BIER Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 15, 2025

Y. Liu
China Mobile
C. Lin
New H3C Technologies
Z. Zhang
ZTE Corporation
Y. Qiu
New H3C Technologies

February 15, 2025

BIER Loop Avoidance using Segment Routing
draft-liu-bier-uloop-05

Abstract

This document provides a mechanism leveraging SR-MPLS/SRv6 to ensure that BIER messages can be forwarded loop-freeness during the IGP reconvergence process following a link-state change event.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on August 15, 2025.

Copyright Notice

Table of Contents

1. Introduction.....2
   1.1. Requirements Language.....4
2. Loop-free convergence process.....4
3. Computing Loop-avoiding Path.....5
   3.1. Explicit Path of Loop-avoiding.....5
   3.2. Calculation Method of Explicit Path.....6
4. Example Application.....7
5. IANA Considerations.....8
6. Security Considerations.....8
7. References.....9
   7.1. Normative References.....9
   7.2. Informative References.....9
8. Acknowledgments.....9
Authors' Addresses.....10

1. Introduction

Forwarding loops happen during the convergence of the IGP, as a result of transient inconsistency among forwarding states of the nodes of the network.

When the network topology changes, loops may appear on new forwarding paths due to the different convergence speeds of each node's routing.

During the multicast packet forwarding process, when the upstream BFR senses that its BFR-NBR is not reachable, the upstream BFR as a PLR node can quickly switch multicast traffic to backup path through the BIER FRR mechanism [I-D.ietf-bier-frr]. If the network fails to recover, multicast traffic will switch back from the backup path to the primary path.

As shown in Figure 1 below, R1 is connected to the multicast source, and all IGP links are symmetric metric. Except for the link cost between R7 and R8, which is 100, the cost of all other links is 1.

The multicast data packet sent from R1 to R9 is initially forwarded along the path R1->R2->R3->R4->R9. When the link between R2 and R3 fails, node R3 fails, or the link failure between R2 and R3 is restored, there may be a loop in packet forwarding between R2 and R7 during the routing convergence process.

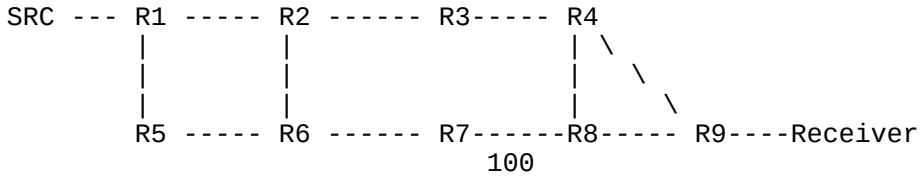

Figure 1

Both of the following scenarios may result in a loop in the forwarding of BIER messages.

* Scenario 1: When a link or node fails, a micro-loop may occur during routing convergence.

  Taking Figure 1 as an example, the fault process is as follows:

  1) When the link between R2 and R3 fails or R3 fails, R2 will detect that the route to R3 is unreachable. If BIER FRR is enabled on R2, based on BIER FRR, R2 will first choose to send messages to backup neighbor R6. If R2 does not enable BIER FRR, R2 will lose packets. At this stage, FRR can enhance network reliability.

  2) R2 floods topology change event through IGP. After receiving IGP messages, each network node recalculates the SPF tree.

  3) After calculating the new unicast SPF tree, issue the new unicast forwarding table entries and update the BIFTs of multicast.

  Normally, after R2 completes the routing convergence to R9, the BFR NBR from R2 to R9 becomes R6. When R2 receives a BIER message sent to R9, R2 searches for BIFT and forwards the message to R6.

  If the convergence speed of R6’s routing is slower than R2, when R2 has completed convergence, R6 is not yet complete. During the convergence process of R6, the BFR BNR recorded in BIFT on R6 to R9 will still be R2. Therefore, after receiving the BIER message, R6 will send the message back to R2 according to the Bitstring in the BIER header. An instant micro-loop between R2 and R6 appears.

* Scenario 2: After fault recovery, a micro-loop may also occur during routing convergence.

  In Figure 1, if the link between R2 and R3 fails, after the entire network routing converges, the BIER packet forwarding path from R1 to R9 becomes R1->R2->R6->R7->R8->R9.

  When the link failure between R2 and R3 is recovered, during the routing convergence process, if the routing of R6 converges faster than R2, before R2 completes the routing convergence, because R2 still records the BFR NBR to R9 as R6, when R2 receives the BIER message sent to R9, it will still forward the message to R6. Resulting in a short period of micro-loop.

This document provides a mechanism leveraging SR-MPLS/SRv6 to ensure that BIER messages can be forwarded loop-freeness during the IGP reconvergence process following a topology change event.

1.1. 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. Loop-free convergence process

Upon a topology change, when a BFR converging for BFERs does not trust the loop-freeness of its post-convergence paths for BFERs, it performs convergence processing as follows.

After computing the new path to BFER, for a predetermined amount of time C (Corresponding to an avoiding uloop timer), BFR installs an explicit path through which packets can be steered to BFER without loop. For example, forwarding through an explicit SR-MPLS/SRv6 path or through a P2MP path. C should be greater than or equal to the worst-case convergence time of a node, network-wide. The determination of "C" is outside the scope of this document. The forwarding path is computed when the event occurs.

After C elapses, BFR installs the normal post-convergence forwarding entry for BFER that ensure the loop-free property.

Taking R1 in Figure 1 as an example, When R1 senses a topology change event between R2 and R3 through unicast routing, it recalculates the forwarding path of multicast packets for the affected nodes. Within the interval C, R1 specifies an explicit path

to send multicast packets to R9 along the link between R7-R8. Before the routing convergence is completed, the multicast traffic is forwarded along the path R1->R6->R7->R8->R9.

3. Computing Loop-avoiding Path

When the link between R2 and R3 fails, unicast routing converges, and all BFRs affected by topology change recalculate the SPF tree. Before the routing of all nodes completes convergence, BFR steers the packet targeting R9 to an explicit path, forwarding it based on the specified node label or SRv6 SID in the explicit path.

When the link failure between R2 and R3 is restored, the process of avoiding micro-loop is also similar.

3.1. Explicit Path of Loop-avoiding

The explicit path of loop-avoiding can be (but not limited to):

* SR policy TE path. Treat each node or adjacent SID on the explicit path as a segment on the SR Policy TE path. During the convergence process, add SRv6 encapsulation to the BIER message, specify the SRH Segment List, and send it to the endpoint of the explicit path. After reaching the endpoint, decapsulate the outer layer SR-MPLS/SRv6 packet header, restore the original BIER packet, and continue forwarding according to the BIER header.

* BIER-TE forwarding path. During the convergence process, each node on the explicit path is treated as a BIER-TE node and forwarded through BIER-TE. If the explicit paths to different BFERs require passing through some identical replication nodes, the BFR nodes on these explicit paths can be arranged as BIER-TE forwarding paths. During routing convergence, BIER packets are first forwarded along the BIER-TE path to the endpoint of the avoiding micro-loop path, then unpacked and the inner layer BIER packets are restored, and finally forwarded according to the BIER header.

  The bit position of nodes on the BIER-TE path can be arranged into the bitstring of the original BIER header, or the BIER-TE header can be encapsulated outside the BIER message.

* P2MP policy forwarding path. If the explicit paths to different BFERs require passing through some identical replication nodes, the BFR nodes on these explicit paths can be arranged as P2MP policy forwarding paths. During the convergence process, multicast messages are forwarded through the path specified by the P2MP policy.

This document focuses on using the SR-MPLS and SRv6 TE path as the path for multicast avoidance loop.

3.2. Calculation Method of Explicit Path

There are currently two methods to calculate the nodes included in the explicit path.

Method 1: Similar to [RFC7490], using the concept of P-Space and Q-Space for TI-LFA generate explicit SR/SRv6-based path from P to Q. The repair list is expressed generally as {P node (NODE SID), all ADJ/End.X SIDs from P node to Q node}.

1) Using BFER as the destination node, BFR calculates the optimal convergence path tree. That is, when the topology changes, unicast routing converges, and BFR calculates a new SPF tree.

2) Find the Q node. On the new SPF tree, traverse the parent nodes starting from the BFER destination node until finding the farthest node from BFER, which is not affected by the link failure and can reach BFER, as the Q node.

3) Find the P node. On the new SPF tree, traverse the parent nodes starting from the Q node until finding a node that is not affected by the topology change on the path from current BFR to that node. This node will be considered as the P node.

4) Calculate the repair segment list path. The repair segment list path is found by computing the explicit SR/SRv6-based path from P to Q when these nodes are not adjacent along the convergence path.

   * For SR-MPLS, the repair list is expressed {Node_SID(P), AdjSID(P->Q)}.

   * For SRv6, the repair list is expressed {END_SID(P), END.X (P->Q)}.

5) After BFR completes routing convergence locally, start an avoiding uloop timer.

6) Before the timeout of the avoiding uloop timer, if BFR receives the BIER message, it sends the BIER message to the endpoint along the explicit path indicated by the repair list.

7) After the endpoint receives the message, remove the explicit path encapsulation, restore the BIER message, and then continue forwarding according to the BIER header.

Method 2: Directly generate a strict explicit path from current BFR to Q node.

* For SR-MPLS, the repair list is expressed {AdjSID(S->Q)}.

* For SRv6, the repair list is expressed {END.X(S->Q)}.

If the P nodes and Q nodes of different BFERs are the same, which means that multicast packets can be forwarded through the same path, it is necessary to merge the multicast forwarding paths to avoid headend replication. Try to place the multicast replication point on the node closest to the multicast receivers.

4. Example Application

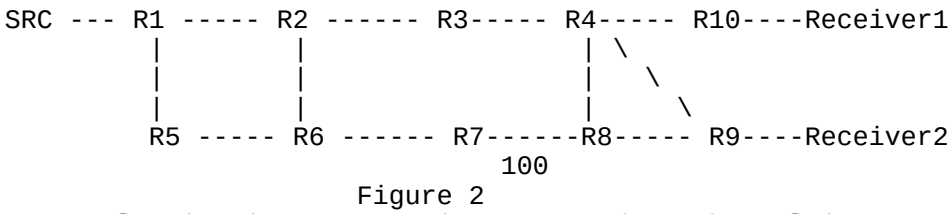

Figure 2

As an example, in Figure 2, R1 is connected to the multicast source, R9 and R10 are BFERs. All IGP links are symmetric metric. Except for the link cost between R7 and R8, the cost of all other links is 1.

The multicast data packet sent from R1 to R9 and R10 is initially forwarded along the path R1->R2->R3->R4->R9/R10. When the link between R2 and R3 fails, the forwarding path of BIER messages from R1 to R9/R10 becomes R1->R2->R6->R7->R8->R9/R10 after the convergence of the entire network routing is completed.

When the link between R2 and R3 recovers, R2 will perceive the topology change and recalculate the SPF tree, and inform IGP neighbors of the topology change event.

All BFRs that receive topology change event need to recalculate the SPF tree. To avoid micro loop during routing convergence, these BFRs also need to calculate the repair path and repair list for each BFER.

Taking BFIR node R1 as an example, the process is as follows:

1) When R1 receives a topology change event between R2 and R3, R1 calculates the new SPF tree and the repair path for each BFER.

2) R1 calculates the Q node for R9, and the result is [R8].

3) R1 calculates the P node for R9, and the result is [R7].

4) The path between P node and Q node is the repair path of R9.

5) R1 repeats the above process to calculate the repair path for R10. The P and Q nodes are also R7 and R8.

BFERs with the same P and Q nodes use the same repair path. Because R9 and R10 have the same repair path, in the avoiding uloop interval, the BIER message is first forwarded to the endpoint of the repair path, and then is replicated at the endpoint.

To R9 and R10, both need to go through the R7->R8 link.

* For SR-MPLS, the repair list of R1 for R9 and R10 considering the fault recovery of link between R2 and R3 or of node R3 is: <NodeSID(R7), AdjSID_R7R8>.

* For SRv6, the repair list of R1 for R9 and R10 considering the fault recovery of link between R2 and R3 or of node R3 is: <End.X_R7R8>.

When the link state between R2 and R3 changes to UP, R1 converges routing and sends the BIER message to R8 along the SR-MPLS/SRv6 path indicated by the repair list in the avoiding uloop interval.

After R8 receives the message, remove the outer SR-MPLS/SRv6 encapsulation, restore the BIER message, and then continue forwarding according to the BIER header.

5. IANA Considerations

No requirements for IANA.

6. Security Considerations

The behavior described in this document is internal functionality to a router that result in the ability to explicitly steer traffic over the post convergence path after a remote topology change in a manner that guarantees loop freeness. Because the behavior serves to minimize the disruption associated with a topology change, it can be seen as a modest security enhancement.

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>.

[RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., So, N., "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", BCP 14, RFC 8174, DOI 10.17487/RFC7490, April 2015, <https://www.rfc-editor.org/info/rfc7490>.

[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

TBD

8. Acknowledgments

The authors would like to thank the following for their valuable contributions of this document:

TBD

Authors' Addresses

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Changwang Lin
New H3C Technologies

Email: linchangwang.04414@h3c.com

Zheng Zhang
ZTE Corporation

Email: zhang.zheng@zte.com.cn

Yuanxiang Qiu
New H3C Technologies

Email: qiuyuanxiang@h3c.com