Workgroup: Routing area
Internet-Draft:
draft-ietf-spring-segment-protection-sr-te-paths-05
Published: 27 September 2023
Intended Status: Informational
Expires: 30 March 2024
Authors: S. Hegde, Juniper Networks Inc.
C. Bowers, Juniper Networks Inc.
S. Litkowski, Cisco Systems
X. Xu, Alibaba Inc.
F. Xu, Tencent

Segment Protection for SR-TE Paths

Abstract

Segment routing supports the creation of explicit paths using Adj-Segment-ID (SID), Node-SIDs, and BSIDs. It is important to provide fast reroute (FRR) mechanisms to respond to failures of links and nodes in the Segment-Routed Traffic-Engineered(SR-TE) path. A point of local repair (PLR) can provide FRR protection against the failure of a link in an SR-TE path by examining only the first (top) label in the SR label stack. In order to protect against the failure of a node, a PLR may need to examine the second label in the stack as well, in order to determine SR-TE path beyond the failed node. This document specifies how a PLR can use the first and second label in the SR-MPLS label stack describing an SR-TE path to provide protection against node failures.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

It is possible for a routing device to completely go out of service abruptly due to power failure, hardware failure or software crashes. Node protection is an important property of the Fast Reroute mechanism. It provides protection against a node failure by rerouting traffic around the failed node. For example, the mechanisms described in Loop Free Alternates ([RFC5286]), Remote Loop Free Alternates ([RFC8102]), and [I-D.ietf-rtgwg-segment-routing-ti-lfa] can be used to provide node protection to ensure minimal traffic loss after a node failure.

Section 2 describes problems with SR-TE paths and the need for a specialized mechanism to provide node protection for SR-TE paths. Section 3 describes the solution applied to paths built using Adj-SIDs and Node-SIDs. In order to distinguish the node failures of the segment endpoints (mid points) in an SR-TE path from the usual node protection mechanisms described in various LFA mechansims, this document uses the term Segment Protection.

2. Node Failures Along SR-TE Paths

The topology shown in Figure 1. illustrates a example network topology with Segment Routing enabled on each node.

```
Node          Node          Node          Node          Node
SID:1         SID:2         SID:3         SID:4         SID:5
+-----+  10   +-----+  10   +-----+  10   +-----+  10   +-----+
| R1 |--------| R2 |--------| R3 |--------| R4 |--------| R5 |
+-----+       +-----+       +-----+       +-----+       +-----+
    \                           \          /
     \ 10                        \ 100    / 60
      \                           \      /
       \   +-----+                 +-----+
        +--| R7 |------------------| R8 |
           +-----+   30            +-----+
          / Node                   Node             Label stack:
         /  SID:7                  SID:8            +------------+
   +-----+                         SRGB:            |  1008 (top)|
   | R6 |                          3000-4000        +------------+
   +-----+                                          |  3005      |
   Node                                             +------------+
   SID:6


       * Numbers on the links represent the symmetric link cost
```

Figure 1: Example topology. The segment index for each node is shown in the diagram. All nodes have SRGB = [1000-2000], except for R8 which has SRGB = [3000-4000]. A label stack that represents the path R1->R7->R8->R4->R5 is shown as well.

2.1. Segment protection for explicit paths with Node-SIDs

Consider an explicit path in the topology in Figure 1 from R1->R5 via R1->R7->R8->R4->R5. This path can be built using the shortest paths from R1-to-R8 and R8-to-R5. The label stack to instantiate this path contains two Node-SIDs 1008 and 3005. The 1008 label will take the packet from R1 to R8 via R7 and get popped. The next label in the stack 3005 will take the packet from R8 to the destination R5 via R4. If the node R8 goes down, it is not possible for R7 to perform FRR without examining the second label in the incoming label stack (3005).

Note that in the absence of a failure, R7 does not need to understand the meaning of the second label (3005) in order to perform normal forwarding. However, in order to support segment protection, R7 will need to understand the meaning of label 3005 in order to determine where the packet is headed after R8.

The mechanisms used to detect whether a node failed or a link failed,is outside the scope of this document. The possible options for node failure detection capabilities of a device and resultant forwarding state is described in section 5.2 in [RFC8679] are applicable to this draft as well.

2.2. Segment Protection for Anycast-SIDs

A prefix segment advertised as a Node-SID may only be advertised by one node in the network. Instead, an anycast prefix segment may be advertised by more than one node. In some situations, one can use Anycast-SIDs to construct SR-TE paths that are protected against node failure, without the need for the mechanism described in this document.

```
+----+   10   +----+   10   +----+   10   +----+   10   +----+
| R1 |--------| R2 |--------| R3 |--------| R4 |--------| R5 |
+----+        +----+        +----+        +----+        +----+
    \                           \          / |
     \ 10                        \100   60/  |
      \                           \      /   |
       \   +----+    30            +----+    |
        +--| R7 |------------------| R8 |    |
           +----+                  +----+    |
          /    \                  Anycast    +
         /      \                 SID:100   /
   +----+        \                         /
   | R6 |         \    40          +----+ /60
   +----+          +---------------| R9 |+          Label stack:
                                   +----+           +------------+
                                  Anycast           |  1100 (top)|
                                  SID:100           +------------+
                                                    |  1005      |
                                                    +------------+
     * Numbers on the links represent the symmetric link cost
```

Figure 2: Topology illustrating use of Anycast-SIDs to protect against node failures. All nodes have SRGB = [1000-2000].

An example of this is shown in Figure 2. In this example, R8 and R9 advertise an Anycast-SID of 100. The label stack in this example = [1100, 1005];. The top label (1100) corresponds to the Anycast-SID advertised by both R8 and R9. In the absence of a failure, the packet sent by R1 with this label stack will follow the path from R1->R5 along R1->R7->R8->R4->R5.

If R7 is performing a per-prefix LFA calculation [RFC5286], then R7 will install a backup next-hop to R9 for this Anycast-SID, protecting against the failure of the primary next-hop to R8. This backup path does not pass through R8, so it is would not be affected by a complete failure of node R8. As illustrated by this example, for some topologies segment-protecting SR-TE paths can be constructed through the use of Anycast-SIDs, as opposed to the mechanism described in this document.

2.3. Segment protection for explicit paths with Adj-SIDs

```
                            Adj-SID:
                            R3-R8:9044

Node-         Node          Node          Node          Node
SID:1         SID:2         SID:3         SID:4         SID:5
+-----+  10   +-----+  10   +-----+  10   +-----+  10   +-----+
| R1 |--------| R2 |--------| R3 |--------| R4 |--------| R5 |
+-----+       +-----+       +-----+       +-----+       +-----+
    \                           \          /              |
     \ 10                        \ 100    / 60            | 10
      \                           \      /                |
       \   +-----+                 +-----+              +-----+
        +--| R7 |------------------| R8 |---------------| R9 |
           +-----+   30            +-----+     10       +-----+
          / Node                   Node                 Node
         /  SID:7                  SID:8                SID:9
   +-----+                         SRGB:
   | R6 |                          3000-4000        Label stack:
   +-----+                                          +------------+
   Node                            Adj-SIDs:        |  1003 (top)|
   SID:6                           R8-R4:9054       +------------+
                                                    |  9044      |
                                                    +------------+
                                                    |  9054      |
                                                    +------------+
                                                    |  1005      |
                                                    +------------+
   * Numbers on the links represent the symmetric link cost
```

Figure 3: Explicit path using an Adj-SID.All nodes have SRGB = [1000-2000], except for R8 which has SRGB = [3000-4000].

Consider an explicit path from R1->R5 via R1->R2->R3->R8->R4->R5. This path can be built using a combination of Node-SIDs and Adj-SIDs, as shown in Figure 3. The diagram shows the label stack needed to instantiate this path, as well as several Adj-SIDs advertised by nodes involved in this path. When a packet leaving R1 with this label stack reaches R3, the top label is 9044, which will take the packet to R8. The next-next-hop in the path is R4. To provide protection for the failure of node R8, R3 would need to send the the packet to R4 without going through R8. However, the only way R3 can learn that the packet needs to go to the R4 is to examine the next label in the stack, label 9054. Since R3 knows that R8 has advertised label 9054 as the adjacency segment for the link from R8 to R4, R3 knows that a backup path can merge back into the original explicit path at R4.

3. Detailed Solution using Context Tables

This section provides a detailed description of how to construct node-protecting backup paths for SR-TE paths using context tables. The end result of this description is externally visible forwarding behavior that can be specified as a packet arriving at a PLR with a particular incoming label stack and leaving the PLR on a particular outgoing interface with a particular outgoing label stack. There may be other methods of arriving at the same externally visible forwarding behavior as described in draft [I-D.ietf-rtgwg-segment-routing-ti-lfa]section 6.2. It is not the intent of this document to exclude other methods, as long as the externally visible forwarding behavior is the same as produced by this method.

3.1. Building Context Tables

[RFC5331] introduced the concept of Context Specific Label Spaces and there are various applications making use of this concept.A context label table on a router represents the Label Forwarding Information Base (LFIB) from the point of view of a particular neighbor . Context tables are built by constructing incoming label mappings advertised by the neighbor and the actions corresponding to those labels. The labels advertised by each node are local to the node and may not be unique across the segment routing domain. The context tables are separate tables built on a per-neighbor basis on every node to ensure they represent LFIBs of a particular neighbor.

When a PLR needs to protect an SR-TE path against the failure of a neighbor N, it creates a context table associated with N. This context table is populated with the following segment routing forwarding entries:

- All the Prefix-SIDs of the network. The programmed incoming label map uses the SRGB of N to compute the input label value. The NHLFE (Next Hop Label Forwarding Entry) is then constructed by looking into all the nexthops for the Prefix-SID and choosing a loop-free path as explained in Section 3.2

- All the Adj-SIDs advertised by N. The NHLFE is constructed as explained in Section 3.3

The following section illustrates how the context table is constructed to allow the PLR to provide node-protecting paths for the next-next hops in the topology shown in Figure 1 and Figure 3.

3.2. Segment protection for Node-SIDs

Figure 4 shows the routing table entries on R7 corresponding to the Node-SIDs to reach R1 and R8 for the topology in Figure 1. In the absence of a failure, a packet with a label stack whose top label is 1008 will have its top label popped by R7 (assuming PHP behavior), and R7 will forward the packet to R8. When the interface to R8 is down, the backup next-hop entry is used. R7 will pop the top label of 1008, and use the context table that R7 computed for R8 to evaluate the next label on the stack.

R7's Routing Table (partial)
Transits routes for Node-SIDs for R1 and R8

In label	Outgoing label action
1001	Primary: pop, fwd to R1 Backup: pop, lookup context.r1
1008	Primary: pop, fwd to R8 Backup: pop, lookup context.r8

R7's Context Table for R8 (context.r8, partial)

In label	Outgoing label action
3004	swap 1004, fwd to R1
3005	swap 1005, fwd to R1
3008	drop

Figure 4: Building node-protecting backup paths for SR-TE paths involving Node-SIDs

R7 builds context table for R8 using the following process. R7 computes the mapping of incoming label to Node-SID that R8 expects to see based on the SRGB advertised by R8. In the example in Figure 1, R7 can determine that R8 interprets in incoming label of 3005 as mapping to the the Node-SID for R5.

R7 then computes a loop-free backup path to reach R5 which is node-protecting with respect to the failure of R8. In this example, the backup path computed by R7 to reach R5 without passing through R8 can be achieved forwarding the packet to R1 with a top label of 1005, corresponding to the Node-SID for R5 in the context of R1's SRGB. The loop-free path computation may be based on a mechanism such as LFA, R-LFA, TI-LFA, or constraint based SPF avoiding failure. To populate the context table for R8, R7 maps the out label actions corresponding to the backup path to R5 to the incoming label 3005. This results in the entry for label 3005 shown in context.r8 in Figure 4.

Therefore, when a packet arrives at R7 with label stack = [1008, 3005], and the link from R7 to R8 has recently failed, R7 will use backup next-hop entry for label 1008 in its main routing table. Based on this entry, R7 will pop label 1008, and use context.r8 to lookup the new top label = 3005. R7 will swap label 3005 for 1005 and forward the packet to R1. This will get the packet to R5 on a node protecting backup path.

Note that R7 activates the node-protecting backup path when it detects that the link to R8 has failed. R7 does not know that node R8 has actually failed. However, the node-protecting backup path is computed assuming that the failure of the link to R8 implies that R8 has failed.

3.3. Segment protection for Adj-SIDs

This section gives an example of how to constuct node-protecting backup paths when the SR-TE path uses Adj-SIDs. Figure 5 shows some of the routing table entries for R3 corresponding to the sample network shown in Figure 3. When the top label of the label stack is an Adj-SID, the PLR needs to recognize that in order to provide a node-protecting backup path, it needs to pop the top label and examine the next label in the context of the next-hop router identified by the top label Adj-SID. In this example, when R3 is constructing its routing table, it recognizes that label 9044 corresponds to a next-hop of R8, so it installs a backup entry, corresponding to the failure of the link to R8, when pops label 9044, and then examines the new top label in the context of R8.

R3's Routing Table (partial)
Transit route for Adj-SID

In label	Outgoing label action
9044	Primary: pop, fwd to R8 Backup: pop, lookup context.r8

R3's Context Table for R8 (context.r8, partial)

In label	Outgoing label action
3005	swap 1005, fwd to R4
9054	pop, fwd to R4

Figure 5: Building node-protecting backup paths for SR-TE paths involving Adj-SIDs

R3 constructs its context table for R8 by determining which labels R8 expects to receive to accomplish different forwarding actions. The entry for incoming label 3005 in context.r8 in Figure 5 corresponds to a Node-SID This entry is computed using the methods described in Section 3.2

The entry for incoming label 9054 in context.r8 corresponds to an Adj-SID. R3 recognizes that R8 has advertised this Adj-SID for the link from R8 to R4 in Figure 3. So R3 determines the outgoing label action needed to reach R4 without passing through R8. This can be accomplished by popping the label 9054, and forwarding the packet directly on the link from R3 to R4.

3.4. Segment protection for edge nodes

The segment protection mechanism described in the previous sections depends on the assumption that the label immediately below the top label in the label stack is understood in the IGP domain.When the provider edge routers exchange service labels via BGP or some other non-IGP mechanism the bottom label is not understood in the IGP domain.

The EPE-SIDs as described in [RFC9086] are used to choose egress interface among a set of egress paths. EPE-SID can be a bottom-most label in a SR-TE path. EPE-SIDs are not understood in the IGP domain. In order to support the procedures described in this document, EPE-SIDs should always be added after Anycast-SID for the nodes that advertised the EPE-SIDs. Same EPE-SID should be configured on all these Anycast nodes so that in case of node failure, the traffic is correctly forwarded by the other protector nodes. If a Node-SID is used instead of an Anycast SID, above the EPE-SID in the label stack, if procedures in this document are in use, it may cause packets to be dropped.

The egress node protection mechanisms described in the draft [RFC8679] is applicable to this usecase and no additional changes will be required for SR based networks

3.4.1. Detailed Example for Segment protection for edge nodes

```
        sid:1    sid:2     sid:3       sid:4      sid:5
 1000-2000   1000-2000 1000-2000   1000-2000  1000-2000
   R2:1024    R3:1034   R8:1044     R5:1064
       R4:2014 =========================
   +----+ 10 +----+ 10 +----+  10   +----+ 10 +----+ Primary
   | PE1|----| R2 |----| R3 |-------| R4  |-- | PE2| context 1.1.1.1
   +----+    +----+    +----+       +----+    +----+\sid 10
       \                  \          /               \+-----+
        \ 10               \ 100    / 60             /| CE1 |
         \                  \      /               /  +-----+
          \   +----+         +----+ R4:1054 +-----+sid 10
           +--| R7 |---------| R8 | --------| PE3 |context 1.1.1.1
              +----+    30   +----+         +-----+ Protector
               /   sid:7       sid:8         sid:9   mirror SID 100
              /    1000-2000   3000-4000     1000-2000
             / 10
          +----+
          | R6 |
          +----+
          sid:6
          1000-2000

                  R4's Context Table for PE2 (context.PE2, partial)
   +=============+==============================================+
   | In label    | Outgoing label action                        |
   +=============+==============================================+
   | 1010        | swap 1100(mirror sid), push 1010 fwd to R8   |
   +-------------+----------------------------------------------+

    * Numbers on the links represent the symmetric link cost
```

Figure 6: Node protection for edge nodes Adj-SIDs

The segment protection mechanisms that are described in previous sections depend on the assumption that the label below the top label in the label stack are understood in the IGP domain. If the edge node goes down, the label below the top label representing the edge node could be BGP service label or labels representing other applications. Service mirroring use case is described in [RFC8402] section 5.1. The Customer edges are multi-homed to provider edges and one of the PE's acts in primary role and the other in protector role. The two PEs advertise a context ip address for each customer site and attaches a Anycast-SID to the context. The protector PE advertises a binding sid with M bit set (Mirror-SID)which implies mirroring capability for the context. Protector PE builds the context table for the BGP service labels advertised by the primary PE for the same context. The BGP service resolves on a transport that has stack of labels with context-sid at the bottom of the label stack. Any penultimate node of PE2 builds a context table for PE2 as explained in the section Section 3.1. This context table contains the sid for the context-id and output action is to pop the top label and replace with the Mirror-SID that the protector PE advertised for the context 1.1.1.1. As shown in the example Section 3.4.1 the SID 10 attached to context-id 1.1.1.1 has been programmed in the context.PE2 on the penultimate router R4. The action is to swap 1010 with Mirror-SID 1100 and push 1010 which is PE2's context SID. When packet reaches PE2, it has top label of 1100 which is a Mirror-SID(context label)on PE2 and directs the protector PE to lookup the context table of Primary PE for the BGP service labels.

4. Determining node can be bypassed

In certain scenarios, the node in the label stack may represent an important function such as firewall filter which must be performed. Bypassing such a functionality may cause major security issues. When segment protection mechanisms described in this document are applied, it's possible that if the firewall goes down, traffic is re-routed via the next label in the stack. There are multiple ways this problem could be solved.

The procedures described in this document should be optional and should be enabled when devices are configured to apply the procedures and examine next label in the stack. The feature should be controllable on a per neighbor granularity. When certain devices offer a critical function, the neighbors of the devices may disable the segment protection for this particular neighbor providing critical functions.

IGP protocol extensions are proposed in [I-D.li-rtgwg-enhanced-ti-lfa] which define a "no bypass" flag for the SIDs. The nodes that indicate critical functions may advertise SIDs with "NB" bit set. Segment protection procedures described in this document should not be applied on these SIDs and in case of failure either link protecting backup paths can be programmed or packet can be dropped with no protection.

5. Hold timers for Node-SID/Prefix-SIDs and Adj-SIDs

SR-TE paths may be computed by a controller or by the head-end router. When there is a node failure in the network, the controller or head-end router has to learn about the failure, recompute the label stacks of any affected SR-TE paths, and get the new label stacks programmed into the forwarding plane of the head-end router. This process may be slow compared to the speed with which routers in the network react to the event. After learning about a node failure, the non-PLR routers in the network will no longer be able to compute a path to reach the failed node. If no special precautions are taken, these non-PLR routers will remove the forwarding entries corresponding the Node-SID and Prefix-SIDs advertised by the failed node. If the head-end router is still sending traffic with that Node-SID/Prefix-SID in the stack, traffic can be blackholed at a non-PLR router. In this case, the node-protection FRR mechanisms do not bring full benefit.

In order to solve the above problem, hold timers are recommended. The hold-timer corresponds to the maximum time that a combination of controller and head-end router or a head-end router alone takes to compute and install label stacks corresponding to a new SR-TE paths in the event of a node failure. The hold times should be applied to forwarding entries for Node-SIDs and Prefix-SIDs that are advertised by a single node in the network. If the Node-SID or Prefix-SID becomes unreachable, the event and resulting 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. The traffic will continue to follow the previous path and get FRR protection on the PLR.

A route corresponding to a global Adj-SID advertised by a node that becomes unreachable should also be left in the forwarding table for the duration of the hold-timer.

The node-protecting backup forwarding entry on the PLR corresponding to the local Adj-SID from the PLR to the failed node should also be left in the forwarding table for the duration of the hold-timer.

The Node-SID/Prefix-SID becoming unreachable is not a single event in IGP. This unreachability is recognized by combining multiple link-down events from the neighbouring nodes. If these link-down events arrive at different time, the remote nodes converge to a state corresponding to the link-down events. When the Node-SID unreachability is finally recognized in a remote node, the previous state may be that of a link-down event. This could lead to hold-down states that are undesirable.

When a network event such as link-up/link-down or metric change event is received, IGP schedules SPF computation. A small configurable delay called spf-delay can be enabled, which will schedule the SPF after the spf-delay time on receiving the first event. In case of a node going down, the spf-delay time coupled with fast-flooding can help to accumulate link-down events reported by all neighbors in one single SPF. This mechansim is on best effort basis and does not guarantee that all link-down events are accumulated before SPF is triggered. If there are flooding delays, the SPF might get triggered before receiving all events related to node going down.

The protection mechanisms are expected to work well when there is single network event. If there are simultaneous network events, the protection mechansims do not guarantee that the traffic will not be impacted. When a node is running hold-down timer and is holding Node-SID and other routes in forwarding plane, if there is another link-down/link-up event or metric change event is received, the hold-down should be aborted and the global convergence procedures should be excecuted.

5.1. Interaction with micro-loop avoidance

During network convergence, the micro-loop avoidance mechansims as described in [I-D.bashandy-rtgwg-segment-routing-uloop] may be applied.For the failed node, all the nodes in the network should consistently detect the failure and maintain the pre-failure shortest path in the forwarding plane so that the traffic can follow pre-failure shortest path and take the node-protecting backup path at the PLR of the failed node.

6. Optimization Considerations

The solution described in this document requires that a PLR build a context table for each neighbor for which node-protection is desired. The context table for each protected neighbor needs to contain route entries for all of the Prefix-SIDs in the network, as well as the route entries corresponding to the Adj-SIDs advertised by the protected neighbor. Although the scale of IGP domain is limited, this may result in considerable additional memory consumption on the routers. It is possible to take advantage of an optimization that allows the PLR to avoid creating context-tables when all of the nodes in the network advertise the same Segment Routing Global Block (SRGB) and all Adj-SIDs in the network are advertised as global Adj-SIDs. In this case, all labels in the stack representing an SR-TE path are globally unique.Protection for node failure cases in such a deployment can be achieved by doing a lookup of the first label and potentially a second lookup of the second label using a common route table with primary and backup entries for all Prefix-SIDs as well as for all of the global Adj-SIDs.

The primary route entries for global Adj-SIDs not advertised by the PLR will be the shortest path to the node advertising the global Adj-SID. The backup route entries for these global Adj-SIDs will generally correspond to the node-protecting backup path to the node advertising the global Adj-SID. However, for a global Adj-SID advertised by the direct neighbor of the PLR the node-protecting backup route entry will correspond to the backup path to the node on the far end of the Adj-SID.

With the common route table constructed in this manner, when the PLR receives a packet whose first label is a global Adj-SID advertised by the failed neighbor of the PLR, the lookup of the first label will produce the correct backup path directly. When the PLR receives a packet whose first label is the Node-SID of the failed neighbor,or an Adj-SID advertised by the PLR corresponding to the failed neighbor, the route entry will instruct the PLR to lookup the second label using the common route table. Finally, when the PLR receives a packet whose first label is a global Adj-SID or a Node-SID advertised by a node which is neither the PLR nor the failed neighbor, then the usual link-protecting backup path will be produced based on a lookup of the first label only.

6.1. Segment Protection Example with Common SRGB

```
Node               Node          Node         Node             Node
sid:1000           sid:1001      sid:1002     sid:1003      sid:1004
+----+2001 2100+----+2102 2201+----+2203   2302+----+2304    2403+----+
| R0 |---------| R1 |---------| R2 |-----------| R3 |------------| R4 |
+----+   1     +----+    1    +----+      1    +----+     1      +----+
    \ 2005                         \ 2206        / 2306          2407 |
     \                              \           /                     |
      \ 1                            \ 10      / 6                  1 |
       \                              \       /                       |
        \                         2602 \     / 2603              2704 |
         \ 2500+----+ 2506        2605+----+2607             2706+----+
          +----| R5 |-----------------| R6 |---------------------| R7 |
               +----+    3            +----+            1        +----+
               Node                          Node                  Node
               sid:1005                      sid:1006           sid:1007
```

* Numbers on the links represent the symmetric link cost
* All nodes have SRGB = [400000-405000] size 5000

R2's Routing Table (partial)

In label	Outgoing label action
4001003	Primary: pop, fwd to R3 Backup: pop, lookup ilm table or ip table based on BOS bit
4001007	Primary: swap 401007, fwd to R6 Backup: Swap 401007, Push 401005(top),fwd R1
4002203	Primary: pop, fwd to R3 Backup: pop, lookup ilm table or ip table based on BOS bit

Label Stack 1:

```
+-------------+
|4001003 (top)|
+-------------+
|   4001007   |
+-------------+
```

Label Stack 2:

```
+-------------+
|4001003 (top)|
+-------------+
|   4001007   |
```

```
+--------------+
```

Figure 7: Common SRGB

The diagram Figure 7 shows an example where optimized Segment Protection mechanism is deployed. All the nodes have a common SRGB of 400000 to 4005000. The Node-SIDs are in the range 1001, 1002 etc and the global Adj-SIDs are in the range 2001, 2005 and so on. R2's partial ILM table consisting of primary and backup nexthops is also shown in the diagram. Node-SID of R3 which is represented by label 4001003 has a primary nexthop pointing to R3 and backup nexthop which pops the label and looks up ILM table with next label in the packet. For Example consider a path from R0 to R7 with a label stack consisting of 4001003 and 4001007. When the node R3 fails, R2 which is the PLR, will pop the label 4001003 and lookup for next label in the same table. Next label in this example is 4001007. Based on the primary nexthop for 4001007, traffic is forwarded to R6. Another example label stack consists of global Adj-SID of 4002203 (Adj-SID from R2->R3). As shown in the partial ILM table on R2, 4002203 also has a backup nexthop which pops the label and looks-up next label in the packet.On R3's failure, traffic will get forwarded via R6.

7. Alternate path protection mechanisms

The current document describes protection mechanisms when nodes that are mid-points in an SR-TE path fail. The solution described here focuses on triggering protection locally on the Point of local repair. There are other path protection mechanisms which provide end-to-end path protection. In end-to-end path protection mechanism, path liveness is monitored using liveness detection protocols such as S-BFD [RFC7880]. A backup path is pre-programmed on the head end of the SR-TE path. When the S-BFD running on a particular SR-TE path detects path failure, the head end of SR-TE path switches the traffic from primary path to backup path. The granularity of failure detection timers configured on the headend depend on the scale of SR-TE tunnels on the device and also capabilty of the device to support fast switchover.

8. Operational Considerations

The procedures described in this document should be configurable and applied only when enabled explicitly. In order to satisfy scenarios described in Section 4, the feature should be controllable on the per neighbor basis. The optimisation procedures described in Section 6, should be applied only when the entire network has a common SRGB and all nodes advertise global Adj-SIDs. This optimization should be applied based on explicit configuration.

9. Security Considerations

The procedures described in this document will in most common cases be deployed inside a single ownership IGP domain. No new security risks are exposed due to the procedures described in this document. The security considerations for SR-MPLS with label stacking is described in detail in [RFC8402] are applicable for this document as well. This document introduces the context table lookup for the labels in the label stack. As described in [RFC8402] MPLS packet filtering at the boundaries ensures the operations on the MPLS labels inside the domain is safe includingcontext table lookup operation. The security procedures applicable to IGP protocols are also applicable to segment routing extensions as described in [RFC8667] and [RFC8665] and ensure required protection for the segment protection procedures described in this document.

10. IANA Considerations

11. Acknowledgments

The authors would like to thank Peter Psenak, Bruno Decraene, Alexander Vainshtein and Huzibo, Dhruv Dhody Ketan Talaulikar for their review and suggestions. Thanks to Bharath R for suggesting Node-SID hold down mechanisms.

12. References

12.1. Normative References

[RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <https://www.rfc-editor.org/info/rfc5286>.

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, DOI 10.17487/RFC5331, August 2008, <https://www.rfc-editor.org/info/rfc5331>.

[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>.

12.2. Informative References

[I-D.bashandy-rtgwg-segment-routing-uloop]
Bashandy, A., Filsfils, C., Litkowski, S., Decraene, B., Francois, P., and P. Psenak, "Loop avoidance using Segment Routing", Work in Progress, Internet-Draft,

draft-bashandy-rtgwg-segment-routing-uloop-15, 18 June 2023, <https://datatracker.ietf.org/doc/html/draft-bashandy-rtgwg-segment-routing-uloop-15>.

[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.li-rtgwg-enhanced-ti-lfa] Li, C., Hu, Z., Zhu, Y., and S. Hegde, "Enhanced Topology Independent Loop-free Alternate Fast Re-route", Work in Progress, Internet-Draft, draft-li-rtgwg-enhanced-ti-lfa-08, 4 May 2023, <https://datatracker.ietf.org/doc/html/draft-li-rtgwg-enhanced-ti-lfa-08>.

[RFC7880] Pignataro, C., Ward, D., Akiya, N., Bhatia, M., and S. Pallagatti, "Seamless Bidirectional Forwarding Detection (S-BFD)", RFC 7880, DOI 10.17487/RFC7880, July 2016, <https://www.rfc-editor.org/info/rfc7880>.

[RFC8102] Sarkar, P., Ed., Hegde, S., Bowers, C., Gredler, H., and S. Litkowski, "Remote-LFA Node Protection and Manageability", RFC 8102, DOI 10.17487/RFC8102, March 2017, <https://www.rfc-editor.org/info/rfc8102>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[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>.

[RFC8679] Shen, Y., Jeganathan, M., Decraene, B., Gredler, H., Michel, C., and H. Chen, "MPLS Egress Protection Framework", RFC 8679, DOI 10.17487/RFC8679, December 2019, <https://www.rfc-editor.org/info/rfc8679>.

[RFC9086] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Patel, K., Ray, S., and J. Dong, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing BGP Egress

Peer Engineering", RFC 9086, DOI 10.17487/RFC9086, August 2021, <https://www.rfc-editor.org/info/rfc9086>.

Authors' Addresses

Shraddha Hegde
Juniper Networks Inc.
Exora Business Park
Bangalore 560103
KA
India

Email: shraddha@juniper.net

Chris Bowers
Juniper Networks Inc.

Email: cbowers@juniper.net

Stephane Litkowski
Cisco Systems

Email: slitkows.ietf@gmail.com

Xiaohu Xu
Alibaba Inc.
Beijing
China

Email: xiaohu.xxh@alibaba-inc.com

Feng Xu
Tencent
China

Email: oliverxu@tencent.com