Workgroup: MBONED WG
Internet-Draft:
draft-ietf-mboned-redundant-ingress-failover-00
Published: 7 April 2022
Intended Status: Informational
Expires: 9 October 2022

Authors: G. Shepherd, Cisco Systems, Inc.
Z. Zhang, Ed., ZTE Corporation
Y. Liu, China Mobile
Y. Cheng, China Unicom

Multicast Redundant Ingress Router Failover

Abstract

This document discusses the redundant ingress router failover in multicast domain.

Status of This Memo

This Internet-Draft will expire on 9 October 2022.

Copyright Notice

Table of Contents

1. Introduction

The multicast redundant ingress router failover is an important issue in multicast deployment. This document tries to do a research on it in the multicast domain. The Multicast Domain is a domain which is used to forward multicast flow according to specific multicast technologies, such as PIM ([RFC7761]), BIER ([RFC8279]), P2MP TE tunnel ([RFC4875]), MLDP ([RFC6388]), etc. The domain may or may not connect the multicast source and receiver directly.

The ingress router is close to the multicast source. The ingress router may connect the multicast source directly, or there may be multiple hops between the ingress router and the multicast source. In the multicast domain, the ingress router is the most adjacent router to the multicast source. It's also called the first-hop router in PIM, or BFIR in BIER, or Ingress LSR in P2MP TE tunnel or MLDP.

The failover function between the multicast source and the ingress router can be achieved by many ways, and it is not included in this document.

The egress router is close to the multicast receiver. The egress router may connect the multicast receiver directly, or there may be multiple hops between the egress router and the multicast receiver. In the multicast domain, the egress router is the most adjacent router to the multicast receiver. It's also called the last-hop router in PIM, or BFER in BIER, or Egress LSR in P2MP TE tunnel or MLDP.

There may be some other function deployed in the multicast domain, such as static configuration, or AMT ([RFC7450]), or SR P2MP Policy ([I-D.ietf-pim-sr-p2mp-policy]).

This document doesn't discuss the details of these technologies. This document discusses the general redundant ingress router failover ways in the multicast domain.

2. Terminology

The following abbreviations are used in this document:

IR: the ingress router which is the most close to the multicast source in the multicast domain.

ER: the egress router which is the most close to the multicast receiver in the multicast domain.

SIR: The IR that is in charge of sending the multicast flow, or the flow from the IR is accepted by the ERs, the IR is called as the Selected-IR, that is SIR in abbreviation.

BIR: The IR that is not in charge of sending the multicast flow, or the flow from the IR is not accepted by the ERs, but the IR replaces the role of SIR once SIR fails. The IR is called as the Backup-IR, that is BIR in abbreviation.

3. Multicast Redundant Ingress Router Failover

```
                            source
                             ...
                       +-----+      +-----+
            +----------+ IR1 +------+ IR2 +---------+
            |multicast +-----+      +-----+         |
            |domain             ...                 |
            |                                       |
            |          +-----+      +-----+         |
            |          | Rm  |      | Rn  |         |
            |          ++---++      +--+--+         |
            |           |   |          |            |
            |     +-----+   +---+      +-----+      |
            |     |             |            |      |
            |   +-v---+      +--v--+      +--v--+   |
            +---+ ER1 +------+ ER2 +------+ ER3 +---+
                +-----+      +-----+      +-----+
                 ...           ...          ...
               receiver      receiver     receiver
```

Figure 1

Usually, a multicast source connects directly, or across multiple hops to two IRs to avoid single node failure. As shown in figure 1, there are two IRs close to a multicast source. The two IRs are UMH (Upstream Multicast Hop) candidates for the ERs.

The two IRs gets multicast flow from the mutlcast source, how to forward the multicast flow to ERs is different according to the technologies deployed in the multicast domain. For example, for PIM which is used in this domain, two PIM Trees that rooted on the two IRs may be built separately.

The IRs works with the other router, such as the ER, in the multicast domain to minimize the multicast flow packet loss during the IR swichover.

3.1. Swichover

There may be some failures occurs in the domain, such as link failure, node failure, if the failed link or node is on the multicast flow forwarding path, there may be multicast flow packet loss.

If there are multiple paths from the IR to the ERs, there is no need to switch IR when some nodes or links fail.

When PIM is used in the domain as multicast forwarding protocol, the forwarding tree for (S, G) or (, G) is built in advance. When a node or link in the forwarding tree fails, the tree is rebuilt partially.

*When BIER is used in the domain as multicast forwarding protocol, there is no need to rebuilt forwarding tree in case of node or link failure, the BIER forwarding recovers along with the IGP routing convergence.

*When P2MP TE tunnel or MLDP is used in the domain as multicast forwarding protocol, the forwarding LSP is built in advance. When a node or link in the LSP fails, the LSP may be rebuilt partially.

*When static multicast tree or SR P2MP policy is used in the domain, the controller needs to re-compute a new forwarding path to bypass the failed node or link.

In some situations, there are some key nodes or links in the network. The multicast path can not be recovered due to the key node or link failure. The IR needs swichover.

```
                  source
                   ...
           +-----+      +-----+
+----------+ IR1 +------+ IR2 +---------+
|          +--+--+      +--+--+         |
|             |            |            |
|          +--+--+      +--+--+         |
|          | Rx  |      | Ry  |         |
|          +-+-+-+      ++---++         |
|            | |         |   |          |
|            | +-----------+ |          |
|            |           | | |          |
|            | +---------+ | |          |
|            | |           | |          |
|          +-v-v-+      +--v-v+         |
|          | Rm  |      | Rn  |         |
|          ++---++      +--+--+         |
|           |   |          |            |
|     +-----+   +---+      +-----+      |
|     |             |            |      |
|   +-v---+      +--v--+      +--v--+   |
+---+ ER1 +------+ ER2 +------+ ER3 +---+
    +-----+      +-----+      +-----+
     ...           ...          ...
   receiver      receiver     receiver
```

Figure 2

For example in figure 2, there is only one path in the network partially. The IR1, Rx are key nodes in the domain, when IR1 or Rx fails, there is no any other path between the IR1 and the ERs.

*When PIM is used in the domain, Rm and Rn may choose Ry as the upstream node to send Join message to build a new tree which rooted with IR2.

*When BIER is used in the domain, IR2 should in charge of the forwarding role to forward the flow to the ERs.

*When P2MP TE tunnel or MLDP is used in the domain, the LSP started from IR2 can be built and replace the used LSP started from IR1 when the used LSP does not work.

*When static multicast tree or SR P2MP policy is used in the domain, the controller should let the IR2 to forward multicast flow to the ERs.

3.2. Failure detection

In order to achieve the successful IR switchover, some methods should be used for monitoring the IR node failure or the path failure between IR and ERs, and the IR can do the switching once the failure occurs. BFD or PING methods can be used for it.

BFD [RFC5880] can be used in all the deployments. Multipoint BFD [RFC8562] can also be used for the failure detection between IR and ERs. BFD for MPLS LSPs [RFC5884] can be used in P2MP TE tunnel or MLDP deployments. BIER BFD [I-D.ietf-bier-bfd] can be used in BIER deployment.

IPv4 PING [RFC0792] and IPv6 PING [RFC4443] can also be used in all the deployments. LSP-Ping [RFC8029] can be used for P2MP TE tunnel or MLDP deployments. BIER PING [I-D.ietf-bier-ping] can be used in BIER deployment.

BIR and ER can detect the SIR node and path failure easily by the BFD and PING methods. If the monitoring is between SIR and ER, how to trigger the switchover quickly is challenging when BIR needs to start forwarding the multicast flow. If the monitoring is between BIR and SIR, the path between BIR and SIR may fail, but the path is not the way from SIR to ERs, BIR may trigger the switchover by mistake, in this case unnecessary duplicate flow occurs. In this case, the ER must support the selective receiving and can be compatible with the IR switchover mistake. In order to minimize the mistaken switchover, the reliability of SIR/BIR detection needs to be enhanced, such as using redundant reliable paths for detection, etc.

4. Stand-by Modes

In case there are more than one IRs can be the UMH, and there is no other path from an IR to ERs in case of the IR fails, the IR needs to be switched.

Usually there are three types of stand-by modes in multicast IR protection. [RFC9026] has some description on it. This document discusses the detail of the three modes here.

The ER may send request to upstream router or IR when it finds the node or path failure. The request from the ER may be the PIM tree building, or BIER overlay protocol signaling, or LSP building, or some other ways to let IR knows whether forwards the multicast flow.

4.1. Cold

In cold standby mode, the ER selects an SIR, for example IR1 in figure 1, as the SIR and signals to it to get the multicast flow.

When the ER finds that the SIR is down, or the ER finds that it cannot receive flow from IR1, the ER signals to IR2 to get the multicast flow.

- *For IR, the IRs, include SIR and BIR, just do the regular operation of forwarding flow according to the request from the ER.

- *For ER, the ER must select an IR as the SIR and signal to it. When the SIR fails or the path between the SIR and ER fails, the ER must signal to the BIR to get the flow.

- *For the intermediate routers, they know nothing about the role of IR, they just do the packet forwarding. There is no duplicate packets in the domain.

In case of the IR switchover, the ER detects the failure of SIR, and signals to the BIR. There is packet loss during the signaling until the ER receives the flow from the BIR.

4.2. Warm

In Warm standby mode, the ER signals to both IR1 and IR2.

In case IR1 is the SIR, IR1 forwards the flow to the ER. The BIR, for example the IR2, must not forward the flow to the ER until the SIR is down.

- *For IR, the IR should take the role of SIR or BIR. The BIR must not forward flow to the ER. When the SIR fails or the path between SIR and ER fails, the BIR must start forwarding the flow to ER. But it's hard to know the failure for BIR itself, some methods should be taken to let the BIR to get the failure notification.

- *For ER, the ER does not select the SIR or BIR. The ER just signal to both of them.

- *For the intermediate routers, they know nothing about the role of IR, they just do the packet forwarding. There is no duplicate packets in the domain.

In case of the IR switchover, the BIR detects the failure of the SIR and switch to SIR. There is packet loss during the IR switchover.

In some deployments, the SIR and BIR may in charge of different multicast flow. For a specific multicast flow, the SIR may be IR1, for another multicast flow, the SIR may be IR2. So the two IRs can share the multicast forwarding load. And another possible deployment is, the two IRs can in charge of different ERs for one multicast flow. For example, IR1 sends the multicast flow to some of the ERs, and IR2 send the multicast flow to the other ERs. In case IR1 detects there is something wrong between IR1 and the ERs, IR1 may notify IR2 to take over the responsibility of forwarding the multicast flow to these ERs that receive flow from IR1 before.

4.3. Hot

In Hot standby mode, the ER signals to both IRs.

Both IRs are sending the flow to the ER. The ER must discard the duplicate flow from one of the IRs.

In this situation, there are no SIR or BIR. Only ER knows which IR is the SIR.

- *For IR, the IR need not to know the roles of SIR or BIR, IR just forwarding the flow according to the request received from ER.

- *For ER, the ER signal to both of the IRs to get the flow. And the ER must discard the duplicated flow from the backup BIR. When the SIR fails or the path between SIR and ER fails, the ER must switch the forwarding plane to forward the flow packet comes from the BIR. To be noted, the ERs may choose different SIR or BIR.

- *For the intermediate routers, they know nothing about the role of IR, they just do the packet forwarding. There are duplicate packets forwarded in the domain.

In case of the IR switchover, the ER detects the failure of the SIR. Because there are duplicate flow packets arrive on the ER, the ER just switch to forward the flow comes from the BIR. There may be packet loss during the switching.

4.4. Summary

The table is a brief comparison among the three modes. The 'SIR failover' means the SIR fails or the path between SIR and ER fails.

role	Cold Mode	Warm Mode	Hot Mode
IR	Forwarding flow according to the request from ER.	Takes the role of SIR or BIR, BIR must not forward flow to ER until SIR failovers.	Need not to know the roles of SIR or BIR, just forwarding flow according to the request from ER.
ER	Must select an IR as SIR to signal the request, signal to the BIR	Does not select the SIR or BIR, just signal to both of them.	Signal to both of SIR and BIR. Discards the duplicate flow

role	Cold Mode	Warm Mode	Hot Mode
	to request the flow when SIR failovers.		from BIR until SIR failover.
Intermediate Router	Knows nothing about SIR or BIR. No duplicated flow is forwarded.	Knows nothing about SIR or BIR. No duplicated flow is forwarded.	Knows nothing about SIR or BIR. Duplicated flow is forwarded.

Table 1

The Cold stand-by mode is the easiest way to implementated, but it takes the longest converge time.

The Hot stand-by mode takes the most less packet loss, but there is duplicated packet forwarding in the domain, more bandwidth is occupied.

The Warm stand-by mode takes the middle packet loss and converge time, but it's hard for BIR to know the failure between SIR and ERs.

So it's hard to say which mode is the best way for multicast redundant ingress router failover, the network administrator should select the most suitable mode according to the network deployment.

5. IANA Considerations

This document does not have any requests for IANA allocation.

6. Security Considerations

This document adds no new security considerations.

7. References

7.1. Normative References

[RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <https://www.rfc-editor.org/info/rfc4875>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC7450] Bumgardner, G., "Automatic Multicast Tunneling", RFC 7450, DOI 10.17487/RFC7450, February 2015, <https://www.rfc-editor.org/info/rfc7450>.

[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>.

[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>.

7.2. Informative References

[I-D.ietf-bier-bfd] Xiong, Q., Mirsky, G., Hu, F., and C. Liu, "BIER BFD", Work in Progress, Internet-Draft, draft-ietf-bier-bfd-01, 8 April 2021, <https://www.ietf.org/archive/id/draft-ietf-bier-bfd-01.txt>.

[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://www.ietf.org/archive/id/draft-ietf-bier-ping-07.txt>.

[I-D.ietf-pim-sr-p2mp-policy] (editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", Work in Progress, Internet-Draft, draft-ietf-pim-sr-p2mp-policy-04, 7 March 2022, <https://www.ietf.org/archive/id/draft-ietf-pim-sr-p2mp-policy-04.txt>.

[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>.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <https://www.rfc-editor.org/info/rfc5884>.

[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>.

[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>.

[RFC9026] Morin, T., Ed., Kebler, R., Ed., and G. Mirsky, Ed., "Multicast VPN Fast Upstream Failover", RFC 9026, DOI 10.17487/RFC9026, April 2021, <https://www.rfc-editor.org/info/rfc9026>.

Authors' Addresses

Greg Shepherd
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose,
United States of America

Email: gjshep@gmail.com

Zheng Zhang (editor)
ZTE Corporation
Nanjing
China

Email: zhang.zheng@zte.com.cn

Yisong Liu
China Mobile
Beijing

Email: liuyisong@chinamobile.com

Ying Cheng
China Unicom
Beijing
China

Email: chengying10@chinaunicom.cn