BIER WG
Internet-Draft
Intended status: Informational
Expires: October 31, 2021

Z. Zhang
G. Mirsky
Q. Xiong
ZTE Corporation
Y. Liu
China Mobile
H. Li
China Telecom

April 29, 2021

BIER (Bit Index Explicit Replication) Redundant Ingress Router Failover
draft-ietf-bier-source-protection-00

Abstract

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

Status of This Memo

Copyright Notice

Table of Contents

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-mld], 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.szcl-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.

```
                         +--------+S1+-------+
                         |                   |
                      +--v----+          +---v---+
               +------+ BFIR1 |..........| BFIR2 +-------+
               .      +-------+          +-------+       .
               .                                         .
               .                 ......                  .
               .                                         .
               . +-------+      +-------+      +-------+ .
               +--|BFER1|-......-|BFER2|-......-|BFER3|--+
                 +--+----+      +---+---+      +--+----+
                    |               |             |
                    v               v             v
                    R1              R2            R3
```

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:

- Node failure monitoring: It is used for BFIR failure detection. The BFER failure detection is out of the scope of this document.

- 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

```
                       +--------+S1+-------+
                       |                   |
                    +--v----+          +---v---+
             +------+ BFIR1 |..........| BFIR2 +-------+
             |      +-----+-+ <------> +-------+       |
             |            |     bfd                    |
             |         +--v---+        +------+        |
             |         | BFR1 |        | BFR2 |        |
             |       +-+---+--+-       +------+        |
             |       |     |   ......                  |
             |       |     +-----+                     |
             |       |           |                     |
             |    +--v---+     +-v----+   +------+     |
             |    | BFRx |     | BFRy |   | BFRz |     |
             |    ++-----+     ++--+--+   +------+     |
             |     |            |  |                   |
             |     |            |  +------------+      |
             |     |            |               |      |
             | +---v---+      +-v-----+      +--v----+ |
             +--|BFER1||......||BFER2||......||BFER3|--+
               +--+----+      +---+---+      +--+----+
                  |               |             |
                  v               v             v
                  R1              R2            R3
```

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:

- S-BFIR monitors all the BFERs;

- 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 frowarding 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 BFER1 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-01 (work in progress), April 2021.

[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-05 (work in progress), February 2021.

[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

Greg Mirsky
ZTE Corporation

Email: gregory.mirsky@ztetx.com

Quan Xiong
ZTE Corporation

Email: xiong.quan@zte.com.cn

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Huanan Li
China Telecom

Email: lihn6@chinatelecom.cn