Routing area
Internet-Draft
Intended status: Informational
Expires: January 31, 2021

S. Hegde
C. Bowers
Juniper Networks Inc.
S. Litkowski
Cisco Systems
X. Xu
Alibaba Inc.
F. Xu
Tencent

July 30, 2020

Node Protection for SR-TE Paths
draft-hegde-spring-node-protection-for-sr-te-paths-07

Abstract

Segment routing supports the creation of explicit paths using adjacency-sids, node-sids, and binding-sids. 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 label stack describing an SR-TE path to provide protection against node failures.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 31, 2021.

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.bashandy-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 adjacency-sids and node-sids.

2. Node Failures Along SR-TE Paths

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

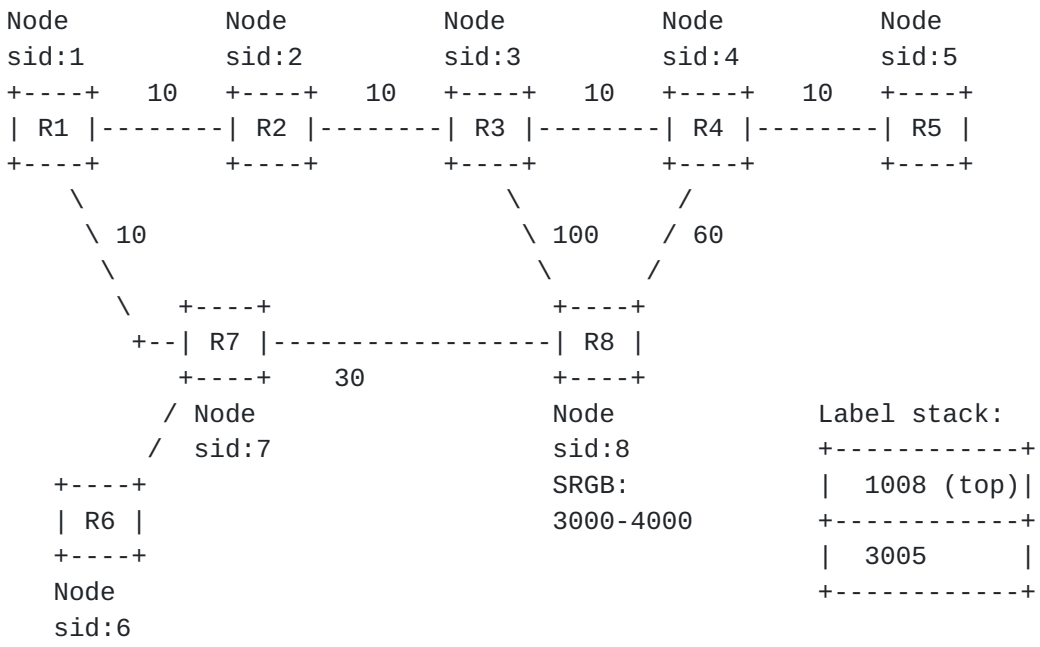

Figure 1: Example topology. The segment index for each node is shown in the diagram. All nodes have SRGB = [1000-2000], except for R8 which has SRGB = [3000-4000]. A label stack that represents the path R1->R7->R8->R4->R5 is shown as well.

2.1. Node protection for node-sid explicit paths

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 node 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. Node-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.

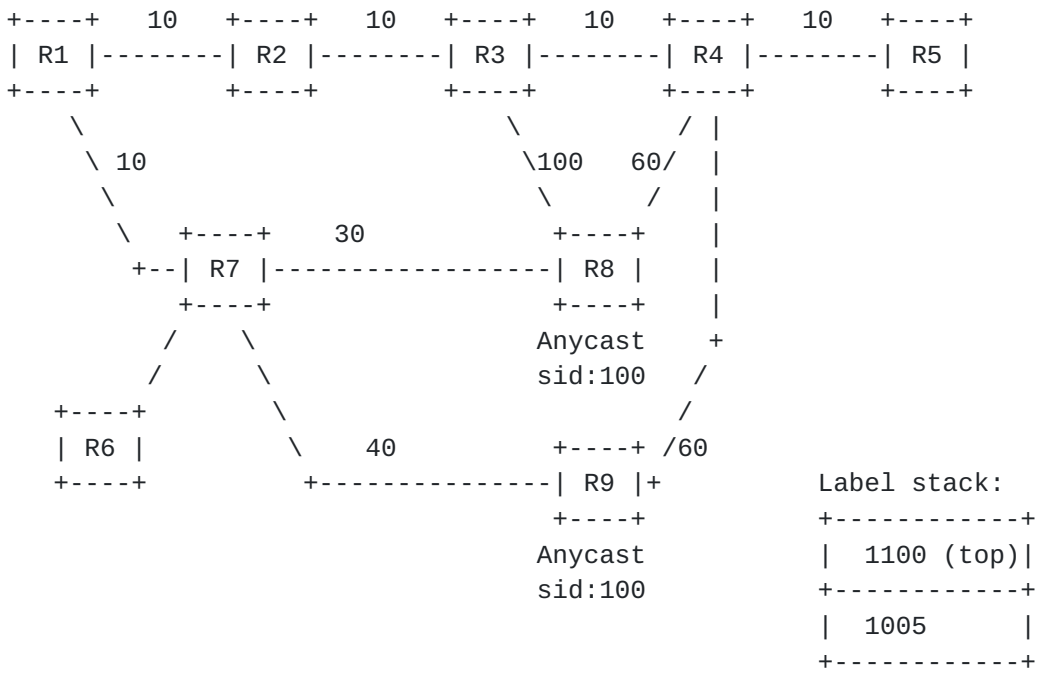

Figure 2: Topology illustrating use of anycast-sids to protect against node failures. All nodes have SRGB = [1000-2000].

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 node-protecting SR-TE paths can be constructed through the use of anycast SIDs, as opposed to the mechanism described in this document.

2.3. Node-protection for adj-sid explicit paths

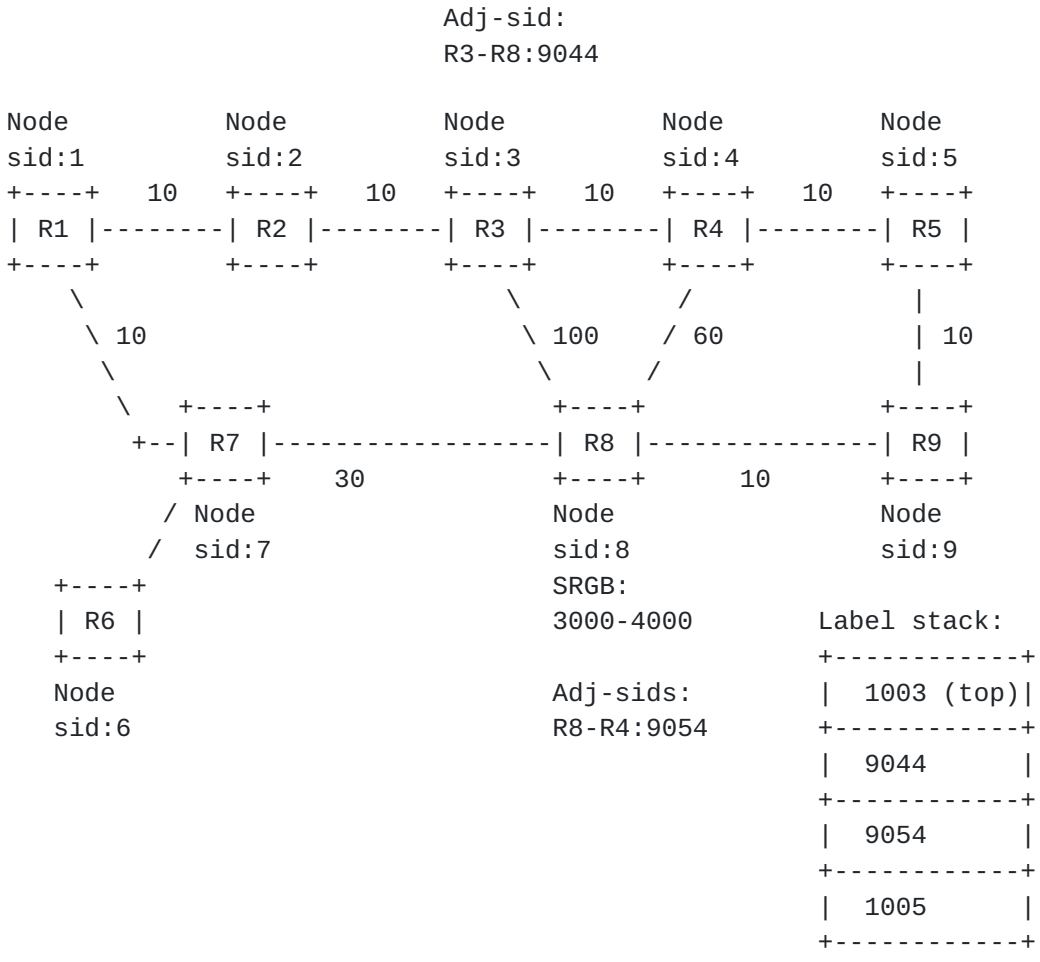

Figure 3: Explicit path using an adjacency sid. All nodes have SRGB = [1000-2000], except for R8 which has SRGB = [3000-4000].

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 adjacency sids, as shown in Figure 3. The diagram shows the label stack needed

to instantiate this path, as well as several adjacency 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.bashandy-rtgwg-segment-routing-ti-lfa]. 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 Adjacency 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. Node 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 showin 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. Node protection for adjacency SIDs

This section gives an example of how to constuct node-protecting backup paths when the SR-TE path uses adjacency 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 adjacency 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 adjacency 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 adjacency 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 adjacency SID. R3 recognizes that R8 has advertised this adjacency 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. Node protection for edge nodes

The node 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 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

4. Hold timers for Node-SID/Prefix-SIDs and Adjacency-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 single node in the network. If the Node-SID or Prefix-SID becomes unreachable, the event and resulting forwarding changes should not 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 adjacency 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 adjacency-SID from the PLR to the failed node should also be left in the forwarding table for the duration of the hold-timer.

4.1. 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.

5. 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 adjacency SIDs advertised by the protected neighbor. This may result in considerable additional memory consumption. 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 SRGB and all adjacency SIDs in the network are advertised as global adjacency 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. 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 procedures applicable to IGP protocols will provide the desired protection.

7. IANA Considerations

8. Acknowledgments

The authors would like to thank Peter Psenak, Bruno Decraene, Alexander Vainshtein and Huzibo for their review and suggestions.

9. References

9.1. 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>.

9.2. Informative References

[I-D.bashandy-rtgwg-segment-routing-ti-lfa] Bashandy, A., Filsfils, C., Decraene, B., Litkowski, S., Francois, P., daniel.voyer@bell.ca, d., Clad, F., and P. Camarillo, "Topology Independent Fast Reroute using Segment Routing", draft-bashandy-rtgwg-segment-routing-ti-lfa-05 (work in progress), October 2018.

[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", draft-bashandy-rtgwg-segment-routing-uloop-09 (work in progress), June 2020.

[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>.

[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>.

[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>.

Authors' Addresses

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