Network Working Group
Internet Draft
Intended status: Informational
Expires: 20 September 2025

Y. Liu
W. Jiang
China Mobile
C. Lin
New H3C Technologies
X. Geng
Huawei Technologies
Y. Liu
ZTE

20 March 2025

Operational Guidance for Protection mechanisms in SRv6 Networks
draft-liu-srv6ops-sr-protection-03

Abstract

This document describes the Operational Guidance for protection of Segment Routing Over IPv6 (SRv6) networks.

Copyright Notice

Table of Contents

1. Introduction

Segment Routing (SR) [RFC8402] leverages the source routing paradigm. An ingress node steers a packet through an ordered list of instructions, called "segments".

SR can be instantiated on the MPLS data plane (MPLS-SR) and the IPv6 data plane (SRv6). On the MPLS-SR data plane, a segment is encoded as an MPLS label, and an ordered list of segments is encoded as a stack of labels. On the SRv6 data plane, a segment is encoded as an IPv6 address (SRv6 SID) [RFC8986], and an ordered list of segments is encoded as an ordered list of SRv6 SIDs in the SR header (SRH) [RFC8754].

This document describes the common failure scenarios and protection mechanisms in SRv6 networks. Then Operational Guidance for protection of SRv6 networks are proposed.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1.2. Terminology

BE: Best Effort

TE: Traffic Engineering

MPLS-SR: Segment Routing over MPLS

SRv6: Segment Routing over IPv6

G-SRv6: Generalized SRv6 Network Programming

2. Forwarding over SRv6 Network

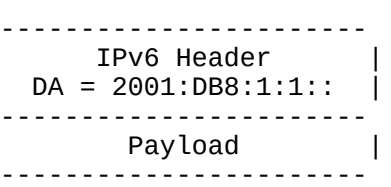
In an SRv6 network, the ingress node steers a packet through an ordered list of segments, which instructs the SRv6 network to

forward the packet via a specific path to the egress node. The forwarding path is either an SRv6 BE path or an SRv6 TE path.

2.1. SRv6 BE Path

An SRv6 BE path is based on shortest path forwarding.

On the SRv6 data plane, the ingress PE encapsulates the payload in an outer IPv6 header where the destination address is the SRv6 Service SID provided by the egress PE. The underlay P nodes between the PEs only need to perform plain IPv6 shortest path forwarding.



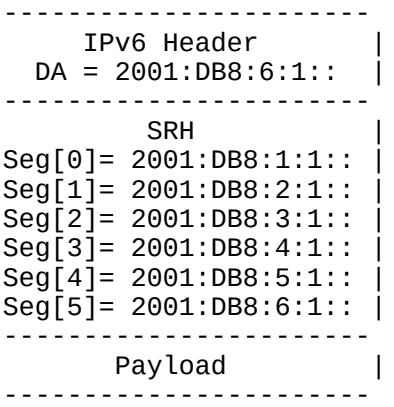
Ingress PE ---> P nodes ---> Egress PE

Figure 1: Forwarding over SRv6 BE

2.2. SRv6 TE Path

In an SRv6 TE path, the ingress PE steers the traffic flow into an SR Policy [RFC9256] with an ordered list of segments associated with that SR Policy. The underlay P nodes whose SIDs are part of the segment list are called endpoint nodes. They will be involved in the forwarding path and execute the function associated with the SID.

On the SRv6 data plane, the ingress PE encapsulates the payload packet in an outer IPv6 header with the Segment Routing Header (SRH) carrying the segment list of the SR policy.

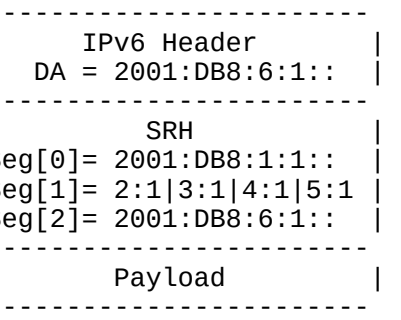


Ingress PE ---> P nodes ---> Egress PE

Figure 2: Forwarding over SRv6 TE

If Compressed Segment List encoding is enabled in the SRv6 network [I-D.ietf-spring-srv6-srh-compression], the segment list in the SRH will be encoded in the compressed way. The compressed SRv6 Segment-List encoding can optimize the packet header length by avoiding the repetition of the Locator-Block and trailing bits with each individual SID.

The G-SRv6 mechanism will be used as an example for the encoding of SRv6 TE path in this document. Figure 3 shows the encapsulation of packet using the G-SRv6 mechanism.



Ingress PE ---> P nodes ---> Egress PE

Figure 3: Forwarding over G-SRv6 Encoded TE

3. Considerations for Protection Mechanisms

Two main categories of protection mechanism in SRv6 networks are described in this section: path protection and egress protection.

Path protection works when the failure occurs along the forwarding path, including SRv6 BE paths and SRv6 TE paths. Path protection is further divided into local protection, which is performed by the node adjacent to the failed component, and end-to-end protection, which is performed by the ingress PE node.

In multi-homed scenarios, egress protection works instead when the failure occurs on the egress PE node, and traffics will be forwarded to another backup Egress PE node. Egress protection can be performed by either local repair or ingress node switchover.

The corresponding liveness check mechanisms are also described along with the protection mechanisms.

3.1. Considerations for Path Protection

3.1.1. Local Proctection

Local protection is performed by the node adjacent to the failed component using fast-reroute techniques [RFC5286] [RFC5714]. The common method of local repair is to provide a repair path for the destination avoiding the failed component.

[I-D.ietf-rtgwg-segment-routing-ti-lfa] describes the Topology Independent Loop-free Alternate Fast Re-route technology (TI-LFA) using Segment Routing, which is able to provide a loop free backup path irrespective of the topologies used in the network. For each

destination in the network, TI-LFA pre-installs a backup forwarding entry for each protected destination ready to be activated upon detection of the failure of a link used to reach the destination.

On the SRv6 data plane, the TI-LFA repair path is encoded as an SRv6 SID list, and encapsulated in the SRH along with an outer IPv6 header.

When local protection occurs, it changes the packet forwarding path, typically applicable in scenarios with low SLA requirements. When configuring local protection, the following factors should be considered:

- The scope of prefixes to be protected, to avoid calculating backup paths for all IGP prefixes and to prevent excessive resource consumption.
- According to the topology plan, configure local protection reasonably. If the network topology is loop-free, there is no need to enable the TI-LFA function.
- Control the interfaces to be protected, enabling local protection only on specific interfaces.
- When multiple backup paths exist, set policies to control the selection of the backup path based on network planning.
- Set a reasonable delay time based on the network scale to avoid temporary congestion; it is usually recommended to set the time between 5 to 30 seconds.
- In cases where multiple points of failure may exist on local links, consider configuring the FRR function with shared risk groups.

3.1.2. Liveness Check for Local Protection

In order to perceive the failures of links and neighbors, a node should monitor the liveness of its adjacent components.

[RFC5880] and [RFC7880] provide widely used mechanisms for liveness check, called Bidirectional Forwarding Detection (BFD) and Seamless Bidirectional Forwarding Detection (S-BFD).

BFD can be associated with the interface state to detect the failure of directly-connected links. Two adjacent nodes may establish BFD or S-BFD sessions between each other, and send BFD control packets to monitor the liveness of each other. In another way, a node may send

BFD echo packets to all the neighbors, and they will reflect the packets back, without establishing BFD sessions.

Other OAM methods, such as Ping, TWAMP or STAMP, may also be used for liveness check for local protection.

When deploying local protection mechanisms, the following factors should be considered:

- Detection Requirements If the detection requirement is less than 100ms, it is recommended to configure BFD (Bidirectional Forwarding Detection).
- Local Interface State Detection Local protection typically monitors the state of the egress interface. This can be achieved by configuring BFD or other OAM mechanisms tied to the interface state.
- BFD Session Optimization If BFD is already configured for IGP neighbor detection, the same BFD session can be used to monitor the next-hop state. This eliminates the need for a separate BFD session for the primary path, reducing the total number of BFD sessions.
- Alternative Detection Protocols In scenarios where TWAMP (Two-Way Active Measurement Protocol) or STAMP (Simple Two-Way Active Measurement Protocol) is already deployed, these protocols can be used not only for link quality detection but also for reachability verification. If performance requirements are not stringent, configuring BFD may not be necessary.

3.1.3. Micro-Loop Avoidance

When a component fails or comes back up, the topology is changed. The routing convergence happens in each node at different times and during a different lapse of time. These transient routing inconsistencies may cause micro-loops.

[I-D.bashandy-rtgwg-segment-routing-uloop] provides a mechanism leveraging segment routing to ensure loop-freeness during the IGP reconvergence process, which relies on the temporary use of SR policies ensuring loop-freeness over the post-convergence paths from the converging node to the destination.

On the SRv6 data plane, the loop-free post-convergence path is encoded as an SRv6 SID list, and encapsulated in the SRH along with an outer IPv6 header.

To effectively configure micro-loop prevention, the following guidelines should be considered:

- Default Duration It is recommended to set the micro-loop prevention duration to 5 seconds as a default value.

- Adjusting Duration for Larger Networks: As the network scale increases, the overall convergence time may also increase. In such cases, the duration of micro-loop prevention can be adjusted accordingly to align with the extended convergence time.

- Local vs. Remote Micro-Loop Prevention: Micro-loops can be caused by either local link changes or remote link changes. Depending on the specific scenario, operators can choose to enable.

3.1.4. End-to-End Protection

End-to-end protection lets the ingress PE node be in charge of the failure recovery. The ingress node should steer the flow from the failed path into another alive path.

In the case of SRv6 TE path, the SR Policy itself allows for multiple candidate paths, of which at any point in time there is a single active candidate path that is provisioned in the forwarding plane and used for traffic steering [RFC9256]. The candidate path with highest preference is selected as the primary path, and the candidate path with second highest preference can be selected as the hot-standby backup. When the primary candidate path fails, switchover to the backup candidate path can be triggered by fast re-route mechanism.

If all the candidate paths fail, the ingress node may use SRv6 BE path for best-effort forwarding as a backup.

To effectively configure End-to-End prevention for SRv6 Policy, the following guidelines should be considered:

- Hot Standby for SRv6 Policy: Enable the hot standby feature for SRv6 Policy to pre-configure the suboptimal candidate paths in the forwarding plane.

- o End-to-End Detection: Ensure that End-to-End detection is enabled for both the primary path and the backup path.

- o Reversion Delay: Configure a reasonable reversion delay to avoid switching back to the primary path too quickly after fault recovery. The recommended value is 5 seconds.

- o Multiple Suboptimal Paths for Sequential Backup: In real deployments, multiple candidate paths may fail simultaneously. It is recommended to configure multiple suboptimal candidate paths to form a sequential backup, enhancing performance in scenarios with multi-point failures.

- o Escape to Best Effort (BE) Path : If all SRv6 Policy paths fail, configure whether to escape to a Best Effort (BE) path based on business requirements.

3.1.5. Liveness Check for End-to-End Protection

It is essential that the ingress PE node should check the end-to-end liveness of paths, including primary path and backup path. So that the ingress PE node can perceive the path failure and then trigger the switchover.

In the case of SRv6 TE path, BFD or S-BFD can be used to monitor the liveness of SR Policy at the level of segment list. If all the BFD sessions associated with segment lists in a candidate path are down, the candidate path is deemed to be failed. If all the candidate paths are failed, the SR Policy is deemed to be failed.

Moreover, If the SRv6 TE path is strict (every hop along the path appearing in the SID list), the reverse path of the BFD packets should be the same with the forward path. Otherwise, the failure in the reverse path may cause the misjudgement of the liveness of SR Policy. To achieve the consistence of forward path and reverse path, the egress node should be instructed to use specific path to send packets back to the ingress node.

Other OAM methods, such as Ping, TWAMP or STAMP, may also be used for liveness check for end-to-end protection, which will not be enumerated here in detail.

Local protection and end-to-end protection may both be used in the same SR network. Since the speed of failure detection for local protection is faster than end-to-end protection, local protection usually performs the local repair in advance, which allows the path to remain alive. In this case, the ingress node will not perceive the failure and does not need to trigger end-to-end protection.

To effectively configure Liveness Check for End-to-End prevention in SRv6 Policy, the following guidelines should be considered:

Consistent Round-Trip Path for Strict Paths When the SRv6 Policy path is a strict path, it is recommended to enable the consistent round-trip path feature for detection packets. This prevents the backup path from being mistakenly marked as DOWN due to inconsistent paths.

- Detection Time for Inconsistent Paths If the round-trip path for detection packets is inconsistent, ensure that the detection timeout for the backup path is longer than that of the primary path. For example, when using BFD, configure:

  Primary path timeout: 50ms

  Backup path timeout: 150ms

- No-Bypass for Detection Packets When local protection is enabled on intermediate devices, it may prevent the End-to-End detection from marking the path as DOWN. To address this, enable the no-bypass feature for detection packets, ensuring they do not take the protection path at intermediate nodes.

- Encapsulation Mode for Detection Packets Configure the encapsulation mode for detection packets, such as Encap mode or Insert mode. For reduced payload overhead, use Insert mode. For full compatibility with detection protocols, use Encap mode.

3.2. Considerations for Egress Protection

If the failure occurs on the egress PE node, the TI-LFA or the hot-standby backup candidate path of SR Policy will not work. To provide protection, the packet should be forwarded to another backup Egress PE node, if it exists.

3.2.1. Local Repair

In the case of egress PE node failure, the local repair node, which is usually the penultimate hop on the SRv6 path, should forward packet to another Egress PE node. If a failure occurs on the link between PE and CE, that PE should work as the local repair node and forward packet to another Egress PE node. That mechanism is beyond the scope of this document.

On the SRv6 data plane, [I-D.ietf-rtgwg-srv6-egress-protection] provides a method to use Mirror SID for egress protection. The

Mirror SID is configured on the backup egress PE to protect the primary egress PE, and it will be used by the repair node to encode the segment list of repair path.

3.2.2. Ingress Node Switchover

If there are multiple egress PE nodes, the ingress PE node receives all their advertisements of the same service, and builds paths for each of them respectively. The ingress PE node may use Fast Reroute (FRR) for these different paths. When the primary egress PE node fails, the ingress node steers the flow to the path belonging to another egress PE node for protection.

BFD can be used to monitor the liveness of the service SID, locator or interface address of the egress PE node. If the BFD session is down, the egress PE node is deemed to be unreachable. The ingress PE node may also use the IGP routes of the locator or interface address of the egress PE node to evaluate if that egress PE node is alive. The IGP convergence is slower than BFD, but it can be useful in some cases. For example, in the BGP-based VPN service network, the ingress node switchover based on IGP convergence of egress PE routes is usually faster than BGP convergence of VPN routes.

Egress protection and path protection may both be used in the same SR network. Among the different paths to the same egress PE node and the paths to different egress PE nodes, one is selected as the primary path and others are used as backup. The priorities of multiple backup paths may be decided by the egress-node-first strategy or the TE-first strategy.

By the Egress-node-first strategy, paths to the primary egress PE nodes are prioritized. For example, if a failure occurs on the primary path, the ingress PE node will select another path still leading to the primary egress PE nodes. Unless all the paths to the primary egress PE node are failed, the ingress PE node would use the path to the backup egress PE node.

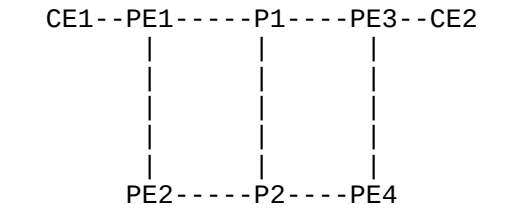
By the TE-first strategy, SRv6 TE paths to any egress PE node have higher priorities than SRv6 BE paths. For example, if a failure occurs on the primary path and there is no other alive SRv6 TE path to the primary egress PE node, the ingress node will select an SRv6 TE path to the backup egress PE node, rather than an SRv6 BE path still leading to the primary egress PE node.

4. Operational Guidance

This section will introduce the operational guidances of protection for SRv6 networks. Section 4.1 describes the single-homed scenario,

and Section 4.2 describes the multi-homed scenario. In the following scenarios, we assume that both SRv6 BE paths or SRv6 TE paths are used in the same network to steer traffics with different requirements.

4.1. Single-homed Scenario



In the single-homed scenario, the combination of following mechanisms can be used for the protection of SR network:

- TI-LFA
- Multiple Candidate Paths
- BE as Backup for TE

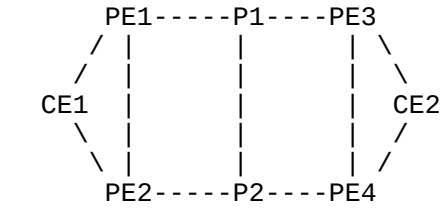
For traffics steered by SRv6 BE paths, protection is performed locally by the node adjacent to the failed component using TI-LFA mechanism. BFD for links and neighbors are used as triggers of TI-LFA.

For traffics steered by SRv6 TE paths, in some cases, end-to-end protection (switchover to backup candidate path) is preferred over local protection (TI-LFA) due to SLA requirements. BFD or S-BFD is enabled to monitor the liveness of candidate paths. If the main candidate path is down, the SR Policy will switch to the backup candidate path. In some other cases, local protection is preferred over backup candidate path due to the requirements of traffic restoring time, like less than 200ms.

- In the cases with high SLA requirements: For strict SRv6 TE path, TI-LFA is not used along the path. For loose SRv6 TE path, local protection only works for the loose part on the path.
- In the cases with fast traffic restoring requirements: TI-LFA preforms local protection in advance. The ingress node will perceive the failure on the main candidate path after routing convergence, and then switch to backup candidate path.

In addition, SRv6 BE path can be used as a final backup for SRv6 TE path in case of multi-point faults. When all candidate paths of an SR Policy are failed, the traffics will be switched to the SRv6 BE path instead of being dropped. Except for the cases where dropping is more preferred due to strong SLA requirements or where there is no requirement of fast traffic restoration for multi-point faults.

4.2. Multi-homed Scenario



In the multi-homed scenario, egress protection is also taken into consideration besides path protection. In addition to the mechanisms mentioned in the previous single-homed scenario, the following ones are also used for the protection of SR network:

o Ingress Node Switchover to Backup Egress Node

The ingress node monitors the liveness of egress nodes, such as enabling BFD for egress nodes, or validating IGP routes of egress nodes. When the failure occurs on the main egress node, the ingress node performs the switchover from the main egress node to the backup egress node. This mechanism works for both the traffics steered by SRv6 TE paths and SRv6 BE paths in the multi-home scenario. Note that, in the multi-homed scenario, the ingress node switchover works among the paths towards different egress nodes. Taking the SRv6 TE paths as an example, the ingress node switches among multiple SR Policies with different endpoints, while in the single-homed scenario the ingress node switches among multiple candidate paths within the same SR Policy.

In the cases with fast traffic restoring requirements, like less than 200ms, the local repair for egress node failure should be deployed.

The path protection is the same as the previous single-homed scenarios.

4.3. Liveness Check

As described in Section 4.1 and 4.2, BFD/S-BFD is used to monitor the liveness of links, neighbors, SR Policies and egress nodes.

The BFD time interval for links and neighbors is recommended to be 10ms * 3 and thus the local protection provided by TI-LFA would restore traffics in less than 50ms.

The BFD time interval for main candidate paths of SR Polices is recommended to be 50ms * 3, while the time interval for backup candidate paths can be relaxed to 100ms * 3. Thus, the end-to-end protection would restore traffics in less than 300ms.

The BFD time interval for egress nodes is recommended to be 50ms * 3.

5. Considerations for SRv6 Segment List Compression

[I-D.ietf-spring-srv6-srh-compression] enables a compressed encoding of the SRv6 Segment List in the SRH, which can reduce the SRv6 encapsulation size. The SRv6 Segment-List compression may have an effect on the protection of SRv6 networks, which is discussed in this section.

5.1. TI-LFA with C-SID

When SRv6 Segment List compression is enabled, the repair node may check the compression capabilities of nodes along the repair path and try to use C-SIDS to encode the repair path.

If NEXT-C-SID flavors are preferred, the TI-LFA repair list consist of the End SID with NEXT-C-SID flavor of the P node and the End.X SID(s) with NEXT-C-SID flavor of the path from P node to Q node, except for the last End.X SID which must not have NEXT-C-SID flavor. In addition, the End SID must be a global C-SID, and the End.X SID(s) can be local C-SID(s).

If REPLACE-C-SID flavors are preferred, the TI-LFA repair list consist of the End SID with REPLACE-C-SID flavor of the P node and the End.X SID(s) with REPLACE-C-SID flavor of the path from P node to Q node, except for the last End.X SID which must not have REPLACE-C-SID flavor.

5.2. Micro-Loop Avoidance with C-SID

If SRv6 Segment List compression is enabled, the converging node may check the compression capabilities of nodes along the post-convergence path and try to use C-SIDs to encode the path.

The TI-LFA mechanism can be used to compute the loop-free post-convergence path. If so, the building of TI-LFA repair list with C-SIDs is similar with the previous section.

6. Considerations for MSD Check

When calculating the label stack in SRv6 TI-LFA and micro-loop prevention scenarios, if the current node does not strictly verify the Maximum SID Depth (MSD) supported by nodes along the path, traffic may fail to forward according to the label stack. To address this issue, the following guidelines should be considered during deployment:

- Enable MSD Strict Check Configure MSD strict check to ensure the current node rigorously verifies the MSD supported by nodes along the path.

- Impact of MSD Strict Check After enabling MSD strict check, if the supported MSD of any node along the path is less than the required label stack depth, the label stack cannot be formed.

By enabling MSD strict check, network operators can ensure that the label stack is compatible with the MSD capabilities of all nodes along the path, preventing forwarding failures and improving network reliability.

7. Considerations for SRv6 Path MTU

SRv6 uses IPv6 as the forwarding plane, it is essential to consider MTU impacts to avoid packet discards and optimize bandwidth utilization. To address this, the following guidelines should be followed when planning and configuring SRv6 Path MTU:

- Reserve Additional SRH Header Length: When configuring SRv6 Path MTU, reserve additional space for scenarios such as TI-LFA FRR, micro-loop prevention, or Egress protection, which introduce extra Segment Routing Header (SRH) length. This reserved length is implemented by configuring a Path MTU reserve value at the SRv6 headend node.

- Active Path MTU Calculation: The Active Path MTU can be calculated by subtracting the reserved value from the configured SRv6 Path MTU.

- Recommended Reserve Value: A reserve value of 72 bytes is recommended to accommodate the additional SRH header length introduced in various scenarios.

8. Security Considerations

TBD.

9. IANA Considerations

This document has no IANA actions.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

[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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

[RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", RFC 9256, DOI 10.17487/RFC9256, July 2022, <https://www.rfc-editor.org/info/rfc9256>.

[I-D.ietf-spring-srv6-srh-compression] Cheng, W., Filsfils, C., Li, Z., Decraene, B., Cai, D., Clad, F., Zadok, S., Guichard, J., Aihua, L., Raszuk, R. and C. Li, " Compressed SRv6 Segment List Encoding in SRH", draft-ietf-spring-srv6-srh-compression-23 (work in progress), February 2025.

[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", draft-ietf-rtgwg-segment-routing-ti-lfa-21 (work in progress), February 2025.

10.2. Informative 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>.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <https://www.rfc-editor.org/info/rfc5714>.

[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>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[I-D.ietf-rtgwg-srv6-egress-protection] Hu, Z., Chen, H., Chen, H., Wu, P., Toy, M., Cao, C., He, T., Liu, L., and X. Liu, "SRv6 Path Egress Protection", Work in Progress, Internet-Draft, draft-ietf-rtgwg-srv6-egress-protection-17, November 2024.

[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-17 (work in progress), June 2024.

Contributors

Mengxiao Chen
H3C
Email: chen.mengxiao@h3c.com

Authors' Addresses

Yisong Liu
China Mobile
China

Email: liuyisong@chinamobile.com

Wenying Jiang
China Mobile
Beijing
China
Email: jiangwenying@chinamobile.com

Changwang Lin
New H3C Technologies
China

Email: linchangwang.04414@h3c.com

Xuesong Geng
Huawei Technologies
China

Email: gengxuesong@huawei.com

Yao Liu
ZTE Corp.
China

Email: liu.yao71@zte.com.cn

Appendix A. Examples

Figure 6 is used as a reference topology to illustrate the deployments of protection in SR networks. PE1 and PE3 are primary PE nodes for VPN service access. PE2 and PE4 are used as backup. The prefix of CE2, along with VPN service SID, is advertised by BGP routes from PE3 and PE4 to PE1 and PE2. The VPN traffic is from CE1 to CE2.

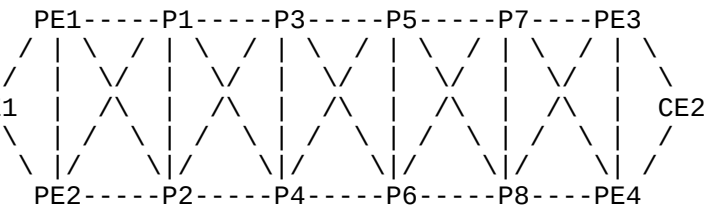


Figure 6: Reference Topology

The link metrics are configured as follows:

- Metrics of PE1-P2, PE2-P1, P1-P4, P2-P3, P3-P6, P4-P5, P5-P8, P6-P7, P7-PE4, P8-PE3, PE1-PE2 and PE3-PE4 links are 11.
- Metrics of all other links are 5.
- Link metrics are bidirectional.

A.1 Example of SR BE Scenario

BE scenario: SR BE paths are used to steer the VPN service. The deployments of protection are as follows:

- All nodes enable TI-LFA for local protection.
- All nodes enable BFD for links and neighbors.
- Ingress PE node enables FRR of SR BE path to backup egress PE node for service protection.
- Ingress PE node enables BFD for egress PE node to monitor the liveness of SR BE path.

Assume that the data plane is MPLS-SR. The MPLS labels are assigned using the following rules (just for the convenience of illustration).

```
NodeID:                 n for PEn, n+10 for Pn
Prefix-SID:             16000 + NodeID
Adj-SID:                24000 + NeigborNodeID
VPN label:              90000 + NodeID
```

For example, the labels assigned on PE1 and P8 are as follows.

```
PE1:
  Prefix-SID:           16001
  VPN label:            90001
  For PE1->P1:
    Adj-SID:            24011
  For PE1->P2:
    Adj-SID:            24012

P8:
  Prefix-SID:           16018
  For P8->P5:
    Adj-SID:            24015
  For P8->P6:
    Adj-SID:            24016
  For P8->P7:
    Adj-SID:            24017
  For P8->PE3:
    Adj-SID:            24003
  For P8->PE4:
    Adj-SID:            24004
```

PE1 installs the SR BE path to PE3 with the label stack of [16003, 90003] as the primary next-hop for the VPN flow. Meanwhile, PE1 also installs the SR BE path to PE4 with the label stack of [16004, 90004] as the backup next-hop.

PE1 enables BFD for Prefix-SID 16003 and 16004 to monitor the liveness of SR BE paths.

TI-LFA is enabled on all nodes. Take P1 for example. The shortest path from P1 to PE3 is via neighbor P3. In order to provide local protection for P3 node failure, P1 computes and installs the repair path P1->P2->P4->P6, using [16014, 24016] as the label stack.

All nodes use BFD to monitor the liveness of links and adjacent nodes.

Under normal circumstances, PE1 encapsulates the VPN payload in a label stack of [16003, 90003].

Assume that a failure occurs on P3. The fail-timer of BFD from P1 to P3 expires, so P1 perceives the failure. When P1 forwards the VPN packet, the TI-LFA repair path is used. Then, P1 pushes [16014, 24016] onto the label stack. The packet is forwarded in the repair path P1->P2->P4->P6 according to the top two labels. So the failure is repaired by local protection.

Assume that a failure occurs on PE3. TI-LFA does not work and the packets along the SR BE path are dropped. Then the BFD session from PE1 to Prefix-SID 16003 is down, so PE1 triggers the switchover to the SR BE path to PE4 and encapsulates the VPN payload in the label stack of [16004, 90004]. After that, the VPN traffic from CE1 to CE2 is recovered.

Assume that a failure occurs on link PE3-CE2. Since the BFD session from PE1 to Prefix-SID 16003 is still alive, PE1 continues to forward the VPN packets to PE3. When PE3 receives the packet, it pops all the labels, looks up the VPN table and forwards the packet to CE2. However, the link PE3-CE2 is failed. So PE3 selects the FRR alternate next-hop which is the SR BE path to PE4. Then PE3 encapsulates the packet in the label stack of [16004, 90004], and forwards it through the link PE3-PE4.

A.2 Example of SR TE Scenario

TE scenario: SR TE paths are used to steer the VPN service. The deployments of protection are as follows:

- In the SR Policy of SR TE strict path, disjoint backup candidate path is used as hot standby for end-to-end protection.
- Ingress PE node uses SR BE paths as backup for end-to-end protection of SR TE paths.
- Ingress PE node enables BFD for SR Policy. In the case of SR TE strict path, the reverse path of BFD packet keeps consistent with forward path.
- Ingress PE node enables BFD for locator of egress PE node to monitor the liveness of SR BE path.
- Ingress PE node enables FRR of paths to backup egress PE node for service protection.
- All nodes enable TI-LFA for local protection. All nodes enable BFD for links and neighbors.

In this scenario, the SR TE strict path is used to steer the VPN traffic flows to the primary egress node PE3, and the SR TE loose path is used for the backup egress node PE4.

Assume that the data plane is SRv6. The SRv6 SIDs are assigned using the following rules (just for the convenience of illustration), with G-SRv6 compression enabled.

NodeID:	An for PEn, Bn for Pn
Locator:	2001:DB8:NodeID::/48
End SID:	Locator:1::
End SID with COC:	Locator:2::
End DT:	Locator:100:: (Only for PE nodes)
End.X SID:	Locator:NeigborNodeID + F1::
End.X SID with COC:	Locator:NeigborNodeID + F2::

For example, the SRv6 SIDs assigned for PE1 and P8 are as follows.

```
PE1:
  Locator:              2001:DB8:A1::/48
  End SID:              2001:DB8:A1:1::
  End SID with COC:     2001:DB8:A1:2::
  End DT:               2001:DB8:A1:100::
  For PE1->P1:
    End.X SID:          2001:DB8:A1:B1F1::
    End.X SID with COC: 2001:DB8:A1:B1F2::
  For PE1->P2:
    End.X SID:         2001:DB8:A1:B2F1::
    End.X SID with COC: 2001:DB8:A1:B2F2::

P8:
  Locator:              2001:DB8:B8::/48
  End SID:              2001:DB8:B8:1::
  End SID with COC:     2001:DB8:B8:2::
  For P8->P5:
    End.X SID:          2001:DB8:B8:B5F1::
    End.X SID with COC: 2001:DB8:B8:B5F2::
  For P8->P6:
    End.X SID:         2001:DB8:B8:B6F1::
    End.X SID with COC: 2001:DB8:B8:B6F2::
  For P8->P7:
    End.X SID:         2001:DB8:B8:B7F1::
    End.X SID with COC: 2001:DB8:B8:B7F2::
  For P8->PE3:
    End.X SID:         2001:DB8:B8:A3F1::
    End.X SID with COC: 2001:DB8:B8:A3F2::
  For P8->PE4:
    End.X SID:         2001:DB8:B8:A4F1::
    End.X SID with COC: 2001:DB8:B8:A4F2::
```

The SR Policies on PE1 are configured as follows:

```
   SR Policy 1 (Strict Path to PE3)
     Candidate Path 1
       Preference: 20
       Segment List: 2001:DB8:A1:B1F2::, 2001:DB8:B1:B3F2::,
   2001:DB8:B3:B5F2::, 2001:DB8:B5:B7F2::, 2001:DB8:B7:A3F1::
     Candidate Path 2
       Preference: 10
       Segment List: 2001:DB8:A1:B2F2::, 2001:DB8:B2:B4F2::,
   2001:DB8:B4:B6F2::, 2001:DB8:B6:B8F2::,2001:DB8:B8:A3F1::

   SR Policy 2 (Loose Path to PE4)
     Candidate Path 1
       Preference: 20
       Segment List: 2001:DB8:B4:2::, 2001:DB8:B8:2::,2001:DB8:A4:1::
```

PE1 installs SR Policy 1, which is the SR TE strict path to PE3, as the primary next-hop for the VPN flow. SR Policy 1 has two disjoint candidate paths. The candidate path with higher preference is selected as the primary candidate path, and the candidate path with lower preference is selected as hot standby backup.

Meanwhile, the SR BE path to PE3, the SR TE loose path to PE4 (SR Policy 2), and the SR BE path to PE4 are also installed as backup next-hops. The priorities of multiple backup paths may be decided by either of the egress-node-first strategy or the TE-first strategy.

Egress-node-first strategy:

- primary: SR TE path to primary egress node PE3 (SR Policy 1)
- backup(1st priority): SR BE path to primary egress node PE3
- backup(2nd priority): SR TE path to backup egress node PE4 (SR Policy 2)
- backup(3rd priority): SR BE path to backup egress node PE4

TE-first strategy:

- primary: SR TE path to primary egress node PE3 (SR Policy 1)
- backup(1st priority): SR TE path to backup egress node PE4 (SR Policy 2)
- backup(2nd priority): SR BE path to primary egress node PE3
- backup(3rd priority): SR BE path to backup egress node PE4

Egress-node-first strategy is used as an example below.

PE1 enables BFD for SR Policy 1 and SR Policy 2 to monitor the liveness of SR TE paths. For SR Policy 1 which is the strict path, the forward and reverse paths of BFD packet should be the same. For example, the primary path of SR Policy 1 is PE1->P1->P3->P5->P7->PE3, so the reverse path should be PE3->P7->P5->P3->P1->PE1. A segment list of such reverse path is installed on PE3. PE1 may send BFD packet with the segment list of SR Policy 1 along with the BSID of reverse path. When the BFD packet is forwarded along the strict path to PE3, PE3 will add an outer IPv6 header with SRH carrying the segment list of [2001:DB8:A3:B7F2::, B7:B5F2, B5:B3F2, B3:B1F2, B1:A1F1], which instructs the packet to be forwarded along the same strict path back to PE1.

PE1 enables BFD for locator 2001:DB8:A3::/48 and 2001:DB8:A4::/48 to monitor the liveness of SR BE paths.

TI-LFA is enabled on all nodes. BFD are used to monitor the liveness of links and adjacent nodes.

Under normal circumstances, PE1 encapsulates the VPN payload in an outer IPv6 header with SRH carrying the segment list of primary candidate path of SR Policy 1 along with the VPN SID advertised by PE3. Using G-SRv6 compression, the segment list will be encoded as [2001:DB8:A1:B1F2::, B1:B3F2, B3:B5F2, B5:B7F2, B7:A3F1, 2001:DB8:A3:100::].

Assume that a failure occurs on P3. The packets are dropped since the failed P3 is on the path. The BFD session of the segment list in the primary candidate path of SR Policy 1 is down, so PE1 triggers the switchover to the backup candidate path of SR Policy 1. Then PE1 encapsulates the VPN payload in an outer IPv6 header with SRH carrying the segment list of [2001:DB8:A1:B2F2::, B2:B4F2, B4:B6F2, B6:B8F2, B8:A3F1, 2001:DB8:A3:100::].

Before the recovery of P3, assume that P8 also fails. The BFD session of the segment list in the backup candidate path of SR Policy 1 is also down. Then PE1 triggers the switchover to the 1st priority backup next-hop which is the SR BE path to PE3. PE1 encapsulates the VPN payload in an outer IPv6 header where the destination address is 2001:DB8:A3:100::.

Assume that a failure occurs on PE3. Both the BFD sessions of SR Policy 1 and locator 2001:DB8:A3::/48 are down, which means the primary next-hop and the 1st priority backup next-hop are down. So PE1 triggers the switchover to the 2nd priority backup next-hop,

which is the SR TE loose path to PE4. Then PE1 encapsulates the VPN payload in an outer IPv6 header with SRH carrying the segment list of [2001:DB8:B4:2::, B8:2, A4:1, 2001:DB8:A4:100::].

Before the recovery of PE3, assume that a failure occurs on P6. The fail-timer of BFD from P4 to P6 expires, so P4 perceives the failure. When P4 forwards the VPN packet, the TI-LFA repair path is used. Then, P4 encapsulates the packet in an outer IPv6 Header with SRH carrying a compressed segment-list of [2001:DB8:B5:2::, B5:B7F1]. The packet is forwarded in the repair path P4->P3->P5->P7 according to the outer IPv6 Header and SRH. So the failure is repaired by local protection.

Before the recovery of PE3, assume that a failure occurs on P8. When P6 forwards the VPN packet to destination address 2001:DB8:B8:2:: which is one of the segments in the segment list of SRH, the TI-LFA on P6 does not work, since the failed node P8 is the destination. So the packets are dropped. The BFD session of SR Policy 2 is down, and PE1 triggers the switchover to the 3rd priority backup next-hop which is the SR BE path to PE4. Then PE1 encapsulates the VPN payload in an outer IPv6 header where the destination address is 2001:DB8:A4:100::. If the routing convergence is not completed at the moment, P6 will use TI-LFA repair path P6->P5->P7->PE4 to forward the packet. After the routing convergence is done, P nodes will forward the packet along new shortest path excluding P8.

Assume that a failure occurs on link PE3-CE2. This is similar with the same failure in section 4.1. The BFD session is still alive, PE1 continues to forward the VPN packets to PE3. PE3 will select the FRR alternate next-hop for CE1 and forward the packet to PE4 with SR BE path.