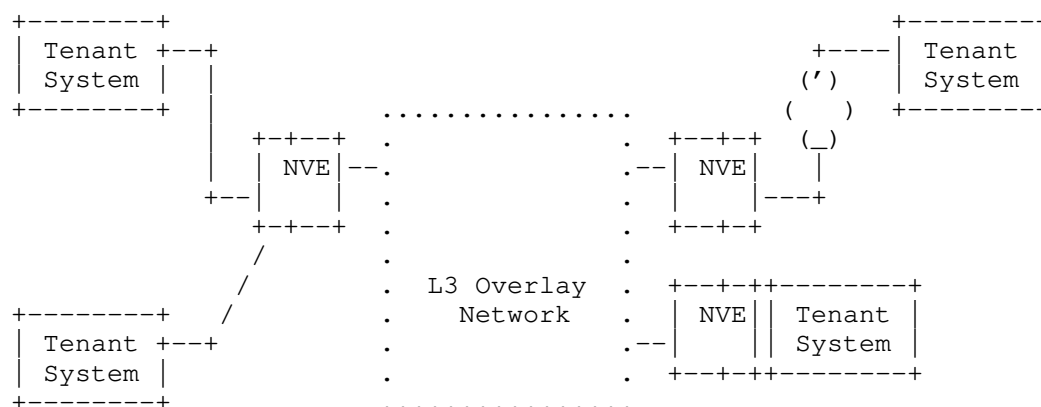


Figure 2: NVO3 Architecture

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  
Diegem 1831  
Belgium

Email: ice@cisco.com



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

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  
Helsinki 00930  
Finland

Email: markus.stenberg@iki.fi



Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: 16 May 2022

H. Bidgoli, Ed.  
J. Kotalwar  
Nokia  
I. Wijnands  
M. Mishra  
Cisco System  
Z. Zhang  
Juniper Networks  
E. Leyton  
Verizon  
12 November 2021

M-LDP Signaling Through BIER Core  
draft-ietf-bier-mlbp-signaling-over-bier-01

Abstract

Consider an end to end Multipoint LDP (mLDP) network, where it is desirable to deploy BIER in portion of this network. It might be desirable to deploy BIER with minimum disruption to the mLDP network or redesign of the network.

This document describes the procedure needed for mLDP tunnels to be signaled over and stitched through a BIER core, allowing LDP routers to run traditional mLDP 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). 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 16 May 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 . . . . .	2
2. Conventions used in this document . . . . .	2
2.1. Definitions . . . . .	3
3. mLDP Signaling Through BIER domain . . . . .	4
3.1. Ingress BBR procedure . . . . .	5
3.1.1. Automatic tLDP session Creation . . . . .	5
3.1.2. ECMP Method on IBBR . . . . .	5
3.2. Egress BBR procedure . . . . .	5
3.2.1. IBBR procedure for arriving upstream assigned label . . . . .	6
4. Datapath Forwarding . . . . .	6
4.1. Datapath traffic flow . . . . .	6
5. Recursive FEC . . . . .	7
6. IANA Consideration . . . . .	7
7. Security Considerations . . . . .	7
8. Acknowledgments . . . . .	7
9. References . . . . .	7
9.1. Normative References . . . . .	7
9.2. Informative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

Some operators that are using mLDP P2MP LSPs for their multicast transport would like to deploy BIER technology in some segment of their network. This draft explains a method to signal mLDP services through a BIER domain, with minimal disruption and operational impact to the mLDP 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 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.

**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 via Targeted LDP.

**EBBR:**

Egress BIER Boundary Router. The egress router in BIER domain from signaling point of view. It terminates the targeted ldp signaling through BIER domain. It also keeps track of all IBBRs that are part of this p2mp tree



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, all BFERs and BFIRs need to be assigned a BFR-id.

### 3. mLDP Signaling Through BIER domain

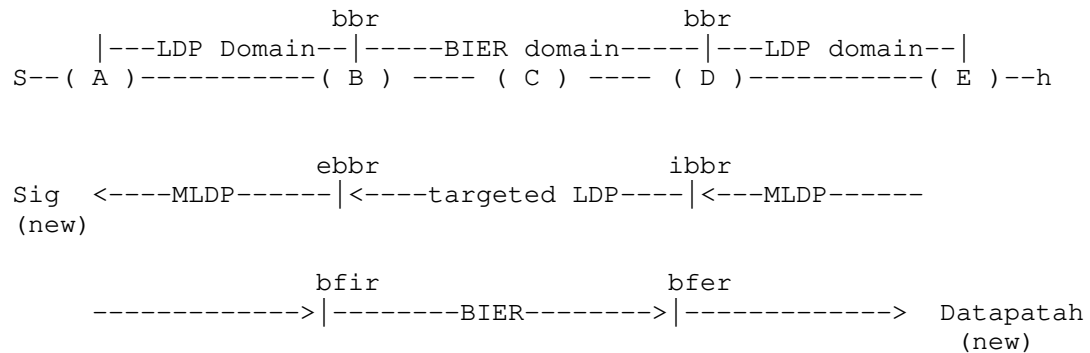


Figure 1: BIER boundry router

As per figure 1, point-to-multipoint and multipoint-to-multipoint LSPs established via mLDP [RFC6388] can be signaled through a bier domain via Targeted LDP sessions. This procedure is explained in [RFC7060] (Using LDP Multipoint Extension on Targeted LDP Sessions).

This documents provides details and defines some needed procedures.

.



### 3.1. Ingress BBR procedure

The Ingress BBR (IBBR) is connected to the mLDP domain on downstream and a bier domain on the upstream. To connect the LDP domains via BIER domain, IBBR needs to establish a targeted LDP session with EBBR closest to the root of the P2MP or MP2MP LSP. To do so IBBR will follow procedures in [RFC7060] in particular the section "6. targeted mLDP with Multicast Tunneling".

The target LDP session can be established manually via configuration or via automated mechanism.

#### 3.1.1. Automatic tLDP session Creation

tLDP session can be signaled automatically from every IBBR to the appropriate EBBR. When mLDP FEC arrives to IBBR from LDP domain, IBBR can automatically start a tLDP Session to the EBBR closest to the Root node. Both IBBR and EBBR should be in auto-discovery mode and react to the arriving tLDP Signaling packets (i.e. targeted hellos, keep-alives etc...) to establish the session automatically.

The Root node address in the mLDP FEC can be used to find the EBBR. To identify the EBBR same procedures as [RFC7060] section 2.1 can be used or the procedures as explained in the [draft-ietf-bier-pim-signaling] appendix A.

#### 3.1.2. ECMP Method on IBBR

If IBBR finds multiple equal cost EBBRs on the path to the Root, it can use a vendor specific algorithm to choose between the EBBRs. These algorithms are beyond the scope of this draft. As an example the IBBR can use the smallest EBBR IP address to establish its mLDP signaling to.

### 3.2. Egress BBR procedure

The Egress BBR (EBBR) is connected to the upstream mLDP domain. The EBBR should accept the tLDP session generated from IBBR. It should assign a unique "upstream assigned label" for each arriving FEC generated by IBBRs.

The EBBR should follow the [RFC7060] procedures with following modifications:

- \* The label assigned by EBBR cannot be Implicit Null. This is to ensure that identity of each p2mp and/or mp2mp tunnel in BIER domain is uniquely distinguished.



- \* The label can be assigned from a domain-wide Common Block (DCB) [draft-zzhang-bess-mvpn-evpn-aggregation-label]
- \* The Interface ID TLV, as per [RFC6389] should includes a new BIER sub-domain sub- tlv (type TBD)

The EBBR will also generate a new label and FEC toward the ROOT on the LDP domain. The EBBR Should stitch this generate label with the "upstream assigned label" to complete the P2MP or MP2MP LSP.

With same token the EBBR should track all the arriving FECs and the IBBRs that are generating these FECs. EBBR will use this information to build the bier header for each set of common FEC arriving from the IBBRs.

#### 3.2.1. IBBR procedure for arriving upstream assigned label

Upon receiving the "upstream assigned label", IBBR should create its own stitching instruction between the "upstream assigned label" and the down stream signaled label.

### 4. Datapath Forwarding

#### 4.1. Datapath traffic flow

On BFIR when the MPLS label for P2MP/MP2MP LSP arrives from upstream, a lookup in ILM table is done and the label is swapped with tLDP upstream assigned label. The BFIR will note all the BFERs that are interested in specific P2MP/MP2MP LSP (as per section 3.2). BFIR will put the corresponding BIER header with bit index set for all IBBRs interested in this stream. 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 [RFC8279]

The BFERs will receive the BIER packet, will look at the protocol of BIER header (MPLS). BFER will remove the BIER header and will do a lookup in the ILM table for the upstream assigned label and perform its corresponding action.

It should be noted that these procedures are also valid if BFIR is the ILER and/or BFER is the ELER as per [RFC7060]



## 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 forwarded via tLDP and the label request.

## 6. IANA Consideration

adf

1. A new BIER sub-domain sub- tlv for the interface ID TLV to be assigned by IANA

## 7. Security Considerations

TBD

## 8. Acknowledgments

Acknowledgments Authors would like to acknowledge Jingrong Xie for his comments and help on this draft.

## 9. References

### 9.1. Normative References

- [draft-ietf-bier-pim-signaling]  
"H. Bidgoli, F. Xu, J. Kotalwar, I. Wijnands, M. Mishra, Z. Zhang", 29 July 2020.
- [draft-zzhang-bess-mvpn-evpn-aggregation-label]  
"Z. Zhang, E. Rosen, W. Lin, Z. Li, I. Wijnands, "MVPN/ EVPN Tunnel Aggregation with Common Labels"", 27 April 2018.
- [RFC6389] "R Aggarwal, JL. Le Roux, "MPLS Upstream Label Assignment for LDP"", November 2011.
- [RFC7060] "M. Napierala, E. Rosen, I. Wijnands", November 2013.
- [RFC8279] "I. Wijnands, E. Rosen, A. ADolganow, T. Prygienda, S. Aldrin", November 2017.

### 9.2. Informative References

- [RFC8401] "Ginsberg, L., Przygienda, T., Aldrin, S., and Z. Zhang, "BIER Support via ISIS"", June 2018.



- [RFC8444] "Psenak, P., Kumar, N., Wijnands, IJ., Dolganow, A., Przygienda, T., Zhang, Z., and S. Aldrin, "OSPF Extensions for Bit Index Explicit Replication", June 2018.
- [RFC8556] "Rosen, E., Ed., Sivakumar, M., Wijnands, IJ., Aldrin, S., Dolganow, A., and T. Przygienda, "Multicast VPN Using Index Explicit Replication (BIER)", April 2018.

#### Authors' Addresses

Hooman Bidgoli (editor)  
Nokia  
Ottawa  
Canada

Email: [hooman.bidgoli@nokia.com](mailto:hooman.bidgoli@nokia.com)

Jayant Kotalwar  
Nokia  
Mountain View,  
United States of America

Email: [jayant.kotalwar@nokia.com](mailto:jayant.kotalwar@nokia.com)

IJsbrand Wijnands  
Cisco System  
Diegem  
Belgium

Email: [ice@cisco.com](mailto:ice@cisco.com)

Mankamana Mishra  
Cisco System  
Milpitas,  
United States of America

Email: [mankamis@cisco.com](mailto:mankamis@cisco.com)

Zhaohui Zhang  
Juniper Networks  
Boston,  
United States of America

Email: [zzhang@juniper.com](mailto:zzhang@juniper.com)



Eddie Leyton  
Verizon

Email: [Edward.leyton@verizonwireless.com](mailto:Edward.leyton@verizonwireless.com)



BIER Working Group  
Internet-Draft  
Intended status: Informational  
Expires: May 19, 2021

G. Mirsky, Ed.  
ZTE Corp.  
N. Kumar  
Cisco Systems, Inc.  
M. Chen  
Huawei Technologies  
S. Pallagatti, Ed.  
VMware  
November 15, 2020

Operations, Administration and Maintenance (OAM) Requirements for Bit  
Index Explicit Replication (BIER) Layer  
draft-ietf-bier-oam-requirements-11

## Abstract

This document describes a list of functional requirement toward  
Operations, Administration and Maintenance (OAM) toolset in Bit Index  
Explicit Replication (BIER) layer of a network.

## 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 May 19, 2021.

## Copyright Notice

Copyright (c) 2020 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. Conventions used in this document . . . . .	2
1.1.1. Terminology . . . . .	2
1.1.2. Requirements Language . . . . .	3
2. Requirements . . . . .	3
3. IANA Considerations . . . . .	4
4. Security Considerations . . . . .	4
5. Contributors . . . . .	4
6. References . . . . .	5
6.1. Normative References . . . . .	5
6.2. Informative References . . . . .	5
Authors' Addresses . . . . .	6

## 1. Introduction

[RFC8279] introduces and explains Bit Index Explicit Replication (BIER) architecture and how it supports forwarding of multicast data packets.

This document lists the OAM requirements for BIER layer of the multicast domain. The list can further be used to for gap analysis of available OAM tools to identify possible enhancements of existing or whether new OAM tools are required to support proactive and on-demand path monitoring and service validation.

### 1.1. Conventions used in this document

#### 1.1.1. Terminology

The term "BIER OAM" used in this document interchangeably with longer version "set of OAM protocols, methods, and tools for BIER layer".

BFR: Bit-Forwarding Router

BFER: Bit-Forwarding Egress Router

BIER: Bit Index Explicit Replication

OAM: Operations, Administration and Maintenance



### 1.1.2. Requirements Language

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. Requirements

This section lists requirements for OAM of BIER layer:

1. The listed requirements MUST be supported with any type of transport layer over which BIER layer can be realized.
2. It MUST be possible to initialize BIER OAM session from any Bit-Forwarding Router (BFR) of the given BIER domain.
3. It SHOULD be possible to initialize BIER OAM session from a centralized controller.
4. BIER OAM MUST support proactive and on-demand OAM monitoring and measurement methods.
5. BIER OAM MUST support unidirectional OAM methods, both continuity check and performance measurement.
6. BIER OAM packets MUST be in-band, i.e., follow exactly the same path as data plane traffic, in the forward direction, i.e., from ingress toward egress endpoint(s) of the OAM test session.
7. BIER OAM MUST support bi-directional OAM methods. Such OAM methods MAY combine in-band monitoring or measurement in the forward direction and out-of-band notification in the reverse direction, i.e., from egress to ingress end point of the OAM test session.
8. BIER OAM MUST support proactive monitoring of BFER availability by a BFR in the given BIER domain, e.g., p2mp BFD active tail support.
9. BIER OAM MUST support Path Maximum Transmission Unit discovery.
10. BIER OAM MUST support Reverse Defect Indication (RDI) notification of the source of continuity checking BFR by Bit-Forwarding Egress Routers (BFERs), e.g., by using Diag in p2mp BFD with active tail support.



11. BIER OAM MUST support active and passive performance measurement methods.
12. BIER OAM MUST support unidirectional performance measurement methods to calculate throughput, loss, delay and delay variation metrics. [RFC6374] provides great details for performance measurement and performance metrics.
13. BIER OAM MUST support defect notification mechanism, like Alarm Indication Signal. Any BFR in the given BIER domain MAY originate a defect notification addressed to any subset of BFRs within the domain.
14. BIER OAM MUST support methods to enable survivability of a BIER layer. These recovery methods MAY use protection switching and restoration.

### 3. IANA Considerations

This document does not propose any IANA consideration. This section may be removed.

### 4. Security Considerations

This document list the OAM requirement for BIER-enabled domain and does not raise any security concerns or issues in addition to ones common to networking.

### 5. Contributors



Erik Nordmark

Email: nordmark@acm.org

Sam Aldrin  
Google

Email: aldrin.ietf@gmail.com

Lianshu Zheng  
Huawei Technologies

Email: vero.zheng@huawei.com

Nobo Akiya  
Big Switch Networks

Email: nobo.akiya.dev@gmail.com

## 6. References

### 6.1. Normative References

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

### 6.2. Informative References

- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <<https://www.rfc-editor.org/info/rfc6374>>.
- [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>>.



Authors' Addresses

Greg Mirsky (editor)  
ZTE Corp.

Email: gregimirsky@gmail.com

Nagendra Kumar  
Cisco Systems, Inc.

Email: naikumar@cisco.com

Mach Chen  
Huawei Technologies

Email: mach.chen@huawei.com

Santosh Pallagatti (editor)  
VMware

Email: santosh.pallagatti@gmail.com



BIER Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: 7 October 2022

G. Mirsky  
Ericsson  
T. Przygienda  
Juniper Networks  
A. Dolganow  
Individual contributor  
5 April 2022

Path Maximum Transmission Unit Discovery (PMTUD) for Bit Index Explicit  
Replication (BIER) Layer  
draft-ietf-bier-path-mtu-discovery-12

Abstract

This document describes Path Maximum Transmission Unit Discovery (PMTUD) in Bit Indexed Explicit Replication (BIER) layer.

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 7 October 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
1.1. Conventions used in this document . . . . .	2
1.1.1. Terminology . . . . .	2
1.1.2. Requirements Language . . . . .	3
2. Problem Statement . . . . .	3
3. PMTUD Mechanism for BIER . . . . .	4
3.1. Data TLV for BIER Ping . . . . .	6
4. IANA Considerations . . . . .	6
5. Security Considerations . . . . .	7
6. Acknowledgment . . . . .	7
7. References . . . . .	7
7.1. Normative References . . . . .	7
7.2. Informative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

In packet switched networks, when a host seeks to transmit data to a target destination, the data is transmitted as a set of packets. In many cases, it is more efficient to use the largest size packets that are less than or equal to the least Maximum Transmission Unit (MTU) for any forwarding device along the routed path to the IP destination for these packets. Such "least MTU" is known as Path MTU (PMTU). Fragmentation or packet drop, silent or not, may occur on hops along the route where an MTU is smaller than the size of the datagram. To avoid any of the listed above behaviors, the packet source must find the value of the least MTU, i.e., PMTU, that will be encountered along the route that a set of packets will follow to reach the given set of destinations. Such MTU determination along a specific path is referred to as path MTU discovery (PMTUD).

[RFC8279] introduces and explains Bit Index Explicit Replication (BIER) architecture and how it supports the forwarding of multicast data packets. [I-D.ietf-bier-ping] introduced BIER Ping as a transport-independent OAM mechanism to detect and localize failures in the BIER data plane. This document specifies how BIER Ping can be used to perform efficient PMTUD in the BIER domain.

## 1.1. Conventions used in this document

## 1.1.1. Terminology

This document uses terminology defined in [RFC8279]. Familiarity with this specification and the terminology used is expected.



### 1.1.2. Requirements Language

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. Problem Statement

[I-D.ietf-bier-oam-requirements] sets forth the requirement to define PMTUD protocol for BIER domain. This document describes the extension to [I-D.ietf-bier-ping] for use in the BIER PMTUD solution.

Current PMTUD mechanisms ([RFC1191], [RFC8201], and [RFC4821]) are primarily targeted to work on point-to-point, i.e. unicast paths. These mechanisms use packet fragmentation control by disabling fragmentation of the probe packet. As a result, a transient node that cannot forward a probe packet that is bigger than its link MTU sends to the packet source an error notification, otherwise the packet destination may respond with a positive acknowledgment. Thus, possibly through a series of iterations, varying the size of the probe packet, the packet source discovers the PMTU of the particular path.

Applying such existing PMTUD solutions are inefficient for point-to-multipoint paths constructed for multicast traffic. Probe packets must be flooded through the whole set of multicast distribution paths over and over again until the very last egress responds with a positive acknowledgment. Consider the multicast network presented in Figure 1, where MTU on all links but one (B, D) is the same. If MTU on the link (B, D) is smaller than the MTU on the other links, using existing PMTUD mechanism probes will unnecessarily flood to leaf nodes E, F, and G for the second and consecutive times and positive responses will be generated and received by root A repeatedly.



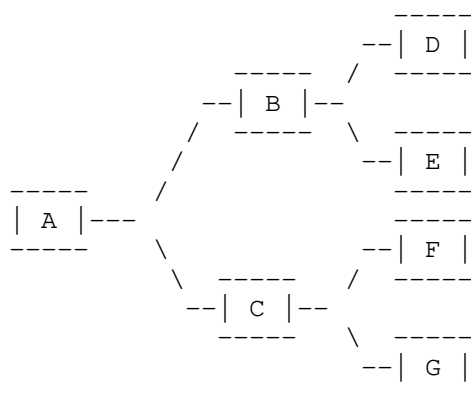


Figure 1: Multicast network

### 3. PMTUD Mechanism for BIER

A BFIR selects a set of BFERs for the specific multicast distribution. Such a BFIR determines, by explicitly controlling a subset of targeted BFERs and transmitting a series of probe packets, the MTU of that multicast distribution tree. In the case of ECMP, BFIR MAY test each path by varying the value in the Entropy field. The critical step is that in case of failure at an intermediate BFR to forward towards the subset of targeted downstream BFERs, the BFR responds with a partial (compared to the one it received in the request) bitmask towards the originating BFIR in error notification. That allows for retransmission of the next probe with a smaller MTU address only towards the failed downstream BFERs instead of all BFERs addressed in the previous probe. In the scenario discussed in Section 2 the second and all following (if needed) probes will be sent only to the node D since MTU discovery of E, F, and G has been completed already by the first probe successfully.

Consider the network displayed in Figure 1 to be a presentation of a BIER domain and all nodes to be BFRs. To discover MTU over BIER domain to BFERs D, F, E, and G BFIR A will use BIER Ping with Data TLV, defined in Section 3.1. Size of the first probe set to  $M_{max}$  determined as minimal MTU value of BFIR's links to BIER domain. As has been assumed in Section 2, MTUs of all links but the link (B, D) are the same. Thus BFERs E, F, and G would receive BIER Echo Request and will send their respective replies to BFIR A. BFR B may pass the packet which is too large to forward over egress link (B, D) to the appropriate network layer for error processing where it would be recognized as a BIER Echo Request packet. BFR B MUST send BIER Echo Reply to BFIR A and MUST include Downstream Mapping TLV, defined in [I-D.ietf-bier-ping] setting its fields in the following fashion:



- \* MTU SHOULD be set to the minimal MTU value among all egress BIER links, logical links between this and downstream BFRs, that could be used to reach B's downstream BFRs;
- \* Address Type MAY be set to any value defined in Section 3.3.4 [I-D.ietf-bier-ping].
- \* I flag MUST be cleared to direct the responding BFR not to include the Incoming SI-BitString TLV in the BIER Echo Response.
- \* Downstream Interface Address field MUST be zeroed.
- \* List of Sub-TLVs MUST include the Egress Bitstring sub-TLV with the list of all BFRs that cannot be reached because the egress MTU turned out to be too small.

The BFIR will receive either of the two types of packets:

- \* a positive Echo Reply from one of BFRs to which the probe has been sent. In this case, the bit corresponding to the BFER MUST be cleared from the bitmask string (BMS);
- \* a negative Echo Reply with bit string listing unreachable BFRs and recommended MTU value MTU'. The BFIR MUST add the bit string to its BMS and set the size of the next probe as min(MTU, MTU')

If a negative Echo Reply is received, the BFIR MUST wait for the expiration of the Echo Request before transmitting the updated Echo Request. If upon expiration of the Echo Request timer BFIR didn't receive any Echo Replies, then the size of the probe SHOULD be decreased. There are scenarios when an implementation of the PMTUD would not decrease the size of the probe. For example, suppose upon expiration of the Echo Request timer BFIR didn't receive any Echo Reply. In that case, BFIR MAY continue to retransmit the probe using the initial size and MAY apply probe delay retransmission procedures. The algorithm used to delay retransmission procedures on BFIR is outside the scope of this specification. The BFIR sends probes using BMS and locally defined retransmission procedures, but not more frequently than after the Echo Request timer expired, until either the bit string is clear, i.e., contains no set bits, or until the BFIR retransmission procedure terminates and PMTU discovery is declared unsuccessful. In the case of convergence of the procedure, the size of the last probe indicates the PMTU size that can be used for all BFRs in the initial BMS without incurring fragmentation.

Thus we conclude that in order to comply with the requirement in [I-D.ietf-bier-oam-requirements]:



- \* a BFR SHOULD support PMTUD;
- \* a BFR MAY use defined per BIER sub-domain MTU value as initial MTU value for discovery or use it as MTU for this BIER sub-domain to reach BFRs;
- \* a BFIR MUST have a locally defined PMTUD probe retransmission procedure.

3.1. Data TLV for BIER Ping

There needs to be a control for probe size in order to support the BIER PMTUD. Data TLV format is presented in Figure 2.

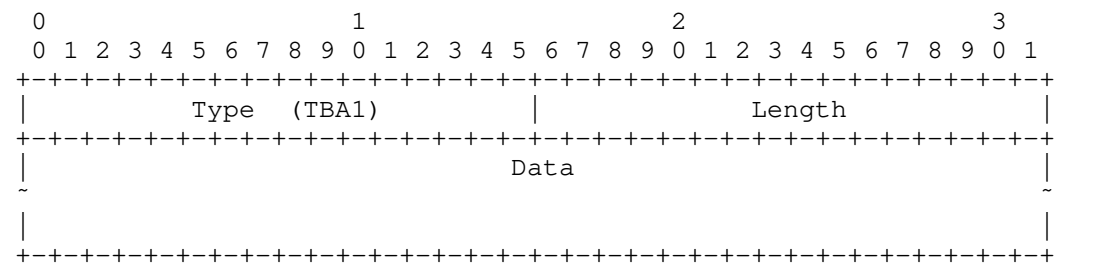


Figure 2: Data TLV format

- \* Type: indicates Data TLV, to be allocated by IANA Section 4.
- \* Length: the length of the Data field in octets.
- \* Data: n octets (n = Length) of arbitrary data. The receiver SHOULD ignore it.

4. IANA Considerations

IANA is requested to assign a new Type value for Data TLV Type from its registry of TLV and sub-TLV Types of BIER Ping as follows:

Value	Description	Reference
TBA1	Data	This document

Table 1: Data TLV Type



## 5. Security Considerations

Routers that support PMTUD based on this document are subject to the same security considerations as defined in [I-D.ietf-bier-ping]

## 6. Acknowledgment

Authors greatly appreciate thorough review and the most detailed comments by Eric Gray.

## 7. References

### 7.1. Normative References

[I-D.ietf-bier-ping]

Kumar, N., Pignataro, C., Akiya, N., Zheng, L., Chen, M., and G. Mirsky, "BIER Ping and Trace", Work in Progress, Internet-Draft, draft-ietf-bier-ping-07, 11 May 2020, <<https://datatracker.ietf.org/doc/html/draft-ietf-bier-ping-07>>.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <<https://www.rfc-editor.org/info/rfc1191>>.

[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>>.

[RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <<https://www.rfc-editor.org/info/rfc4821>>.

[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>>.

[RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <<https://www.rfc-editor.org/info/rfc8201>>.

### 7.2. Informative References

[I-D.ietf-bier-oam-requirements]

Mirsky, G., Kumar, N., Chen, M., and S. Pallagatti, "Operations, Administration and Maintenance (OAM)



Requirements for Bit Index Explicit Replication (BIER) Layer", Work in Progress, Internet-Draft, draft-ietf-bier-oam-requirements-11, 15 November 2020, <<https://datatracker.ietf.org/doc/html/draft-ietf-bier-oam-requirements-11>>.

[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>>.

#### Authors' Addresses

Greg Mirsky  
Ericsson  
Email: [gregimirsky@gmail.com](mailto:gregimirsky@gmail.com)

Tony Przygienda  
Juniper Networks  
Email: [prz@juniper.net](mailto:prz@juniper.net)

Andrew Dolganow  
Individual contributor  
Email: [adolgano@gmail.com](mailto:adolgano@gmail.com)



BIER Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: 2 October 2022

G. Mirsky  
Ericsson  
L. Zheng  
Individual Contributor  
M. Chen  
G. Fioccola  
Huawei Technologies  
31 March 2022

Performance Measurement (PM) with Marking Method in Bit Index Explicit  
Replication (BIER) Layer  
draft-ietf-bier-pmmm-oam-12

#### Abstract

This document describes the applicability of a hybrid performance measurement method for packet loss and packet delay measurements of a multicast service through a Bit Index Explicit Replication 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 2 October 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. Conventions used in this document . . . . .	3
2.1. Terminology . . . . .	3
2.2. Requirements Language . . . . .	3
3. OAM Field in BIER Header . . . . .	3
4. Theory of Operation . . . . .	4
4.1. Single-Marking Enabled Measurement . . . . .	5
4.2. Double-Marking Enabled Measurement . . . . .	6
4.3. Operational Considerations . . . . .	7
5. IANA Considerations . . . . .	7
6. Security Considerations . . . . .	7
7. Acknowledgement . . . . .	8
8. References . . . . .	8
8.1. Normative References . . . . .	8
8.2. Informative References . . . . .	9
Authors' Addresses . . . . .	9

## 1. Introduction

[RFC8279] introduces and explains the Bit Index Explicit Replication (BIER) architecture and how it supports the forwarding of multicast data packets. [RFC8296] specified that in the case of BIER encapsulation in an MPLS network, a BIER-MPLS label, the label that is at the bottom of the label stack, uniquely identifies the multicast flow. [I-D.fioccola-rfc8321bis] and [I-D.fioccola-rfc8889bis] describe a hybrid performance measurement method, according to the classification of measurement methods in [RFC7799]. The method, called Packet Network Performance Monitoring (PNPM), can be used to measure packet loss, latency, and jitter on live traffic complies with requirements R-5 and R-12 listed in [I-D.ietf-bier-oam-requirements]. Because this method is based on marking consecutive batches of packets, the method is often referred to as a marking method. Terms PNPM and "marking method" in this document are used interchangeably.



This document defines how the marking method can be used on the BIER layer to measure packet loss and delay metrics of a multicast flow in an MPLS network.

## 2. Conventions used in this document

### 2.1. Terminology

This document uses the terms related to the Alternate Marking Method as defined in [I-D.fioccola-rfc8321bis], [I-D.fioccola-rfc8889bis]. This document uses the terms related to the Bit Indexed Explicit Replication as defined in [RFC8296].

### 2.2. Requirements Language

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. OAM Field in BIER Header

[RFC8296] defined the two-bits long field, referred to as OAM. The OAM field can be used for the marking performance measurement method. Because the setting of the field to any value does not affect forwarding and/or quality of service treatment of a packet, using the OAM field for PNPM in BIER layer can be viewed as the example of the hybrid performance measurement method.

Figure 1 displays the interpretation of the OAM field defined in this specification for the use of the PNPM method. The context of interpretation of the OAM field MAY be signaled via the control plane or configured using an extension to the BIER YANG data model [I-D.ietf-bier-bier-yang]. These extensions are outside the scope of this document.

```

      0
      0  1
+--+--+--+
| S | D |
+--+--+--+

```

Figure 1: OAM field of BIER Header format

where:



\* S - Single-Marking flag;

\* D - Double-Marking flag.

#### 4. Theory of Operation

The marking method can be used in the multicast environment supported by BIER layer. Without limiting any generality consider multicast network presented in Figure 2. Any combination of markings can be applied to a multicast flow by the Bit Forwarding Ingress Router (BFIR) at either ingress or egress point to perform node, link, segment or end-to-end measurement to detect performance degradation defect and localize it efficiently.

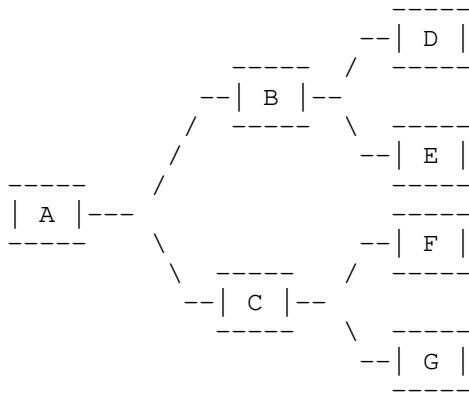


Figure 2: Multicast network



Using the marking method, a BFIR creates distinct sub-flows in the particular multicast traffic over BIER layer. Each sub-flow consists of consecutive blocks of identically marked packets. For example, a block of N packets, with each packet being marked as X, is followed by the block of M packets with each packet being marked as Y. These blocks are unambiguously recognizable by a monitoring point at any Bit Forwarding Router (BFR) and can be measured to calculate packet loss and/or packet delay metrics. The marking method can be used on multiple flows concurrently. Demultiplexing of monitored flows might be achieved using n-tuple, for example, two-tuple as combination of the values in the Entropy and BFIR-id fields [RFC8296]. Also, that can be achieved by using an explicit Flow Identifier. The definition of the Flow Identifier is outside the scope of this specification. It is expected that the marking values be set and cleared at the edge of BIER domain. Thus for the scenario presented in Figure 2 if the operator initially monitors the A-C-G and A-B-D segments he may enable measurements on segments C-F and B-E at any time.

#### 4.1. Single-Marking Enabled Measurement

As explained in [I-D.fioccola-rfc8321bis], marking can be applied to delineate blocks of packets based either on the equal number of packets in a block or based on the equal time interval. The latter method offers better control as it allows a better account for capabilities of downstream nodes to report statistics related to batches of packets and, at the same time, time resolution that affects defect detection interval.

If the Single-Marking measurement is used to measure packet loss, then the D flag MUST be set to zero on transmit and ignored by the monitoring point.

The S flag is used to create sub-flows to measure the packet loss by switching the value of the S flag every N-th packet or at certain time intervals. Delay metrics MAY be calculated with the sub-flow using any of the following methods:

- \* First/Last Packet Delay calculation: whenever the marking, i.e., the value of S flag changes, a BFR can store the timestamp of the first/last packet of the block. The timestamp can be compared with the timestamp of the packet that arrived in the same order through a monitoring point at a downstream BFR to compute packet delay. Because timestamps collected based on the order of arrival this method is sensitive to packet loss and re-ordering of packets (see Section 4.3 for more details).



- \* Average Packet Delay calculation: an average delay is calculated by considering the average arrival time of the packets within a single block. A BFR may collect timestamps for each packet received within a single block. Average of the timestamp is the sum of all the timestamps divided by the total number of packets received. Then the difference between the average packet arrival time calculated for the downstream monitoring point and the same metric but calculated at the upstream monitoring point is the average packet delay on the segment between these two points. This method is robust to out of order packets and also to packet loss on the segment between the measurement points (packet loss may cause a minor loss of accuracy in the calculated metric because the number of packets used is different at each measurement point). This method only provides a single metric for the duration of the block, and it doesn't give the minimum and maximum delay values. This limitation of producing only the single metric could be overcome by reducing the duration of the block. As a result, the calculated value of the average delay will better reflect the minimum and maximum delay values of the block's duration time.

#### 4.2. Double-Marking Enabled Measurement

Double-Marking method allows measurement of minimum and maximum delays for the monitored flow, but it requires more nodal and network resources. If the Double-Marking method used, then the S flag is used to create the sub-flow, i.e., mark blocks of packets. The D flag is used to mark single packets within a block to measure delay and jitter.

The first marking (S flag alternation) is needed for packet loss and also for average delay measurement. The second marking (D flag is put to one) creates a new set of marked packets that are fully identified over the BIER network, so that a BFR can store the timestamps of these packets; these timestamps can be compared with the timestamps of the same packets on a second BFR to compute packet delay values for each packet. The number of measurements can be easily increased by changing the frequency of the second marking. On the other hand, the higher frequency of the second marking will cause a higher volume of the measurement data being transported through the BIER domain. An operator should consider and balance both effects. This method is useful to measure not only the average delay but also the minimum and maximum delay values and, in wider terms, to know more about the statistic distribution of delay values.



#### 4.3. Operational Considerations

For the ease of operational procedures, the initial marking of a multicast flow is performed at BFIR, and cleared, by way of removing BIER encapsulation from a payload packet, at the edge of the BIER domain by BFERs.

Since at the time of writing this specification, there are no proposals to using auto-discovery or signaling mechanism to inform downstream nodes what methodology is used each monitoring point MUST be configured beforehand.

Section 5 [I-D.fioccola-rfc8321bis] provides a detailed analysis of how packet re-ordering and the duration of the block in the Single-Marking mode of the marking method impact the accuracy of the packet loss measurement. Re-ordering of packets in the Single-Marking mode will be noticeable only at the edge of a block of packets (re-ordering within the block cannot be detected in the Single-Marking mode). If the extra delay for some packets is much smaller than half of the duration of a block, then it should be easier to attribute re-ordered packets to the proper block and thus maintain the accuracy of the packet loss measurement.

Selection of a time interval to switch the marking of a batch of packets should be based on the service requirements. In the course of the regular operation, reports, including performance metrics like packet loss ratio, packet delay, and inter-packet delay variation, are logged every 15 minutes. Thus, it is reasonable to maintain the duration of the measurement interval at 5 minutes with 100 measurements per each interval. To support these measurements, marking of the packet batch is switched every 3 seconds. In case when performance metrics are required in near-real-time, the duration interval of a single batch of identically marked packets will be in the range of tens of milliseconds.

#### 5. IANA Considerations

This document sets no requirements to IANA. This section can be removed before the publication.

#### 6. Security Considerations

Regarding using the marking method, [I-D.fioccola-rfc8321bis] stressed two types of security concerns. First, the potential harm caused by the measurements, is a lesser threat as [RFC8296] defines OAM field used by the marking method so that the value of "two bits have no effect on the path taken by a BIER packet and have no effect on the quality of service applied to a BIER packet." Second security



concern, potential harm to the measurements can be mitigated by using policy, suggested in [RFC8296], to accept BIER packets only from trusted routers, not from customer-facing interfaces.

All the security considerations for BIER discussed in [RFC8296] are inherited by this document.

## 7. Acknowledgement

Comments from Alvaro Retana helped improve the document and are much appreciated.

Reviews and comments from Quan Xiong and Xiao Min are thankfully acknowledged.

## 8. References

### 8.1. Normative References

- [I-D.fioccola-rfc8321bis]  
Fioccola, G., Cociglio, M., Mirsky, G., Mizrahi, T., Zhou, T., and X. Min, "Alternate-Marking Method", Work in Progress, Internet-Draft, draft-fioccola-rfc8321bis-03, 23 February 2022, <<https://datatracker.ietf.org/doc/html/draft-fioccola-rfc8321bis-03>>.
- [I-D.fioccola-rfc8889bis]  
Fioccola, G., Cociglio, M., Sapio, A., Sisto, R., and T. Zhou, "Multipoint Alternate-Marking Method", Work in Progress, Internet-Draft, draft-fioccola-rfc8889bis-03, 23 February 2022, <<https://datatracker.ietf.org/doc/html/draft-fioccola-rfc8889bis-03>>.
- [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>>.
- [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>>.
- [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>>.



## 8.2. Informative References

- [I-D.ietf-bier-bier-yang]  
Chen, R., Hu, F., Zhang, Z., Dai, X., and M. Sivakumar,  
"YANG Data Model for BIER Protocol", Work in Progress,  
Internet-Draft, draft-ietf-bier-bier-yang-07, 8 September  
2020, <<https://datatracker.ietf.org/doc/html/draft-ietf-bier-bier-yang-07>>.
- [I-D.ietf-bier-oam-requirements]  
Mirsky, G., Kumar, N., Chen, M., and S. Pallagatti,  
"Operations, Administration and Maintenance (OAM)  
Requirements for Bit Index Explicit Replication (BIER)  
Layer", Work in Progress, Internet-Draft, draft-ietf-bier-  
oam-requirements-11, 15 November 2020,  
<[https://datatracker.ietf.org/doc/html/draft-ietf-bier-  
oam-requirements-11](https://datatracker.ietf.org/doc/html/draft-ietf-bier-oam-requirements-11)>.
- [RFC7799] Morton, A., "Active and Passive Metrics and Methods (with  
Hybrid Types In-Between)", RFC 7799, DOI 10.17487/RFC7799,  
May 2016, <<https://www.rfc-editor.org/info/rfc7799>>.
- [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>>.

## Authors' Addresses

Greg Mirsky  
Ericsson  
Email: [gregimirsky@gmail.com](mailto:gregimirsky@gmail.com)

Lianshu Zheng  
Individual Contributor  
Email: [veronique\\_zheng@hotmail.com](mailto:veronique_zheng@hotmail.com)

Mach Chen  
Huawei Technologies  
Email: [mach.chen@huawei.com](mailto:mach.chen@huawei.com)

Giuseppe Fioccola  
Huawei Technologies  
Email: [giuseppe.fioccola@huawei.com](mailto:giuseppe.fioccola@huawei.com)



BIER  
Internet-Draft  
Intended status: Standards Track  
Expires: July 8, 2021

Z. Zhang  
Juniper Networks  
N. Warnke  
Deutsche Telekom  
I. Wijnands  
Cisco Systems  
D. Awduche  
Verizon  
January 4, 2021

Tethering A BIER Router To A BIER incapable Router  
draft-ietf-bier-tether-01

Abstract

This document specifies optional procedures to optimize the handling of Bit Index Explicit Replication (BIER) incapable routers, by attaching (tethering) a BIER router to a BIER incapable router.

Requirements Language

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.

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 July 8, 2021.



## 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 . . . . .	2
2. Additional Considerations . . . . .	3
3. Egress Protection . . . . .	5
4. Specification . . . . .	6
4.1. IGP Signaling . . . . .	6
4.2. BGP Signaling . . . . .	7
5. Security Considerations . . . . .	8
6. IANA Considerations . . . . .	8
7. Contributors . . . . .	8
8. Acknowledgements . . . . .	9
9. Normative References . . . . .	9
Authors' Addresses . . . . .	10

## 1. Introduction

Consider the scenario in Figure 1 where router X does not support BIER.

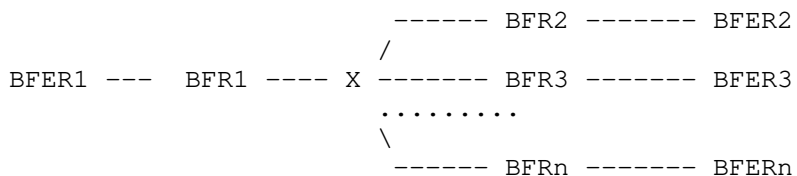


Figure 1: Deployment with BIER incapable routers

For BFR1 to forward BIER traffic towards BFR2...BFRn, it needs to tunnel individual copies through X. This degrades to "ingress" replication to those BFRs. If X's connections to BFRs are long



distance or bandwidth limited, and  $n$  is large, it becomes very inefficient.

A solution to the inefficient tunneling from BFRs is to attach (tether) a BFRx to  $X$  as depicted in Figure 2:

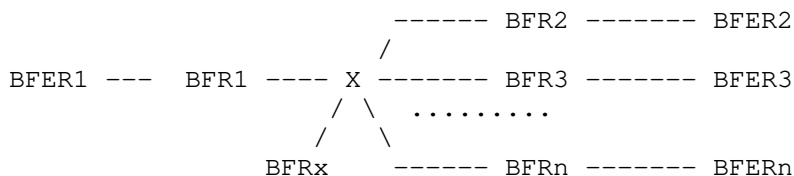


Figure 2: Tethered BFRx

Instead of BFR1 tunneling to BFR2, ..., BFRn directly, BFR1 will get BIER packets to BFRx, who will then tunnel to BFR2, ..., BFRn. There could be fat and local pipes between the tethered BFRx and  $X$ , so ingress replication from BFRx is acceptable.

For BFR1 to tunnel BIER packets to BFRx, the BFR1-BFRx tunnel need to be announced in Interior Gateway Protocol (IGP) as a forwarding adjacency so that BFRx will appear on the Shortest Path First (SPF) tree. This needs to happen in a BIER specific topology so that unicast traffic would not be tunneled to BFRx. Obviously this is operationally cumbersome.

Section 6.9 of BIER architecture specification [RFC8279] describes a method that tunnels BIER packets through incapable routers without the need to announce tunnels. However that does not work here, because BFRx will not appear on the SPF tree of BFR1.

There is a simple solution to the problem though. BFRx could advertise that it is  $X$ 's helper and other BFRs will use BFRx (instead of  $X$ 's children on the SPF tree) to replace  $X$  during its post-SPF processing as described in section 6.9 of BIER architecture specification [RFC8279].

## 2. Additional Considerations

While the example shows a local connection between BFRx and  $X$ , it does not have to be like that. As long as packets can arrive at BFRx without requiring  $X$  to do BIER forwarding, it should work.

Additionally, the helper BFRx can be a transit helper, i.e., it has other connections (instead of being a stub helper that is only



connected to X), as long as BFRx won't send BIER packets tunneled to it back towards the tunnel ingress. Figure 3 below is a simple case:

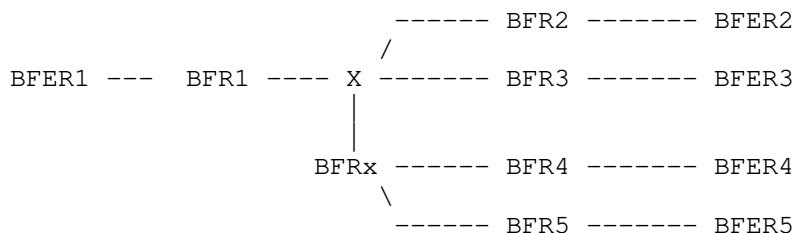


Figure 3: A Safe Transit Helper

In the example of Figure 4, there is a connection between BFR1 and BFRx. If the link metrics are all 1 on the three sides of BFR1-X-BFRx triangle, loop won't happen but if the BFRx-X metric is 3 while other two sides of the triangle has metric 1 then BFRx will send BIER packets tunneled to it from BFR1 back to BFR1, causing a loop.

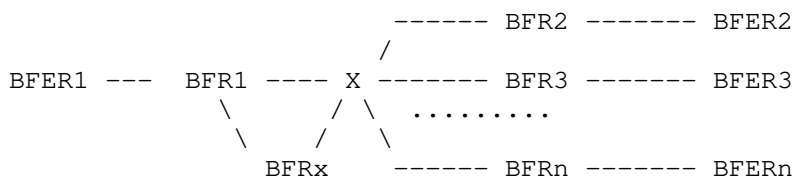


Figure 4: Potential looping situation

This can easily be prevented if BFR1 does an SPF calculation with the helper BFRx as the root. For any BFERn reached via X from BFR1, if BFRx's SPF path to BFERn includes BFR1 then BFR1 must not use the helper. Instead, BFR1 must directly tunnel packets for BFERn to X's BFR (grand-)child on BFR1's SPF path to BFERn, per section 6.9 of [RFC8279].

Notice that this SPF calculation on BFR1 with BFRx as the root is not different from the SPF done for a neighbor as part of Loop-Free Alternate (LFA) calculation. In fact, BFR1 tunneling packets to X's helper is not different from sending packets to a LFA backup.

Also notice that, instead of a dedicated helper BFRx, any one or multiple ones of BFR2..N can also be the helper (as long as the connection between that BFR and X has enough bandwidth for replication to multiple helpers through X). To allow multiple



helpers to help the same non-BFR, the "I am X's helper" advertisement carries a priority. BFR1 will choose the helper advertising the highest priority among those satisfying the loop-free condition described above. When there are multiple helpers advertising the same priority and satisfying the loop-free condition, any one or multiple ones could be used solely at the discretion of BFR1. However, if multiple ones are used, it means that multiple copies may be tunneled through X.

The situation in Figure 5 where a helper BFRxy helps two different non-BFRs X and Y also works. It's just a special situation of a transit helper.

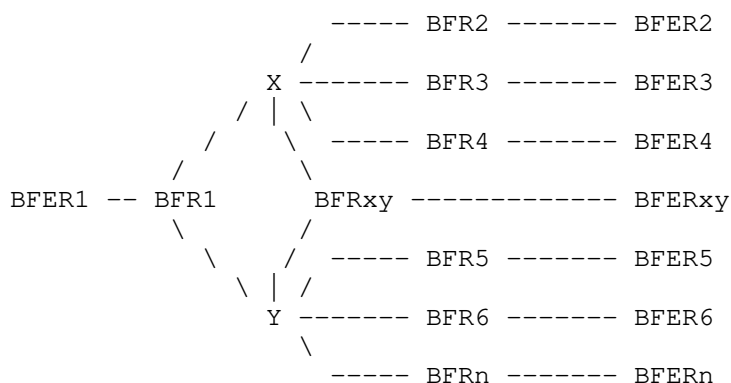


Figure 5: One Helper for multiple helped

### 3. Egress Protection

The same principal can be used for egress protection. Consider the following:

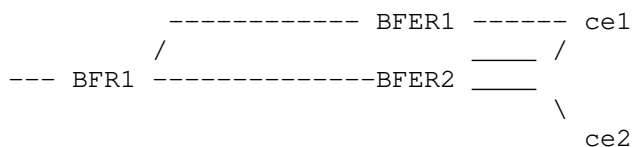


Figure 6: Egress Protection

ce1 is multi-homed to BFER1 and BFER2. Suppose both ce1 and ce2 need to receive certain multicast traffic and the copy for ce1 in normal situation follows the BFR1-BFER1-ce1 path while the copy for ce2 follows the BFR1-BFER2-ce2 path (i.e. the packet that BFR1 receives



has two bits set for BFER1 and BFER2 respectively). If BFR1 detects the node failure of BFER1, in-flight traffic with BFER1 bit set is redirected to BFER2, who will then deliver to ce1 only. Note that for the same multicast payload, BFER2 would receive two copies (before the BFIR converges), one with the BFER1 bit set and one with the BFER2 bit set. BFER2 will deliver the copy with the BFER1 bit to ce1 upon detection of node failure of BFER1, but will not deliver the same to ce2.

If ce2 is also multi-homed to BFER1 and BFER2, then BFER1 and BFER2 could egress-protect each other. Each announces that it is the helper node of the other, and the fact that each is capable of BIER indicates that it is for egress protection only.

#### 4. Specification

The procedures in this document apply when a BFRx is tethered to a BIER incapable router X as X's helper for BIER forwarding.

##### 4.1. IGP Signaling

Suppose that the BIER domain uses BIER signaling extensions to ISIS [RFC8401] or OSPF [RFC8444]. The helper node (BFRx) MUST advertise one or more BIER Helped Node sub-sub-TLVs (one for each helped node). The value is BIER prefix of the helped node (X) followed by a one-octet priority field, and one-octet reserved field. The length is 6 for IPv4 and 18 for IPv6 respectively.

The post-SPF processing procedures in Section 6.9 of the BIER architecture specification [RFC8279] are modified as following for BIER tethering purpose.

At step 2, the removed node is added to an ordered list maintained with each child that replaces the node. If the removed node already has a non-empty list maintained with itself, add the removed node to the tail of the list and copy the list to each child.

At the end, the calculating node BFR-B would use a unicast tunnel to reach next hop BFRs for some BFERs. The next hop BFR has an ordered list created at step 2 above, recording each BIER incapable node replaced by their children along the way. For a particular BFER to be reached via a tunnel to the next hop BFR, additional procedures are performed as following.

- o Starting with the first node in the ordered list of incapable nodes, say N1, check if there is one or more helper nodes for N1. If not, go the next node in the list.



- o Order all the helper nodes of N1 based descending (priority, BIER prefix). Starting with the first one, say H1, check if BFR-B could use H1 as LFA next hop to reach the BFER. If yes, H1 is used as the next hop BFR for the BFER and the procedure stops. If not, go to the next helper in order.
- o If none of the helper nodes of N1 can be used, go to the next node in the list of incapable nodes.

If the above procedure finishes without finding any helper, then the original BFR next hop via a tunnel is used to reach the BFER.

#### 4.2. BGP Signaling

Suppose that the BIER domain uses BGP signaling [I-D.ietf-bier-idr-extensions] instead of IGP. BFR1..N advertises BIER prefixes that are reachable through them, with BIER Path Attributes (BPA) attached. There are three situations regarding X's involvement:

- (1) X does not participate in BGP peering at all
- (2) X re-advertises the BIER prefixes but does not do next-hop-self
- (3) X re-advertises the BIER prefixes and does next-hop-self

With (1) and (2), the BFR1..N will tunnel BIER packets directly to each other. It works but not efficiently as explained earlier. With (3), BIER forwarding will not work, because BFR1..N would try to send BIER packets to X though X does not advertise any BIER information. If Tunnel Encapsulation Attribute (TEA) [I-D.ietf-idr-tunnel-encaps] is used as specified in [I-D.zzhang-bier-multicast-as-a-service] with (3), then it becomes similar to (2) - works but still not efficiently.

To make tethering work well with BGP signaling, the following can be done:

- o Configure a BGP session between X and its helper BFRx. X re-advertises BIER prefixes (with BPA) to BFRx without changing the tunnel destination address in the TEA.
- o BFRx advertises its own BIER prefix with BPA to X, and sets the tunnel destination address in the TEA to itself. X then re-advertises BFRx's BIER prefix to BFR1..N, without changing the tunnel destination address in the TEA.



- o For BIER prefixes (with BIER Path Attribute) that X re-advertises to other BFRs, the tunnel destination in the TEA is changed to the helper BFRx.

With the above, BFR1..N will tunnel BIER packets to BFRx (following the tunnel destination address in the TEA), who will then tunnel packets to other BFRs (again following the tunnel destination address in the TEA). Notice that what X does is not specific to BIER at all.

## 5. Security Considerations

This specification does not introduce additional security concerns beyond those already discussed in BIER architecture and OSPF/ISIS/BGP extensions for BIER signaling.

## 6. 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
----	----
TBD1	BIER Helped Node

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

Type	Name
----	----
TBD2	BIER Helped Node

## 7. Contributors

The following also contributed to this document.

Zheng(Sandy) Zhang  
ZTE Corporation

EMail: zzhang\_ietf@hotmail.com

Hooman Bidgoli  
Nokia  
EMail: hooman.bidgoli@nokia.com



## 8. Acknowledgements

The author wants to thank Eric Rosen and Antonie Przygienda for their review, comments and suggestions.

## 9. Normative References

[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-07 (work in progress), September 2019.

[I-D.ietf-idr-tunnel-encaps]

Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-20 (work in progress), November 2020.

[I-D.zzhang-bier-multicast-as-a-service]

Zhang, Z., Rosen, E., Awduche, D., and L. Geng, "Multicast/BIER As A Service", draft-zzhang-bier-multicast-as-a-service-01 (work in progress), May 2020.

[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>>.

[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>>.

[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>>.

[RFC8444] Psenak, P., Ed., Kumar, N., Wijnands, IJ., Dolganow, A., Przygienda, T., Zhang, J., and S. Aldrin, "OSPFv2 Extensions for Bit Index Explicit Replication (BIER)", RFC 8444, DOI 10.17487/RFC8444, November 2018, <<https://www.rfc-editor.org/info/rfc8444>>.



Authors' Addresses

Zhaohui Zhang  
Juniper Networks

EMail: zzhang@juniper.net

Nils Warnke  
Deutsche Telekom

EMail: Nils.Warnke@telekom.de

IJsbrand Wijnands  
Cisco Systems

EMail: ice@cisco.com

Daniel Awduche  
Verizon

EMail: daniel.awduche@verizon.com



BIER WG  
Internet-Draft  
Intended status: Informational  
Expires: August 11, 2021

Z. Zhang  
G. Mirsky  
Q. Xiong  
ZTE Corporation  
Y. Liu  
China Mobile  
February 7, 2021

BIER (Bit Index Explicit Replication) Redundant Ingress Router Failover  
draft-zhang-bier-source-protection-02

## Abstract

This document describes a failover in the Bit Index Explicit Replication domain with a redundant ingress router.

## 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 August 11, 2021.

## 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 . . . . .	2
2. Keywords . . . . .	3
3. The Redundant BFIR Failover Analysis . . . . .	3
3.1. Node Failure Monitoring . . . . .	4
3.2. Monitoring of the Working Path for a Failure . . . . .	5
4. BFD and Ping . . . . .	7
4.1. BIER Ping . . . . .	7
4.2. BIER BFD . . . . .	8
5. IANA Considerations . . . . .	9
6. Security Considerations . . . . .	9
7. References . . . . .	9
7.1. Normative References . . . . .	9
7.2. Informative References . . . . .	9
Authors' Addresses . . . . .	11

## 1. Introduction

Bit Index Explicit Replication (BIER) [RFC8279] is an architecture that provides 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).

Redundant Ingress Router Failover is not specific to the BIER environment. Redundant Ingress Router Failover means that to avoid single node failure, two or more ingress routers, BFIRs in a BIER environment, can be connected to the same multicast flow's source node. One of BFIRs is selected to forward the flow from a multicast source node to egress routers, i.e., BFER in a BIER environment. The BFERs may choose the primary BFIR for the given multicast flow based on local policies. BFERs in the same multicast group may select the same or different BFIR. The BFIR and the path in use are referred to as working, while all alternative available BFIRs and paths that can be used to receive the same multicast flow are referred to as protection.

When either the working BFIR or the working path fails, a BFER can select one of the protecting BFIRs to recover the multicast flow. The shorter the detection time, the faster the flow recovers.

This document discusses the functions that can be used to detect the failure to trigger redundant ingress router failover in the BIER environment.



## 2. Keywords

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. The Redundant BFIR Failover Analysis

According to the BIER architecture [RFC8279], BIER overlay protocols, which among others include MVPN [RFC8556], MLD [I-D.ietf-bier-mlld], PIM [I-D.ietf-bier-pim-signaling], are used to exchange the multicast flow information. Based on that, a BFER selects the UMH (Upstream Multicast Hop) BFIR as the ingress router. The BFIR selected as the UMH through a BIER overlay protocol learns of BFERs which have chosen it to receive the particular multicast flow. BIER transport is used to deliver the multicast packet to the destination BFERs. The detection of a defect in the BIER transport layer ensures that the source flow protection is uninterrupted. The switchover is performed at the BIER overlay layer. Upon detecting the failure, an update in the BIER overlay can trigger BFIR re-selection by BFERs.

As described in [I-D.szczl-mboned-redundant-ingress-failover], the root standby modes, i.e., Cold Standby, Warm Standby, and Hot Standby, can be used in the BIER environment. In Warm and Hot Standby modes, the protection BFIR needs to learn through BIER overlay protocols the identities of BFERs in the particular multicast group. In the Hot Standby mode, BFER receives duplicate flows from the selected active BFIR and protection BFIR, BFER accepts the flow packet from the selected active BFIR, identified, for example, by BFIR-id in the BIER header, discards the multicast packet from the protection BFIR.

The most important elements in the redundant BFIR failover mechanism are failure detection and switchover. Note that the scope of the failure detection includes node and working path. Similarly, BFIR switching and BFER switching are considered in the switchover scenario.

The selected BFIR is referred to as Selected BFIR (S-BFIR), and the backup BFIR is referred to as Backup BFIR (B-BFIR). For simplicity, only one B-BFIR is considered in the following use case.



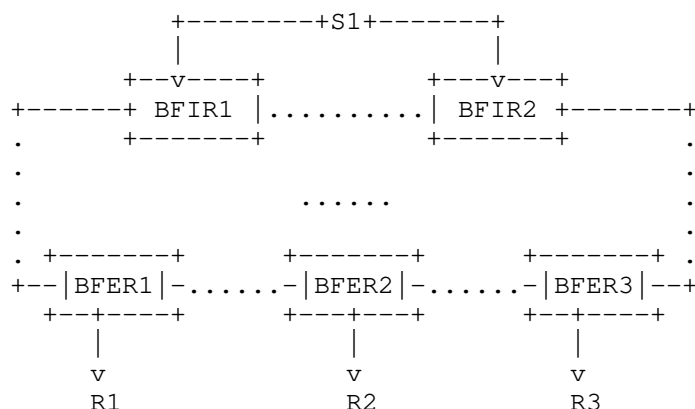


Figure 1: An Example of the Redundant BFIR Failover

In Figure 1, a multicast source S1 is connected to BFIR1 and BFIR2. In some deployments, only BFIR1 advertises S1 flow information to BFERs using a BIER overlay protocol, such as, among others, BGP (MVPN), MLD, or PIM. For this example, all BFERs that are directed to receive the S1 flow will select BFIR1 as the S-BFIR, and BFIR2 is considered as the B-BFIR. In some other deployments, BFIR1 and BFIR2 both advertise S1 flows to BFERs using a BIER overlay protocol. As a result, some BFERs may select BFIR1 as their S-BFIR, other BFERs may select BFIR2 as S-BFIR, BFIR1 and BFIR2 are responsible for different sub-groups of BFERs, and they, respectively, are the B-BFIR for the second sub-set of BFERs. We do not distinguish these two cases strictly.

There are two types of failure monitoring:

- o Node failure monitoring: It is used for BFIR failure detection. The BFER failure detection is out of the scope of this document.
- o Working path failure monitoring: It is used for BIER transport path failure detection. It is used for the monitoring among BIER domain edge routers, which include BFIR and BFER, through BIER forwarding.

### 3.1. Node Failure Monitoring

For example, consider when S1 is connected to BFIR1 and BFIR2 on a shared media segment. BFIR1 is acting as S-BFIR for the multicast flow transmitted by S1. BFIR2 can monitor BFIR1 node failure using a BFD session [RFC5880] built over the shared media segment. Also, can use ping methods, including, for example, IPv4 ping [RFC0792], IPv6



ping [RFC4443], and LSP-Ping [RFC8029] in a network with either IPv4, IPv6, or MPLS data plane, respectively.

In case there is no shared media segment interconnecting S1, BFIR1, and BFIR2, BFIR2 may monitor the state of BFIR1 using a BIER BFD session [I-D.ietf-bier-bfd] or a ping protocol across the BIER domain. A ping protocol listed above or BIER ping [I-D.ietf-bier-ping] can be used. In case there is no direct connection between BFIR1 and BFIR2, multiple hops will be traversed. Similarly, any of the listed above path continuity checking methods can be used by a BFER to monitor the path to and state of S-BFIR. The case when the S-BFIR monitors the working path to a BFER is considered further in the document in more details.

The monitoring case between S-BFIR and B-BFIR, referred to as the Warm Standby mode, is described in section 4.2 [I-D.szcl-mboned-redundant-ingress-failover]. For code and Hot Standby modes described in Sections 4.1 and 4.3 [I-D.szcl-mboned-redundant-ingress-failover], the monitoring between S-BFIR and B-BFIR may not be necessary.

For the monitoring between BFIR and BFERs, the BFIR node failure detection is also be combined with working path failure detection.

### 3.2. Monitoring of the Working Path for a Failure



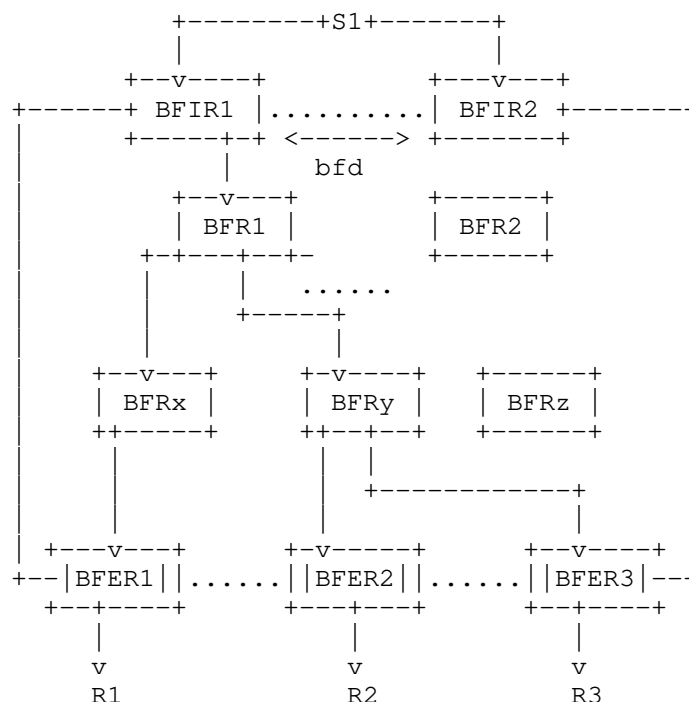


Figure 2: An Example of the Monitoring of the Working Path

In the case of a node failure detection, the path between B-BFIR and S-BFIR may not be the same as the path traversed by the data flow. For example, in Figure 2, the path from BFIR1 (S-BFIR) to all the BFERs is different from the path between BFIR1 and BFIR2 (B-BFIR). In Warm Standby mode, if the path between BFIR2 and BFIR1 is broken, BFIR2 will detect the failure and interpret that as if BFIR1 is down. As a result, BFIR2 will take on the role of S-BFIR. But the path from BFIR1 to all or some of the BFERs may be working well and is not affected by the defect between BFIR1 and BFIR2. In this situation, the B-BFIR switches to S-BFIR unnecessarily, and that causes packet duplication in the network and at BFERs.

For the failure detection between BIER edge routers, which include BFIR and BFER, the path of a test packet is steered from BFIR to BFER is the same as the path traversed by the monitored flow. In this way, the BFER simultaneously monitors S-BFIR for node and working path failure.

There are two options to monitor the working multicast distribution tree in BIER:



- o S-BFIR monitors all the BFERs;
- o BFER monitors the S-BFIR.

In the BIER transport environment, the defect detection is based on a BIER-specific mechanism, e.g., BIER Ping [I-D.ietf-bier-ping], BIER BFD [I-D.ietf-bier-bfd]. BIER BFD [I-D.ietf-bier-bfd] reduces the number of BFD sessions between S-BFIR and each of BFERs. Only one multipoint BFD session will be built among S-BFIR and all the BFERs and B-BFIR. When MVPN is used as the BIER overlay protocol, BFD Discriminator attribute, defined in Section 3.1.6 in [I-D.ietf-bess-mvpn-fast-failover], can be used to bootstrap the multipoint BFD session between a BFIR and BFERs. In this situation, only S-BFIR sends the BFD Discriminator attribute and transmits periodic BFD Control messages, BFER and B-BFIR can monitor S-BFIR, S-BFIR doesn't monitor BFER and B-BFIR.

Consider when S-BFIR monitors paths to and state of all BFERs in the particular multicast group. Once S-BFIR detects that a BFER is unreachable, S-BFIR notifies B-BFIR and the latter may start forwarding that multicast packets to that BFER. The monitoring can be achieved by a P2P BFD session between S-BFIR and each of BFERs. Alternatively, a P2MP BFD session with active tails between S-BFIR and BFERs can be used. This behavior can be used for the Warm Standby mode.

When BFER monitors S-BFIR, a B-BFIR can also monitor S-BFIR. Consider that a BFER or B-BFIR detects the failure of the S-BFIR. In the Cold Standby mode, the BFER MUST select B-BFIR as the new S-BFIR and signal to B-BFIR using a BIER overlay protocol as soon as possible. In the Hot Standby mode, the BFER MUST switch to accept and forward the multicast flow received from B-BFIR. In the Warm Standby mode, B-BFIR becomes the S-BFIR and begins to forward the flow to BFERs.

#### 4. BFD and Ping

BFD and Ping can be used in failure detection, but there are differences between them. A network administrator can select the appropriate mechanism according to the actual network.

##### 4.1. BIER Ping

[I-D.ietf-bier-ping] describes the mechanism and basic BIER Operation, Administration, and Maintenance packet format that can be used to perform failure detection and isolation on the BIER data plane without any dependency on other layers like the IP layer.



In the example of Figure 1, BFER can monitor the status of BFIR and the path status between BFER and BFIR. BFER1 sends the BIER Ping packet to BFIR1. Suppose BFER1 does not receive several consecutive responses from BFIR1 in an expected period (may be multiple of the average round-trip time). In that case, the BFER1 concludes the BFIR1 as a failed UMH, and BFER1 selects BFIR2 as the UMH. In the Cold Standby mode, BFER1 signals to BFIR2 to start receiving the multicast flow. In the Hot Standby mode, BFER begins to accept the flow from BFIR2. If B-BFIR monitors S-BFIR in the Warm Standby mode and detects the failure, B-BFIR takes the role of S-BFIR and begins to forward the flow.

In this example, BFER1, BFER2, BFER3, and B-BFIR send the BIER ping packets to BFIR1 separately. The timeout period MAY be set to different values depending on the local performance requirement on each BFER. In the Warm Standby mode, if the timeout period is different on BFER and B-BFIR, and the period on B-BFIR is longer than BFER, and multicast packets could be lost.

In the general case of a more complex BIER topology, it cannot be guaranteed that the path used from BFIR1 to BFER1 is the same as in the reverse direction, i.e., from BFER1 to BFIR1. If that is not guaranteed and the paths are not co-routed, then this method may produce false results, both false negative and false positive. The former is when ping fails while the multicast path and flow are OK. The latter is when the multicast path has a defect, but ping works. Thus, to improve the consistency of this method of detecting a failure in multicast flow transport, the path that the echo request from BFER1 traverses to BFIR1 must be co-routed with the path that the monitored multicast flow traverses through the BIER domain from BFIR1 to BFER1.

#### 4.2. BIER BFD

[I-D.ietf-bier-bfd] describes the application of P2MP BFD in a BIER network. And it describes the procedures for using such a mode of BFD protocol to verify multipoint or multicast connectivity between a sender (BFIR) and one or more receivers (BFER and a redundant BFIR).

In the same example, BFIR1 sends the BIER Echo request packet to BFERs to bootstrap a p2mp BFD session. After BFER1, BFER2 and BFER3 receive the Echo request packet with BFD Discriminator and the Target SI-Bitstring TLVs, BFERs creates the BFD session of type MultipointTail [RFC8562] to monitor the status of BFIR1 and the working path. If BFERs have not received a BFD packet from BFIR1 for the Detection Time [RFC8562], BFER1 will treat BFIR1 as a failed UMH. In the Cold Standby mode, BFER1 re-selects UMH and then signals to BFIR2. As a result, BFIR2 begins to forward the multicast flow. In



the Hot Standby mode, BFER1 switches to accept the flow from BFIR2. B-BFIR (BFIR2) monitors S-BFIR (BFIR1) in the Warm Standby mode, using the same p2mp BFD session. After B-BFIR detects the failure, it takes on the role of S-BFIR and begins to forward the multicast flow to BFERs.

## 5. IANA Considerations

This document does not have any requests for IANA allocation. This section can be deleted before the publication of the draft.

## 6. Security Considerations

Security considerations discussed in [RFC8279], [RFC8562], [I-D.ietf-bier-ping], [I-D.ietf-bess-mvpn-fast-failover] and [I-D.ietf-bier-bfd] apply to this document.

## 7. References

### 7.1. Normative References

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

### 7.2. Informative References

- [I-D.ietf-bess-mvpn-fast-failover]  
Morin, T., Kebler, R., and G. Mirsky, "Multicast VPN Fast Upstream Failover", draft-ietf-bess-mvpn-fast-failover-15 (work in progress), January 2021.
- [I-D.ietf-bier-bfd]  
Xiong, Q., Mirsky, G., hu, f., and C. Liu, "BIER BFD", draft-ietf-bier-bfd-00 (work in progress), November 2020.
- [I-D.ietf-bier-mld]  
Pfister, P., Wijnands, I., Venaas, S., Wang, C., Zhang, Z., and M. Stenberg, "BIER Ingress Multicast Flow Overlay using Multicast Listener Discovery Protocols", draft-ietf-bier-mld-04 (work in progress), March 2020.



- [I-D.ietf-bier-pim-signaling]  
Bidgoli, H., Xu, F., Kotalwar, J., Wijnands, I., Mishra, M., and Z. Zhang, "PIM Signaling Through BIER Core", draft-ietf-bier-pim-signaling-11 (work in progress), November 2020.
- [I-D.ietf-bier-ping]  
Nainar, N., Pignataro, C., Akiya, N., Zheng, L., Chen, M., and G. Mirsky, "BIER Ping and Trace", draft-ietf-bier-ping-07 (work in progress), May 2020.
- [I-D.szcl-mboned-redundant-ingress-failover]  
Shepherd, G., Zhang, Z., Liu, Y., and Y. Cheng, "Multicast Redundant Ingress Router Failover", draft-szcl-mboned-redundant-ingress-failover-00 (work in progress), October 2020.
- [RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, DOI 10.17487/RFC0792, September 1981, <<https://www.rfc-editor.org/info/rfc792>>.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <<https://www.rfc-editor.org/info/rfc4443>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <<https://www.rfc-editor.org/info/rfc8029>>.
- [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>>.
- [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>>.



[RFC8562] Katz, D., Ward, D., Pallagatti, S., Ed., and G. Mirsky,  
Ed., "Bidirectional Forwarding Detection (BFD) for  
Multipoint Networks", RFC 8562, DOI 10.17487/RFC8562,  
April 2019, <<https://www.rfc-editor.org/info/rfc8562>>.

Authors' Addresses

Zheng Zhang  
ZTE Corporation

Email: [zhang.zheng@zte.com.cn](mailto:zhang.zheng@zte.com.cn)

Greg Mirsky  
ZTE Corporation

Email: [gregimirsky@gmail.com](mailto:gregimirsky@gmail.com)

Quan Xiong  
ZTE Corporation

Email: [xiong.quan@zte.com.cn](mailto:xiong.quan@zte.com.cn)

Yisong Liu  
China Mobile

Email: [liuyisong@chinamobile.com](mailto:liuyisong@chinamobile.com)