Workgroup: Network Working Group
Internet-Draft:
draft-hu-spring-segment-routing-proxy-
forwarding-17
Published: 9 February 2022
Intended Status: Standards Track
Expires: 13 August 2022
Authors: Z. Hu, Huawei Technologies
H. Chen, Futurewei
J. Yao, Huawei Technologies
C. Bowers, Juniper Networks
Y. Zhu, China Telecom
Y. Liu, China Mobile

SR-TE Path Midpoint Restoration

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

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.

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 (the Proxy Forwarder) will forward traffic on behalf of the failed node. A router 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 the affected traffic reaches a Proxy Forwarder, it sends the traffic on the post-failure shortest path to the node immediately following the failed node in the segment list.

For a binding segment of a possible failed node, the node advertises the information about the binding segment, including the binding SID and the list of SIDs associated with the binding SID, to its direct neighbors only. Note that the information is not advertised in the network domain.

After the node fails and the IGP converges on the failure, the traffic with the binding SID of the failed node will reach its neighbor having SR Proxy Forwarding capability. Once receiving the traffic, the neighbor swaps the binding SID with the list of SIDs associated with the binding SID and sends the traffic along the post-failure shortest path to the first node in the segment list.

3. Extensions to IGP for Proxy Forwarding

This section describes the semantic of extensions to IGP for advertising the SR proxy forwarding capability of a node in a network domain and the information about each binding segment (including its binding SID and the list of SIDs associated) of a node to its direct neighbors.

3.1. Extensions to OSPF

3.1.1. Advertising Proxy Forwarding

When a node P is able to do SR proxy forwarding for all its neighboring nodes for protecting the failures of these nodes, P advertises its SR proxy forwarding capability in its router information opaque LSA. The LSA contains a Router Functional Capabilities TLV with one bit (called PF bit) set to one indicating that P is capable of doing SR proxy forwarding for all its neighbors.

For a node X in the network, it learns the prefix/node SID of node N, which is originated and advertised by node N. It creates a proxy prefix/node SID of node N for node P if node P is capable of doing SR proxy forwarding for node N. The proxy prefix/node SID of node N for node P is a copy of the prefix/node SID of node N originated by node N, but stored under (or say, associated with) node P. The route to the proxy prefix/node SID is through proxy forwarding capable nodes.

In normal operations, node X prefers to use the prefix/node SID of node N. When node N fails, node X prefers to use the proxy prefix/ node SID of node N. Thus node X will forward the traffic targeting to the prefix/node SID of node N to node P when node N fails, and node P will do a SR proxy forwarding for node N and forward the traffic 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.

If node P can not do a SR proxy forwarding for all its neighboring nodes, but for some of them, then it advertises the node SID of each of the nodes as a proxy node SID, indicating that it is able to do proxy forwarding for the node SID.

A new TLV, called Proxy Node SIDs TLV, is defined for node P to advertise the node SIDs of some of its neighboring nodes. P originates an Extended Prefix Opaque LSA containing this new TLV.

3.1.2. 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 segments, node A advertises an LSA containing the binding (i.e., the binding SID and the list of the segments) in a binding segment TLV. The LSA is advertised only to each of the node A's neighboring nodes. For OSPFv2, the LSA is a opaque LSA of LS type 9 (i.e., a link local scope LSA).

Alternatively, when a protocol (such as PCE or BGP running on a controller) supports sending a binding on a node A to A, this protocol may be extended to send the binding with node A to A's neighbors if the controller knows the neighbors and there are protocol (PCE or BGP) sessions between the controller and the neighbors.

3.2. Extensions to IS-IS

3.2.1. Advertising Proxy Forwarding

When a node P has the capability to do SR proxy forwarding for all its neighbors for protecting the failures of them, P advertises its SR proxy forwarding capability in its LSP. The LSP contains a Router Capability TLV including a SR capabilities sub-TLV. One bit (called PF bit) in the Flags field of the sub-TLV is set to one indicating that P is capable of doing SR proxy forwarding for all its neighbors.

If node P can not do SR proxy forwarding for all its neighbors, but for some of them, then it advertises the node SID of each of the (some) neighbors as a proxy node SID, indicating that P is able to do proxy forwarding for the node SID. P uses the SID/Label Binding TLV defined in [RFC8667] to advertise the node SID of its neighbor.

3.2.2. Advertising Binding Segment

For supporting binding SID proxy forwarding, a new IS-IS TLV, called Binding Segment TLV, is defined. It contains a binding SID and a list of segments (SIDs). This TLV is advertised in Circuit Scoped Link State PDUs (CS-LSP) [RFC7356].

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 N has SRGB = [N000-N999]. 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 for RT3 following 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 all its neighboring nodes, it must advertise this capability. If the PLR can not do SR proxy forwarding for all its neighboring nodes, but for some of them, for example, RT3, then it uses proxy Node SIDs TLV to advertise the prefix-SID learned from RT3. The TLV contains the Sub-TLV/value for the prefix/node SID of RT3 as a proxy SID. When RT3 fails, RT2 needs to maintain the Sub-TLV/value for a period of time. When the proxy forwarding table corresponding to the fault node is deleted, the Sub-TLV/value is withdrawn. The nodes in the network (for example, RT1) learn the prefix/node SID TLV advertised by RT3 and the proxy Node SIDs TLV advertised by RT2. When RT3 is normal, the nodes prefer prefix/node SID TLV. When the RT3 fails, the proxy prefix/node SIDs TLV advertised by RT2 is preferred.

For binding-SID 100, which is associated with segment list {30034, 40045}, RT3 advertises the binding (i.e., 100 bond to {30034, 40045}) to its 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. The entry is used when RT3 fails.

4.2. Building Proxy Forwarding Table

A SR proxy node P needs to build an independent proxy forwarding table for each neighbor N. 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.

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 node N as a backup forwarding entry to the adjacency-SID and Node-SID of node N. When node N fails, the proxy forwarding table needs to be maintained for a period of time, which is recommended for 30 minutes.

Node RT3 in Figure 1 is node N, and node RT2 is node P (PLR). RT2 builds the proxy forwarding table for RT3. RT2 calculates the proxy forwarding table for RT3, 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

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 the node RT3 fails, the proxy forwarding SID implied or advertised by the RT2 is preferred to forward the traffic of the RT1 to the PLR node RT2. Node RT2 acts as a PLR node and uses Binding-SID to query 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.

c. RT2 uses Binding-SID:100 (label 2003 has pop) as the in-label to lookup the Next Label record of the Proxy Forwarding Table, the behavior found is to swap to Segment list {30034, 40045}.

d. RT2 swaps Binding-SID:100 to Segment list {30034, 40045}, and uses the 30034 to lookup the Next Label record of the Proxy Forwarding table again. The behavior found is to forward the packet to RT4.

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

f. RT2 forwards packets to RT7. RT7 queries the local 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. IANA Considerations

7. Acknowledgements

The authors would like to thank Peter Psenak, Acee Lindem, Les Ginsberg, Bruno Decraene and Jeff Tantsura for their comments to this work.

8. References

8.1. 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>.

[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>.

8.2. 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-08, 21 January 2022, <https://www.ietf.org/archive/id/draft-ietf-rtgwg-segment-routing-ti-lfa-08.txt>.

[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-02, 21 January 2022, <https://www.ietf.org/archive/id/draft-ietf-spring-segment-protection-sr-te-paths-02.txt>.

[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-16, 28 January 2022, <https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-16.txt>.

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 proxy forwarding TLV advertised by the RT2 is preferred to direct the traffic of the RT1 to the PLR node RT2. 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