Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 February 2024

Z. Hu
Huawei Technologies
H. Chen
Futurewei
J. Yao
Huawei Technologies
C. Bowers
Juniper Networks
Y. Zhu
China Telecom
Y. Liu
China Mobile

21 August 2023

SR-TE Path Midpoint Restoration
draft-hu-spring-segment-routing-proxy-forwarding-24

Abstract

Segment Routing Traffic Engineering (SR-TE) supports explicit paths using segment lists containing adjacency-SIDs, node-SIDs and binding-SIDs. The current SR FRR such as TI-LFA provides fast re-route protection for the failure of a node along a SR-TE path by the direct neighbor or say point of local repair (PLR) to the failure. However, once the IGP converges, the SR FRR is no longer sufficient to forward traffic of the path around the failure, since the non-neighbors of the failure will no longer have a route to the failed node. This document describes a mechanism for the restoration of the routes to the failure of a SR-MPLS TE path after the IGP converges. It provides the restoration of the routes to an adjacency segment, a node segment and a binding segment of the path. With the restoration of the routes to the failure, the traffic is continuously sent to the neighbor of the failure after the IGP converges. The neighbor as a PLR fast re-routes the traffic around 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

Copyright Notice

Table of Contents

1. Introduction

Segment Routing Traffic Engineering (SR-TE) is a technology that implements traffic engineering using a segment list. SR-TE supports the creation of explicit paths using adjacency-SIDs, node-SIDs, anycast-SIDs, and binding-SIDs. A node-SID in the segment list defining an SR-TE path indicates a loose hop that the SR-TE path should pass through. When the node fails, the network may no longer be able to properly forward traffic on that SR-TE path.

[I-D.ietf-rtgwg-segment-routing-ti-lfa] describes an SR FRR mechanism that provides fast re-route protection for the failure of a node on a SR-TE path by the direct neighbor or say point of local repair (PLR) to the failure. However, once the IGP converges, the SR FRR is no longer sufficient to forward traffic of the path around the failure, since the non-neighbors of the failure will no longer have a route to the failed node and drop the traffic.

To solve this problem, [I-D.ietf-spring-segment-protection-sr-te-paths] proposes that a hold timer should be configured on every router in a network. After the IGP converges on the event of a node failure, if the node-SID of the failed node becomes unreachable, the forwarding changes should not be communicated to the forwarding planes on all configured routers (including PLRs for the failed node) until the hold timer expires. This solution may not work for some cases such as some of nodes in the network not supporting this solution.

This document describes a proxy forwarding mechanism for the restoration of the routes to the failure of a SR-MPLS TE path after the IGP converges. It provides the restoration of the routes to an adjacency segment, a node segment and a binding segment on a failed node along the path. With the restoration of the routes to the failure, the traffic for the SR-MPLS TE path is continuously sent to the neighbor of the failure after the IGP converges. The neighbor as a PLR fast re-routes the traffic around the failure.

1.1. Terminology

SR: Segment Routing.

PLR: Point of Local Repair.

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

LSA: Link State Advertisement in OSPF.

LS: Link State, which is LSP or LSA.

2. Proxy Forwarding

In the proxy forwarding mechanism, each neighbor of a possible failed node advertises its SR proxy forwarding capability in its network domain when it has the capability. This capability indicates that the neighbor (Proxy Forwarder) will forward traffic on behalf of the failed node. A router (non-neighbor) receiving the SR Proxy Forwarding capability from the neighbors of a failed node will send traffic using the node-SID of the failed node to the nearest Proxy Forwarder after the IGP converges on the event of the failure.

Once receiving the traffic, the Proxy Forwarder sends the traffic on the post-failure shortest path to the node immediately following the failed node in the segment list.

For a binding SID of a possible failed node, the information about the binding, including the binding SID and the list of SIDs associated with the binding SID, is advertised to the neighbors of the node.

After the node fails and the IGP converges on the failure, the non-neighbors of the failed node send the traffic with the node-SID of the failed node followed by the binding SID to the neighbor (Proxy Forwarder) of the failed node. Once receiving the traffic with the node-SID of the failed node, the Proxy Forwarder finds the forwarding entry for the node-SID of the failed node in its Routing Table and pops the node-SID. According to the action in the entry, the Proxy Forwarder will swap the binding SID with the list of SIDs associated with the binding SID and send the traffic along the post-failure shortest path to the first node in the segment list.

3. Protocol Extensions/Re-uses for Proxy Forwarding

This section describes the semantic of protocol extensions/re-uses for advertising the information about each binding segment (including its binding SID and the list of SIDs associated with the binding SID) of a node and the SR proxy forwarding capability of a node in a network domain.

3.1. Advertising Binding Segment

For a binding segment (or binding for short) on a node A, which consists of a binding SID and a list of SIDs, the binding (i.e., the binding SID and the list of the SIDs) with the ID of node A is advertised.

There are different types of IDs of node A. For example, node A’s name, BGP router ID, and IGP ID (OSPF router ID or ISIS system ID) are IDs of node A. The IGP ID of node A MUST be used as the ID of node A. When OSPF runs in the network, a OSPF router ID is an IGP ID; when ISIS runs in the network, an ISIS system ID is the IGP ID. PCE and others know which IGP (OSPF or ISIS) runs in the network and can obtain the IGP ID of a node.

When a protocol (such as PCE or BGP running on a controller) supports sending a binding on node A to A, we may extend this protocol to send the binding to A’s neighbors if the controller knows the neighbors and there are protocol (PCE or BGP) sessions between the controller and the neighbors. Alternatively, we may extend YANG and IGP to advertise the binding to A’s neighbors.

Note: how to send bindings on node A to A’s neighbors via which protocol is out of the scope of this document.

3.2. Advertising Proxy Forwarding

When a node P is able to do SR proxy forwarding for its neighboring nodes for protecting the failures of these nodes, P advertises its SR proxy forwarding capability for these nodes. P advertises the mirror SID [RFC8402] for a node N (Neighboring node of P) using IS-IS extensions [RFC8667] to indicate P’s capability for N.

Node N advertises its node-SID to every node in the network. In normal operations, a non-neighbor node X of node N sends the packet with the node-SID of node N to node N. When node N fails, node X sends the packet with the node-SID of node N to node P, and node P does a SR proxy forwarding for node N and forwards the packet towards its final destination without going through node N.

Note that the behaviors of normal IP forwarding and routing convergences in a network are not changed at all by the SR proxy forwarding. For example, the next hop used by BGP is an IP address (or prefix). The IGP and BGP converge in normal ways for changes in the network. The packet with its IP destination to this next hop is forwarded according to the IP forwarding table (FIB) derived from IGP and BGP routes.

Similar to IS-IS [RFC8667], OSPF should be extended for advertising mirror SID to indicate the capability. Note that OSPF extensions is out of the scope of this document.

4. Proxy Forwarding Example

This section illustrates the proxy forwarding for a binding SID through an example. The proxy forwarding for a node-SID and an adjacency SID can refer to [I-D.ietf-spring-segment-protection-sr-te-paths] or Appendix.

Figure 1 is an example network topology used to illustrate the proxy forwarding mechanism for a binding SID. Each node RTi has SRGB = [i000-i999]. RT1 is an ingress node of SR domain. RT3 is a failure node. RT2 is a Point of Local Repair (PLR) node, i.e., a proxy forwarding node. Label Stack 1 uses a node-SID and a binding-SID. The Binding-SID with label = 100 at RT3 represents the ECMP-aware path RT3->RT4->RT5. So Label Stack 1, which consists of the node-SID of RT3 followed by Binding-SID = 100, represents the ECMP-aware path RT1->RT3->RT4->RT5.

```
             Node SID:2      Node SID:3
             +-----+          +-----+
             |     |----------+     |
           / |RT2  |          | RT3 |\
          /  +-----+          +-----+ \
         /      | \           /|     \
        /       |  \         / |      \
       /        |   \       /  |       \
      /         |    \     /   |        \
     /          |     \   /    |         \
 Node SID:1     |      \ /     |          \Node SID:4    Node SID:5
+-----+         |       \ /    |           +-----+       +-----+
|     |         |        X     |           |     |-------|     |
| RT1 |         |       / \    |           | RT4 |       | RT5 |
+-----+         |      /   \   |           +-----+       +-----+
   \            |     /     \  |           /
    \           |    /       \ |          /
     \          |   /         \|         /
      \         |  /           \        /
       \        | /             \|     /
        \       |/               |    /
         \   +-----+           +-----+ /
          \  |     |           |     |/
           \ | RT6 |-----------| RT7 |
             +-----+           +-----+
             Node SID:6        Node SID:7

+-----------------+  +---------------+
|    Node SRGB    |  | Adj-SID       | +--------+ +--------+ +--------+
+-----------------+  +---------------+ |Label   | |Label   | |Label   |
| RT1:[1000-1999] |  |RT1->RT2:10012|  |Stack 3|  |Stack 2|  |Stack 1|
+-----------------+  +---------------+ +--------+ +--------+ +--------+
| RT2:[2000-2999] |  |RT2->RT3:20023|  | 10012  | | 1003   | | 1003   |
+-----------------+  +---------------+ +--------+ +--------+ +--------+
| RT3:[3000-3999] |  |RT3->RT6:30036|  | 20023  | | 3004   | | 100    |
+-----------------+  +---------------+ +--------+ +--------+ +--------+
| RT4:[4000=4999] |  |RT3->RT7:30037|  | 30034  | | 4005   |  100 is
+-----------------+  +---------------+ +--------+ +--------+ binding SID
| RT5:[5000-5999] |  |RT3->RT4:30034|  | 40045  |            to
+-----------------+  +---------------+ +--------+           {30034,40045}
| RT6:[6000-6999] |  |RT7->RT4:70074|
+-----------------+  +---------------+
| RT7:[7000-7999] |  |RT4->RT5:40045|
+-----------------+  +---------------+
```

Figure 1: Topology of SR-TE Path

4.1. Advertising Proxy Forwarding

If the Point of Local Repair (PLR), for example, RT2, has the capability to do SR proxy forwarding for its neighboring nodes such as RT3, RT2 advertises this capability to all the other nodes in the network. When RT3 fails, RT2 needs to maintain its SR proxy forwarding capability for a period of time. When the proxy forwarding table corresponding to the fault node is deleted, the capability is withdrawn.

Every node advertises its node-SID to all the other nodes in the network. For example, RT3 advertises its node-SID to all the other nodes. The other nodes (e.g., RT1) learn RT3's node-SID and the proxy forwarding capability of RT2, which is a neighbor of RT3. When RT3 is normal, the nodes (e.g., RT1) prefer the route to RT3 for the traffic with RT3's node-SID. When the RT3 fails, the nodes use the route to RT2 (proxy forwarder for RT3) for the traffic with RT3's node-SID.

For RT3's binding-SID 100, which is associated with segment list {30034, 40045}, the binding (i.e., 100 bond to {30034, 40045}) with RT3's ID is advertised to RT3's neighbors RT2, RT4 and RT7. RT2 as PLR uses the binding to build an entry for proxy forwarding for binding-SID 100 in its Proxy Forwarding Table for RT3. RT2 uses the entry when RT3 fails.

4.2. Building Proxy Forwarding Table

A SR proxy node P (e.g., RT2) needs to build an independent proxy forwarding table for each neighbor N (e.g., RT3). The proxy forwarding table for node N contains the following information:

1: Node N's SRGB range and the difference between the SRGB start value of node P and that of node N,

2: Every adjacency-SID of N and Node-SID of the node pointed to by node N's adjacency-SID, and

3: Every binding-SID of N and the label stack associated with the binding-SID.

Node P (PLR) uses a proxy forwarding table based on the next segment to find a backup forwarding entry for the adjacency-SID and Node-SID of node N. When node N fails, node P maintains the proxy forwarding table for N for a period of time, which is recommended for 30 minutes.

RT2 (as P) in Figure 1 builds the proxy forwarding table for RT3 (as N) as shown in Figure 2.

In-label	SRGBDiffValue	Next Label	Action	Map Label
2003	-1000	30034	Fwd to RT4	2004
		30036	Fwd to RT6	2006
		30037	Fwd to RT7	2007
		100	Swap to { 30034, 40045 }	

Figure 2: RT2's Proxy Forwarding Table for RT3

1: The difference (SRGBDiffValue) between the SRGB start value of RT2 (P) and that of RT3 (N) is -1000 since the SRGB start value of RT2 is 2000 and that of RT3 is 3000.

2: RT3 has adjacency-SIDs 30034, 30036 and 30037 for the adjacencies from RT3 to RT4, RT6 and RT7 respectively. The node-SIDs of RT4, RT6 and RT7 are 2004, 2006 and 2007 respectively (i.e., the node-SIDs of the nodes pointed to by RT3's adjacency-SIDs 30034, 30036 and 30037 are 2004, 2006 and 2007 respectively). RT2 builds a forwarding entry for each of RT3's adjacency-SIDs 30034, 30036 and 30037. The entry contains the adjacency-SID (e.g., 30034) in Next Label column, forward (fwd) to adjacent node (e.g., fwd to RT4) in Action column, and the node-SID of the adjacent node (e.g., 2004) in Map Label column.

3: RT3 has binding-SID 100, which is associated with label stack {30034, 40045}. RT2 builds a forwarding entry for binding-SID 100 in the proxy forwarding table for RT3. The entry contains binding-SID 100 in Next Label column and "Swap to {30034, 40045}" in Action column.

4.3. Proxy Forwarding for Binding Segment

This Section shows through example how a proxy node uses the SR proxy forwarding mechanism to forward traffic to the destination node when a node fails and the next segment of label stack is a binding-SID.

As shown in Figure 1, Label Stack 1 {1003, 100} represents SR-TE loose path RT1->RT3->RT4->RT5, where 100 is a Binding-SID, which represents segment list {30034, 40045}.

When RT3 fails, RT1 forwards the packet with RT3’s node-SID 1003 to RT2, which is the proxy forwarder for RT3. RT2 acts as a PLR and uses Binding-SID to query the proxy forwarding table locally built for RT3. RT2 gets the label forwarding path to RT3’s next hop node (RT4), which bypasses RT3. The specific steps are as follows:

a. RT1 swaps label 1003 to out-label 2003 to RT3.

b. RT2 receives the label forwarding packet whose top label of label stack is 2003 (RT3’s node-SID) and finds the forwarding entry for 2003 in its Routing Table. The action in the entry is to lookup the Proxy Forwarding table for RT3 due to RT3 failure. RT2 pops label 2003.

c. RT2 uses Binding-SID:100 to lookup the forwarding entry (Next Label record) in the Proxy Forwarding Table. The action in the entry is to swap to Segment list {30034, 40045}.

d. RT2 swaps Binding-SID:100 to Segment list {30034, 40045}, and uses 30034 (RT3’s Adjacency-SID for the adjacency from RT3 to RT4) to lookup the forwarding entry (Next Label record) in the Proxy Forwarding table again. The action in the entry is to forward the packet to RT4.

e. RT2 queries its Routing Table to RT4, using primary or backup path to RT4. The next hop is RT7.

f. RT2 forwards the packet to RT7. RT7 queries its routing table to forward the packet to RT4.

5. Security Considerations

The extensions to OSPF and IS-IS described in this document result in two types of behaviors in data plane when a node in a network fails. One is that for a node, which is a upstream (except for the direct upstream) node of the failed node along a SR-TE path, it continues to send the traffic to the failed node along the SR-TE path for an extended period of time. The other is that for a node, which is the direct upstream node of the failed node, it fast re-routes the traffic around the failed node to the direct downstream node of the failed node along the SR-TE path. These behaviors are internal to a network and should not cause extra security issues.

6. Acknowledgements

The authors would like to thank Peter Psenak, Acee Lindem, Les Ginsberg, Bruno Decraene, Joel Halpern and Jeff Tantsura 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>.

[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>.

[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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

7.2. Informative References

[I-D.ietf-rtgwg-segment-routing-ti-lfa] Litkowski, S., Bashandy, A., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, Internet-Draft, draft-ietf-rtgwg-segment-routing-ti-lfa-11, 30 June 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-rtgwg-segment-routing-ti-lfa-11>.

[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", Work in Progress, Internet-Draft, draft-ietf-spring-segment-protection-sr-te-paths-04, 10 March 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-spring-segment-protection-sr-te-paths-04>.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-22, 22 March 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-spring-segment-routing-policy-22>.

Appendix A. Proxy Forwarding for Adjacency and Node Segment

This Section shows through example how a proxy node forward traffic to the destination node when a node fails and the next segment of label stack is an adjacency-SID or node-SID.

A.1. Next Segment is an Adjacency Segment

As shown in Figure 1, Label Stack 3 {10012, 20023, 30034, 40045} uses only adjacency-SIDs and represents the SR-TE strict explicit path RT1->RT2->RT3->RT4->RT5. When RT3 fails, node RT2 acts as a PLR, and uses next adjacency-SID (30034) of the label stack to lookup the proxy forwarding table built by RT2 locally for RT3. The path returned is the label forwarding path to RT3’s next hop node RT4, which bypasses RT3. The specific steps are as follows:

a. RT1 pops top adjacency-SID 10012, and forwards the packet to RT2;

b. RT2 uses the label 20023 to identify the next hop node RT3, which has failed. RT2 pops label 20023 and queries the Proxy Forwarding Table corresponding to RT3 with label 30034. The query result is 2004. RT2 uses 2004 as the incoming label to query the label forwarding table. The next hop is RT7, and the incoming label is changed to 7004.

c. So the packet leaves RT2 out the interface to RT7 with label stack {7004, 40045}. RT7 forwards it to RT4, where the original path is rejoined.

d. RT2 forwards packets to RT7. RT7 queries the local routing table to forward the packet to RT4.

A.2. Next Segment is a Node Segment

As shown in Figure 1, Label Stack 2 {1003, 3004, 4005} uses only node-SIDs and represents the ECMP-aware path RT1->RT3->RT4->RT5, where 1003 is the node-SID of RT3.

When the node RT3 fails, the non-neighbors (e.g., RT1) of RT3 prefer the route to the proxy SID implied/advertised by RT2 (proxy forwarder for RT3). Node RT2 acts as a PLR node and queries the proxy forwarding table locally built for RT3. The path returned is the label forwarding path to RT3's next hop node RT4, which bypasses RT3. The specific steps are as follows:

a. RT1 swaps label 1003 to out-label 2003 to RT3.

b. RT2 receives the label forwarding packet whose top label of label stack is 2003, and searches for the local Routing Table, the behavior found is to lookup Proxy Forwarding table due to RT3 failure, RT2 pops label 2003.

c. RT2 uses 3004 as the in-label to lookup Proxy Forwarding table, The value of Map Label calculated based on SRGBDiffValue is 2004. and the query result is forwarding the packet to RT4.

d. Then RT2 queries the Routing Table to RT4, using the primary or backup path to RT4. The next hop is RT7.

e. RT2 forwards the packet to RT7. RT7 queries the local routing table to forward the packet to RT4.

f. After RT1 convergences, node-SID 1003 is preferred to the proxy SID implied/advertised by RT2.

Authors' Addresses

Zhibo Hu
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China
Email: huzhibo@huawei.com

Huaimo Chen
Futurewei
Boston, MA,
United States of America
Email: Huaimo.chen@futurewei.com

Junda Yao
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.

Beijing
100095
China
Email: yaojunda@huawei.com

Chris Bowers
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA, 94089
United States of America
Email: cbowers@juniper.net

Yongqing
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
510000
China
Email: zhuyq8@chinatelecom.cn

Yisong
China Mobile
510000
China
Email: liuyisong@chinamobile.com