Workgroup: SPRING
Internet-Draft:
draft-hegde-spring-auto-edge-protection-00
Published: 13 March 2023
Intended Status: Standards Track
Expires: 14 September 2023
Authors: S. Hegde
Juniper Networks Inc.
K. Szarkowicz
Juniper Networks Inc.
J. Zhang
Juniper Networks Inc.
D. Voyer
Bell Canada

Auto Edge Protection

Abstract

This document specifies procedures to automatically establish context based forwarding for providing fast reroute during egress node and egress link failures. It describes how to detect multi-homed services and establish context for forwarding. It also defines procedures to avoid conflicts among routers while establishing context.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 14 September 2023.

Table of Contents

1. Introduction

If the PE (egress node) or the PE-CE link (egress link) fails, the CE can switch the CE->PE traffic to the other PE within 10s of

miliseconds. Protection of PE->CE traffic, on the other hand, is handled on PE side. Certain critical customer services require high reliability and failover times in 10s of milliseconds for both CE->PE and PE->CE traffic directions. The Node and link failures, inside a domain are covered using Fast Reroute (FRR) mechanisms as described in [RFC4090], [RFC5286], [RFC7490], and [RFC7812]. These protection mechanisms are widely deployed. The sub-second convergence for PE->CE traffic direction can be achieved for egress node as well egress link failures as described in [RFC8679].

Multiple customers may be multi-homed to different set of PEs. For example, customer1 may be multi-homed to PE1 and PE2 while customer2 may be multi-homed to PE1 and PE3. When PE1 goes down, the protected traffic of customer1 need to be sent to PE2 while the customer2 traffic need to be sent to PE3. In order to achieve this requirement [RFC8679] suggests use of "context-IDs". These context-IDs are an indication of the Primary/protector PE group and need to be allocated based on multi-homed customer information. The adoption of the egress protection mechanisms [RFC8679] have been slow due to the operational overhead of configuring a context for each multi-homed service.

[I-D.hegde-rtgwg-egress-protection-sr-networks] describes a mechanism that statically allocates context as well as same service labels/SIDs on multi-homed PEs. This mechanism may be used without requiring protocol extensions but has the same limitation of operational complexity of managing context as well as service SIDs.

The recent growth of cloud technologies has given rise to virtualized service instances which are deployed and managed very dynamically. There may be more than one service instance deployed in the datacenter where one service instance acts as primary and the other as backup. In case of primary service instance connectivity failure, the traffic must to switch to backup service instance which has all the relevant application data backed-up. The primary/backup service instances may be brought up and moved around in a datacenter based on various system parameters such as CPU, memory, load etc. It is difficult to provide egress protection by manually configuring context in such environments.

The motivation of this document is to define procedures to automate the process of identifying multi-homed services and allocating a context id specific to the multi-homed PE group. This document focusses on segment routing based tunnelling such as SR-MPLS and SRv6 and defines necessary adaptation of the egress protection framework [RFC8679] to SR networks.

2. Edge protection Architecture

This section describes some of the basic principles used in edge protection architecture.

2.1. Procedures for detection of multi-homing

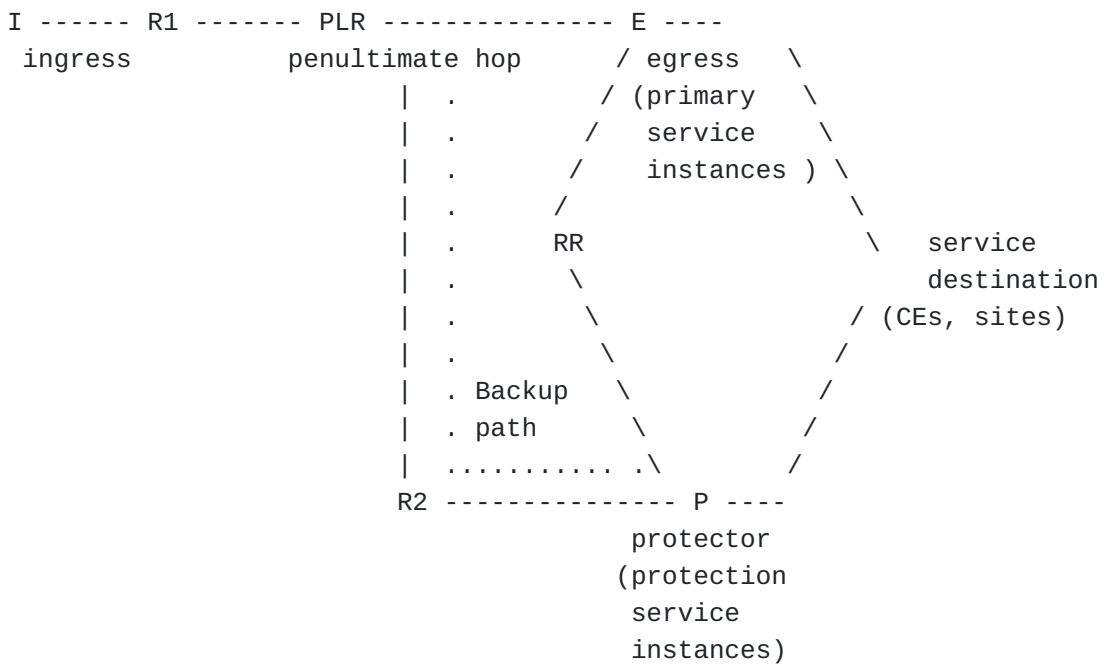

Figure 1: Detection of multi-homing

Figure 1 shows a typical multi-homed CE to two PEs E and P.

- E and P exchange service prefixes via Route Reflector RR.

- CE advertises a service prefix S1 to E, and E advertises S1 to P and I via RR. (Note: E and P might require implementation of draft-ietf-idr-best-external in order to advertise towards RR prefix S1 received from CE – via EBGP, despite receiving already prefix S1 from RR – via IBGP)

- E and P receives prefix S1 from each other as well as CE

- E and P detect multi-homing of S1. E and P are part of the multi-home group.

- There can be more than two PEs in the multi-home group;

- Multi-homed PE group is identified by the participating PEs and the primary, secondary or weights associated with each PE.

2.2. Context-id

- Context-ID is an anycast IPv4 or IPv6 address.

- Context-ID is a unique anycast IPv4 or IPv6 address that is assigned to the multi-homed PE group.

- Context-ID is used to advertise the service multi-homing context into the network

- Context-ID MUST be routable inside the network domain.

- Context-IDs may be summarized at the ABRs in a SRv6 based network

- Context-IDs may be advertised in BGP as part of newly defined attributes

- Context-IDs are advertised in IGP as prefixes

2.2.1. Context-id pools

- Context-ID pool is a set of IPv4 or IPv6 addresses assigned to a router

- Context-ID pool assigned on every router needs to be disjoint w.r.t to other router's context-ID pool.

- Context-IDs are dynamically allocated from the context-ID pools

- Context-ID pools are advertised in IGP, and if there is a conflict, it is logged

- In order to participate in auto-edge protection, a node MUST have a valid, non-conflicting context-id pool allocated to it.

2.2.1.1. Context-id management

- On detection of multi-homing Section 2.1, router allocates a context-ID from its context-ID pool, unless the service advertisement already contains an allocated context-ID

- If the service advertisement already contains an allocated context-ID, the same context-ID is used by the PE that receives the advertisement.

- In order to minimize the context-ID fluctuations, if a PE stops advertising the service prefix, other PEs MAY continue to use the same context-ID for a configurable amount of time.

2.3. Context-id database

- context-id database stores the allocated context-ids and the associated key information on which PE group the context-ID is allocated.Context-ID database on a PE contains all context-ids allocated for the multi-home PE group where this particular PE participates.

- If there is another service prefix that is also advertised by same PE group and same primary/secondary combination, then the same context-id is used for this service prefix as well.

- The context-id database contents may vary based on type of forwarding plane being used.

- In case of SRv6, the context-id need to be different for each flex-algo but in case of SR-MPLS same context-id may be used for each flex-algo but different SIDs for each flex-algo need to be allocated.

2.4. Scale Considerations

Each context-ID is advertised as a prefix in IGP. The number of context-IDs allocated in an IGP domain depends on the number of multi-homed PE groups in the domain. If the number combinations of the multi-homed PE group based on service attachments is large, it may result into significant increase in the number of prefixes in the IGP domain. In order to contain the number of context-IDs, operators may want to provide the auto edge protection feature to certain niche services or to some premium customers. The services that require such protection may be indicated by local configuration and/or with specific BGP community advertisement.

In cases where scale is not a concern, the auto-edge protection may be applied in general to all services.

2.5. Solution Overview

In order to support the auto edge protection feature, the network need to be provisioned with the dis-joint context-id pool on each PE that is going to participate in the auto-edge protection feature. The PE nodes that participate in the auto-edge protection feature MUST be upgraded to support the feature.

When a new multi-homed service is provisioned, the primary/protector role for the PE needs to be identified and configured along with the

service. In case the service requires the traffic to be equally load-balanced among multi-homed PEs, that needs to be indicated through configuration. The service configuration should also have indication that it requires auto edge protection. When service requiring auto edge protection is to be advertised, existance of same service prefix from another PE is checked.

```
 - If there is already an advertised service
    - that already contains a context-ID then
      - the same context-ID is used.
      - Service prefix BGP next-hop attribute is set to context-ID
      - Context-ID is updated in context-database
      - IGP is informed about context-ID and SID for distribution
        of the context-ID and SID.
      - The service SID/label that the other PE allocated, same servi
        is also allocated. In case of MPLS, the service Label is prog
        context table corresponding to the multi-homed PE.
                    -  The Context-label is made to point to the con

 - Otherwise, a new context-ID and context-SID is allocated.
      - Service prefix BGP next-hop attribute is set to context-ID
      - Context-ID is updated in context-database
      - IGP is informed about context-ID and SID for distribution
        of the context-ID and SID.
      - A new  service SID/label  is allocated.
        In case of MPLS, the service Label is programmed in a
        global MPLS table and points to the customer VPN.
                    - Context-label points to the global MPLS table.
 - If there is no other PE advertising the same service prefix,
    - then the service follows the usual
        service advertisement mechanism without context-ID.
```

Figure 2: Context-ID allocation

The procedures described above ensure that the service labels, context-ID and context-SID advertisements and forwarding plane programming is all set up correctly such that any PE receives the traffic, it will be sent to the correct CE.

The ingress encapsulates the traffic in a tunnel. The tunnel will be based on the Context-ID. The data forwarding inside the domain is based on the context-ID. If the Primary PE goes down, the PLR forwards the traffic to the backup PE. The service label/SID in backup PE points to the correct CE and traffic is delivered to the CE.

3. Edge protection in MPLS networks

3.1. Anycast SID

- SR-MPLS networks use an MPLS dataplane and require SIDs to be associated with the IP addresses for MPLS forwarding

- Context-ID is an anycast IP4/IPv6 address in SR-MPLS networks. Anycast IP address pool need to be allocated for every participating PE.

- SR-MPLS networks also require a SID to be associated with the anycast IP address.

- Anycast SID is dynamicaly allocated when the context-ID is allocated

- Anycast SIDs are global SIDs and MUST avoid conflict with other SIDs allocated by other nodes in the same domain

- Anycast SIDs assigned to a Anycast IP4/IPv6 address are derived from the dedicated anycast-sid pool on each router. This will avoid two PE routers allocating the same SID for different Anycast IP address and avoids conflicts.

- Context-ID and the anycast-SID associated with the context-ID are advertised as typical anycast IP and associated anycast-SID in IGP. No new extensions are required for advertising the context-id in IGP.

3.2. Anycast SID pools

- A dedicated global index pool allocated to each node eligible to participate in auto-edge protection

- The dedicated index pool is derived from the Global SRGB on each router.

- The anycast-sid pool is advertised in the IGP.

- Each node validates that the anycast-SID pool is disjoint from other such advertisements.

- If there is an overlap, the nodes whose anycast-sid pool overlaps, stop participating in auto-edge protection, and advertise the services with usual loopbacks as BGP next-hop attribute.

3.3. Theory of operation

3.3.1. Egress Node protection

* Segment routing natively supports ECMP and hence there can be one primary and one or more protector. There can also be all multi-homed PEs in the group are ECMP and protect each other.

* Context-ID is allocated for the multi-home group. The anycast IP address is originated in the IGP on these PEs and advertised in IGP. The metric advertised for the anycast prefix is based on the primary/secondary role. The primary PE advertises the anycast prefix with a metric of 1 and secondary advertises with maximum usable metric (MAX-METRIC-1). For ECMP group, all PEs advertise a metric of 1.

* When multi-homing is detected, the node that detects the multi-homing allocates a context-ID and an anycast SID from its context-ID pool and Anycast SID pool. These values do not conflict with someone else's context-ID and anycast SID allocation as each node has been assigned a unique pool.

* The service prefix is advertised with context-ID and anycast sid information in newly defined edge protection attributes.

* The service prefixes are advertised with locally allocated service label.

* IGP advertises the Context-ID and the associated anycast-SID in the Prefix-SID advertisement.

* Anycast SID for the Context-ID anycast address is advertised as a ultimate-hop-popping (UHP) label in IGP.

* The Anycast UHP label is programmed with POP and Lookup action on the node that successfully allocated the context-ID.

* The Anycast UHP label is programmed with the POP and lookup into corresponding context-table action on one or more nodes in the multi-homed PE group.

* When multi-homing is detected on a node, if the service advertisement already contains context-ID and anycast-sid information, the same context-id and anycast-sid is used in the receiving router.

* In cases where two PE nodes independently allocate different context-ID for the same context, the node that allocated numerically higher context-ID will win and the other nodes MUST release the context-ids allocated by them.

* The nodes that use other PE allocated context-ID will also use that PEs service label and build context-table. The context-table will have the service label allocated by the other PE and will point to correct VPN table. The PE then sets nexthop attribute to the anycast address and uses the same service label that the first PE advertised.

* The anycast SID is programmed to point to the context-table on the PE that adopted another PE's context-ID.

*IGP routing tables get built on all nodes for the context-ID IP address and the anycast-SID. TI-LFA paths also get programmed on all nodes. This is based on current protocol definitions and procedures no changes required.

* The ingress will receive the service advertisement along with context-id information and use the asociated context-id tunnel to forward the traffic. If there is failure of the primary node, PLR will switch the traffic towards the protector based on the context-id route in the IGP. As the backup paths are already programmed in forwarding plane this switch is expected to be similar to local FRR and in the order of tens of millisecs.

* In order to support the mechanism described in this section, Primary and protector PE nodes need to be upgraded. Binding-SID advertisement (For mirroring context [RFC8402]) is not required and PLR does not require software upgrade.

3.3.2. Egress Node protection with mirror SID (Binding-SId with mirroring context)

* An alternate optional mechanism uses mirror SID advertisement. Mirror-SID functionality is similar to the context-label as described in [RFC8679].

* The nodes with protector role build "context based forwarding" for the nodes they are protecting. This procedure is the same as described in [RFC8679] section 5.7.

* Service prefix is advertised with a locally allocated label from both primary and protector PE.

* As this mechanism uses different service label, only primary/ protector mode can be supported. ECMP mode cannot be supported.

* Protectors allocate a "context-label" for the context table and advertise as mirror-sid for the context-id. The mirror SID advertisement is optional.

* IGP routing tables get built on all nodes for the context-ID IP address and the anycast-SID. TI-LFA paths also get programmed on all nodes.

* On the PLR nodes, the backup path for the context-ID and anycast-SID will have a bottom label context label and top labels will be TI-LFA computed labels for backup path to the protector. The protector chosen would be based on metric and will follow the post convergence path. In cases where there are multiple protectors and are equidistant, both backup paths will be installed with corresponding context label.

* On primary node failure, the PLR will switch the traffic to protector which will have a context-label just above the service label. Traffic arrives at the protector with a context-label on top. Looks up the corresponding context-table with service label which points to the correct VRF table and traffic gets forwarded to the CE.

* This mechanism required primary, protector PEs and PLR to be upgraded. This mechanism cannot support ECMP mode of PEs.

3.3.3. Egress link Protection

The egress link protection mechanisms follow the procedures described in [RFC8679] sec 6 and this document does not propose any change.

3.4. Egress Protection in MPLS networks Example

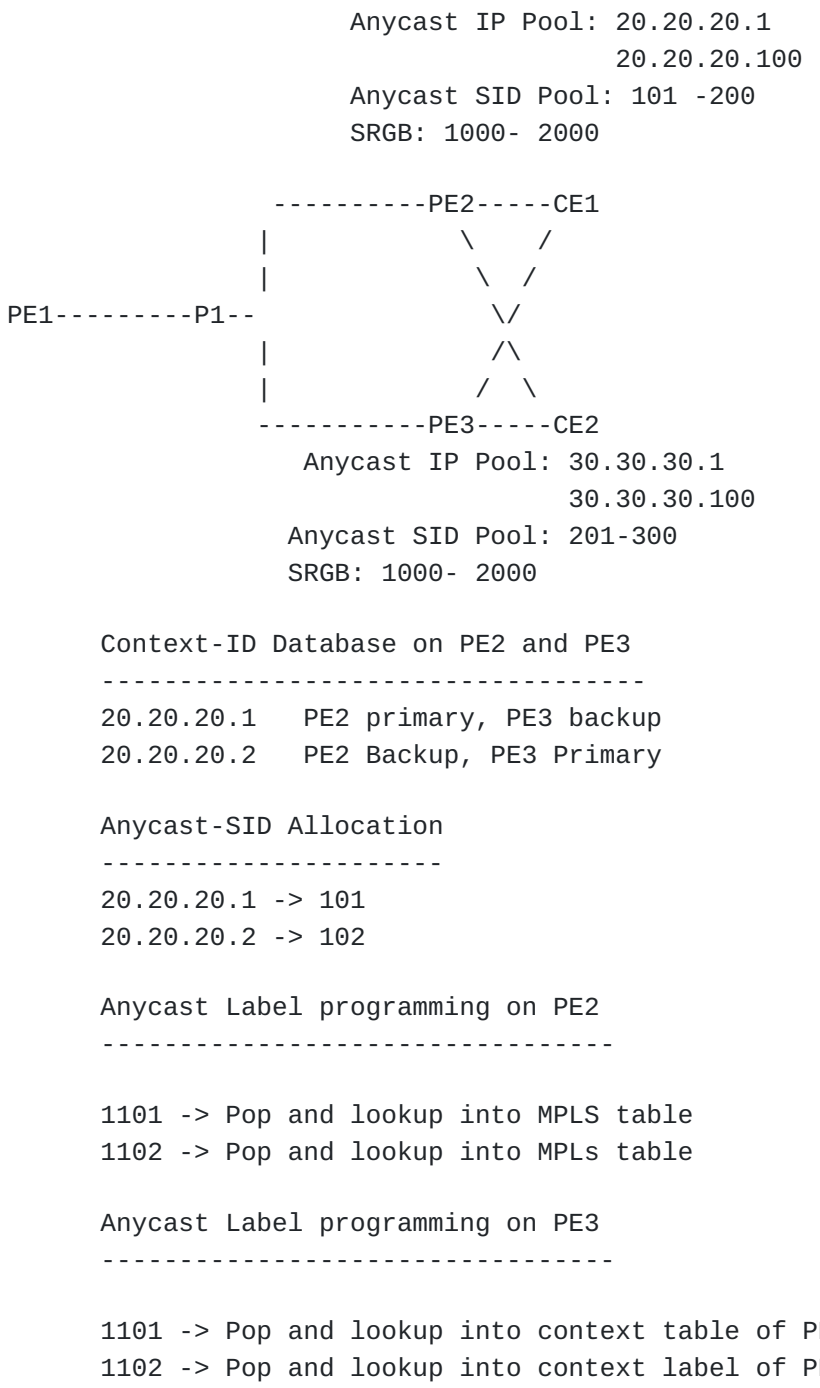

Figure 3: SR-MPLS network

The above diagram Figure 3 shows and example MPLS network with PE1,PE2,PE3 edge routers and P transit router. PE2 and PE3 routers participate in the auto-edge protection feature. On PE2, anycast IP address pool of 20.20.20.1 to 20.20.20.100 is configured and an anycast SID pool of 101 to 200 is configured. Anycast SID pool is a small pool of SIDs from the SRGB. PE3 has an anycast IP address of

30.30.30.1 to 30.30.30.100 and anycast SID pool of 201 to 300. The anycast IP address pool and the anycast SID pool both are disjoint between PE2 and PE3 and, also do not conflict with other IP address and SIDs in the network. CE1 is multi-homed to PE2 and PE3 with PE2 as primary and PE3 as backup. CE2 is also multi-homed to PE2 and PE3 with PE3 as primary and PE2 as backup.

In this example, when CE1 connects to PE2 and PE3, PE2 detects the multi-homing first and allocates the context-ID of 20.20.20.1 and a SID of 101 from its pool. PE3 receives the advertisement from PE2 and associates the same context-ID and SID for the multi-homed PE group (PE2 primary, PE3 backup). PE3 creates a context table for PE2 programs the service label from PE2's advertisement into a context table. This service label in the context table points to the VPN table to which CE1 belongs.

PE3 also advertises the same service label that PE2 allocated for CE1 prefix and sets the next-hop attribute to the context-ID 20.20.20.1. As the context-ID is used as the next-hop, the lookup for the service label on PE3 happens on the context-table of PE2. Both PE2 and PE3 advertise the anycast IP address 20.20.20.1 and the SID 101 with no-PHP bit set in IGP. The diagram Figure 3 shows the IGP anycast label programming on PE2 and PE3. On PE2, the anycast label 1101 has an action to POP the label and lookup in the global MPLS ILM table. On PE3, anycast label 1101 has an action to POP the label and lookup in PE2's context table.

Similarly, when CE2 connects to PE2 and PE3, they detect multi-homing. We have taken a case when PE2 first detects multi-homing. The scenario will work in a similar way if PE3 first detects multi-homing. As the multi-homed PE group is (PE2 backup, PE3 Primary), a new context-ID and anycast SID is allocated on PE2. As shown in diagram Figure 3, 20.20.20.2 is the context-ID and 102 is the anycast SID. The anycast Label 1102 programming on PE2 and PE3 is shown in Figure 3. Note that on the node that allocates context-ID, SID and service labels from its own pool have the anycast label pointing to the global MPLS ILM table while the node that inherits the context-ID, SID and the service label allocated by other nodes pool, has the anycast label pointing to the context-table of the node from which it inherits the context-ID.

4. Edge protection in SRV6 networks

4.1. Anycast locator/uN pools

* Anycast locator pool is configured on every node. In case of micro-sid, uN pool is configured. uN pool is taken from the global SID space.

* Anycast locator/uN pools should be disjoint and should not overlap across nodes.

* Every node advertises its configured locator pool/uN pool in IGP.

* If there is overlap in the advertised pools, the nodes that advertise overlapping pools stop participating in auto-edge protection.

* The anycast locators/uN will be allocated from this pool.

* In SRv6, separate SID pool is not required as all the allocated SIDs will be from the locator/uN and will not conflict due to the disjointness of the locator pool.

4.2. Egress Node Protection

* Disjoint locator pools/uN are configured on each node and advertised in IGP. Eligibility to participate in auto-edge protection is evaluated.

* Service prefix is advertised with single homed locator and SID.

* When the same service prefix is received on another PE via CE as well multi-homing is detected.

* The node that detects multi-homing (it could be primary or could be protector) allocates a new locator/uN from the pool.

* The node allocates an END SID and service SID END.DT4/END.DT6/ uDT4/uDT6 from the new locator.

* Service SID and the context-ID information is advertised in the service prefix.

* Locator/uN and the END SID are advertised in IGP.

* All other PEs part of the multi-home group receive the service advertisement and, also detect multi-homing. They allocate the same anycast locator/anycast uN and the same service SID and advertise the service with the new SID and the context information.

* All the PEs in multi-home group also advertise the anycast locator/anycast uN and the END SID.

* As all the PEs are advertising same service SID, there is no need to build context table and no need to allocate mirror SID

* IGPs build the primary path and the backup path for the anycast locator/uN. These are standard procedures, and no change is

required. Only the PE routers need to be upgraded to support this feature. P routers can continue to run older versions.

* The ingress receives the service prefix with context-id information and uses the Tunnel corresponding to the anycast locator associated with the Service SID.

* The tunnel could be best effort IGP tunnel or SRv6-TE policy [RFC9256] , the procedures for protection are identical for both.

* If the egress node goes down, the PLR has the backup path programmed to the anycast locator advertised by the protector PE. The traffic will switch to the alternate PE and the failover will be local FRR protection where the failover time in the order of 10s millisecs can be guaranteed.

4.3. Egress link Protection

* When the service prefix is received from another PE via iBGP peer as well as from CE via eBGP peer, the backup paths towards the PE advertisement is created through service multipath procedures. The path from PE to the other PE will be primary path and will not be a TI-LFA computed path.

* When the primary CE-PE link fails, traffic will be FRR switched towards the other PE.

* As the same service SIDs are allocated by both PEs, the context table procedures are not required.

4.4. Egress Protection in SRv6 networks Example

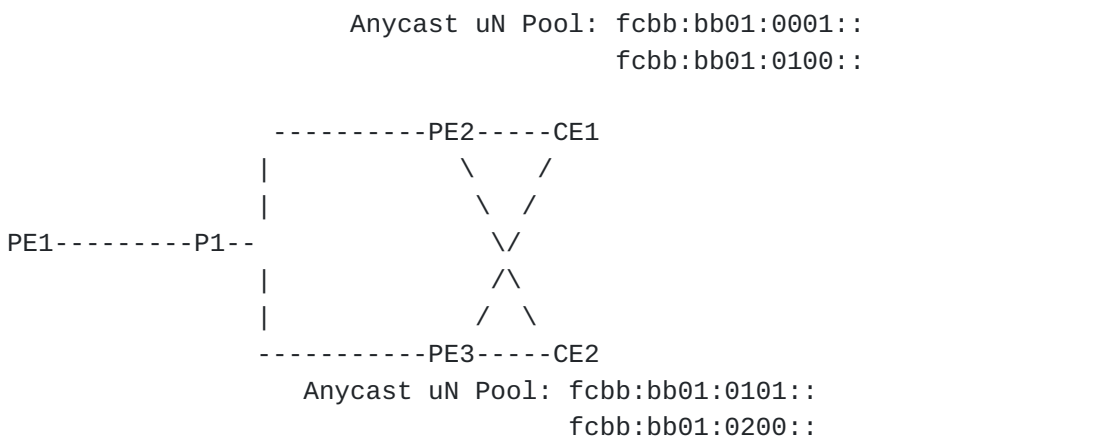

Figure 4: SRv6 network

The diagram Figure 4 shows a SRv6 network provisioned with micro-sids. The uN pools are configured on PE2 and PE3 which participate in the auto-edge protection. The UN pools are disjoint between PE2 and PE3. CE1 and and CE2 are multi-homed to PE2 and PE3. CE1 uses (PE2 primary, PE3 backup). CE2 uses (PE2 backup, PE3 primary).

The example in Figure 4 shows the detection of multi-homing for both CE1 and CE2 first happening on PE2. PE2 allocates an anycast uN fcbb:bb01:001 for CE1 (PE2 primary, PE3 backup) and an anycast uN fcbb:bb01:001 for CE2 (PE2 backup, PE3 primary). When PE3 receives CE1 advertisement from PE2, it also detects multi-homing and associates the already allocated uN pool and the service SID from PE2. IGP advertises the anycast uN as locator in IGP. As the anycast uN pools are unique to each multi-homed group pair, and the service SIDs are allocated from this unique pool, there is no possibility of

SID conflict among the PEs in the multi-homed PE group. If two or more PEs simultaneously allocate a uN, the node that allocated numerically highest uN wins. There is no need to build the context table and advertise mirror SID in case of SRv6 networks. The service SID programming looks identical on both PEs as shown in Figure 4.

5. Backward Compatibility

6. Operational Considerations

7. Security Considerations

TBD

8. IANA Considerations

9. Acknowledgements

10. Contributors

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[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>.

11.2. Informative References

[I-D.hegde-rtgwg-egress-protection-sr-networks] Hegde, S., Lin, W., and S. Peng, "Egress Protection for Segment Routing (SR) networks", Work in Progress, Internet-Draft, draft-hegde-rtgwg-egress-protection-sr-networks-02, 2 March 2022, <https://datatracker.ietf.org/doc/html/draft-hegde-rtgwg-egress-protection-sr-networks-02>.

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090,

DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

[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>.

[RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, DOI 10.17487/RFC7490, April 2015, <https://www.rfc-editor.org/info/rfc7490>.

[RFC7812] Atlas, A., Bowers, C., and G. Enyedi, "An Architecture for IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR)", RFC 7812, DOI 10.17487/RFC7812, June 2016, <https://www.rfc-editor.org/info/rfc7812>.

[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>.

Authors' Addresses

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

Email: shraddha@juniper.net

Krzysztof Szarkowicz
Juniper Networks Inc.

Email: kszarkowicz@juniper.net

Jeffrey Zhang
Juniper Networks Inc.

Email: zzhang@juniper.net

Daniel Voyer
Bell Canada

Email: daniel.voyer@bell.ca