Internet Engineering Task Force (IETF)
Request for Comments: 7916
Category: Standards Track
ISSN: 2070-1721

S. Litkowski, Ed.
B. Decraene
Orange
C. Filsfils
K. Raza
Cisco Systems
M. Horneffer
Deutsche Telekom
P. Sarkar
Individual Contributor

July 2016

Operational Management of Loop-Free Alternates

Abstract

Loop-Free Alternates (LFAs), as defined in RFC 5286, constitute an IP Fast Reroute (IP FRR) mechanism enabling traffic protection for IP traffic (and, by extension, MPLS LDP traffic). Following early deployment experiences, this document provides operational feedback on LFAs, highlights some limitations, and proposes a set of refinements to address those limitations. It also proposes required management specifications.

This proposal is also applicable to remote-LFA solutions.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7916.

Table of Contents

1. Introduction

Following the first deployments of Loop-Free Alternates (LFAs), this document provides feedback to the community about the management of LFAs.

- Section 3 provides real use cases illustrating some limitations and suboptimal behavior.
- Section 4 provides requirements for LFA simulations.
- Section 5 proposes requirements for activation granularity and policy-based selection of the alternate.
- Section 6 expresses requirements for the operational management of LFAs and, in particular, a policy framework to manage alternates.
- Section 7 details some operational considerations of LFAs, such as IS-IS overload bit management and troubleshooting information.

1.1. Requirements Language

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

2. Definitions

- Per-prefix LFA computation: Evaluation for the best alternate is done for each destination prefix, as opposed to the "per-next-hop" simplification technique proposed in Section 3.8 of [RFC5286].
- PE router: Provider Edge router. These routers connect customers to each other.
- P router: Provider router. These routers are core routers without customer connections. They provide transit between PE routers, and they form the core network.
- Core network: subset of the network composed of P routers and links between them.
- Core link: network link part of the core network, i.e., a link between P routers.
- Link-protecting LFA: alternate providing protection against link failure.

- Node-protecting LFA: alternate providing protection against node failure.

- Connected alternate: alternate adjacent (at the IGP level) to the Point of Local Repair (PLR) (i.e., an IGP neighbor).

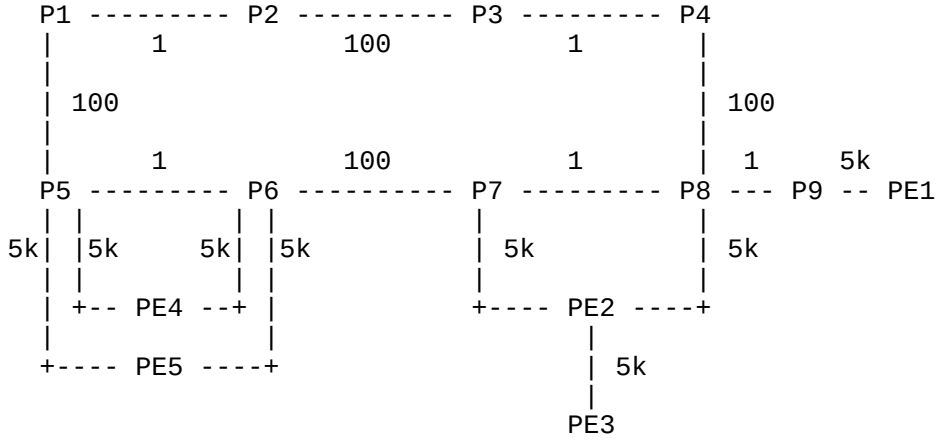
- Remote alternate: alternate that does not share an IGP adjacency with the PLR.

3. Operational Issues with Default LFA Tiebreakers

[RFC5286] introduces the notion of tiebreakers when selecting the LFA among multiple candidate alternate next hops. When multiple LFAs exist, [RFC5286] has favored the selection of the LFA that provides the best coverage against the failure cases. While this is indeed a goal, it is one among multiple goals, and in some deployments this leads to the selection of a suboptimal LFA. The following sections detail real use cases related to such limitations.

Note that the use case for LFA computation per destination (per-prefix LFA) is assumed throughout this analysis. We also assume in the network figures that all IP prefixes are advertised with zero cost.

3.1. Case 1: PE Router Protecting against Failures within Core Network



Px routers are P routers using n * 10 Gbps links.
PEs are connected using links with lower bandwidth.

Figure 1

In Figure 1, let us consider the traffic flowing from PE1 to PE4. The nominal path is P9-P8-P7-P6-PE4. Let us now consider the failure of link P7-P8. As the P4 primary path to PE4 is P8-P7-P6-PE4, P4 is not an LFA for P8 (because P4 will loop traffic back to P8), and the only available LFA is PE2.

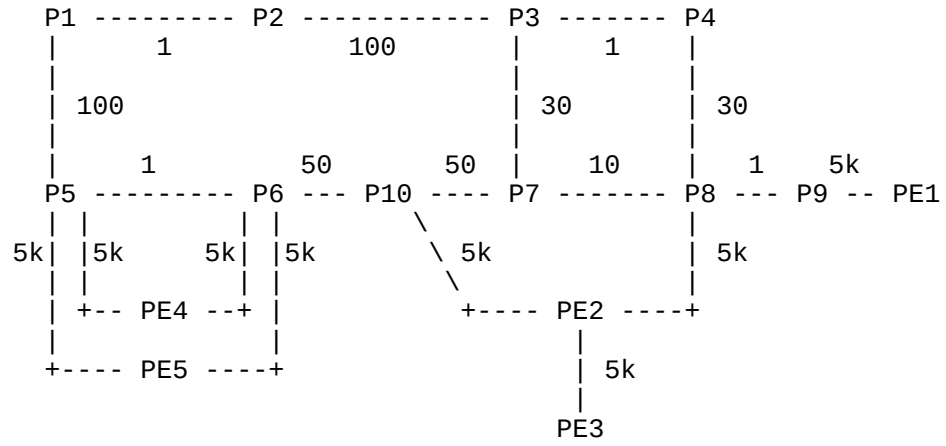
When the core link P8-P7 fails, P8 switches all traffic destined to PE4/PE5 towards the node PE2. Hence, a PE node and PE links are used to protect against the failure of a core link. Typically, PE links have less capacity than core links, and congestion may occur on PE2 links. Note that although PE2 is not directly affected by the failure, its links become congested, and its traffic will suffer from the congestion.

In summary, in the case of P8-P7 link failure, the impact on customer traffic is:

- From PE2’s point of view:
  - without LFA: no impact.
  - with LFA: traffic is partially dropped (but possibly prioritized by a QoS mechanism). It must be highlighted that in such a situation, traffic not affected by the failure may be affected by the congestion.
- From P8’s point of view:
  - without LFA: traffic is totally dropped until convergence occurs.
  - with LFA: traffic is partially dropped (but possibly prioritized by a QoS mechanism).

Besides the congestion aspects of using a PE router as an alternate to protect against a core failure, a service provider may consider this to be a bad routing design and would want to prevent it.

3.2. Case 2: PE Router Chosen to Protect against Core Failures while P Router LFA Exists

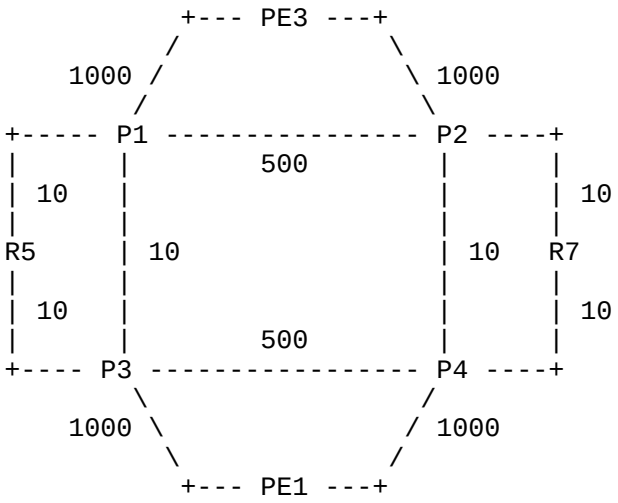


Px routers are P routers meshed with n * 10 Gbps links.
PEs are meshed using links with lower bandwidth.

Figure 2

In Figure 2, let us consider the traffic coming from PE1 to PE4. The nominal path is P9-P8-P7-P10-P6-PE4. Let us now consider the failure of the link P7-P8. For P8, P4 is a link-protecting LFA and PE2 is a node-protecting LFA. PE2 is chosen as the best LFA, due to the better type of protection that it provides. Just as in case 1, this may lead to congestion on PE2 links upon LFA activation.

3.3. Case 3: Suboptimal P Router Alternate Choice



Px routers are P routers.
P1-P2 and P3-P4 links are 1 Gbps links.
All other inter-Px links are 10 Gbps links.

Figure 3

In Figure 3, let us consider the failure of link P1-P3. For destination PE3, P3 has two possible alternates:

- P4, which is node-protecting
- R5, which is link-protecting

P4 is chosen as the best LFA, due to the better type of protection that it provides. However, for bandwidth capacity reasons, it may not be desirable to use P4. A service provider may prefer to use high-bandwidth links as the preferred LFA. In this example, preferring the shortest path over the type of protection may achieve the expected behavior, but in cases where metrics do not reflect the bandwidth, this technique would not work and some other criteria would need to be involved when selecting the best LFA.

3.4. Case 4: No-Transit LFA Computing Node

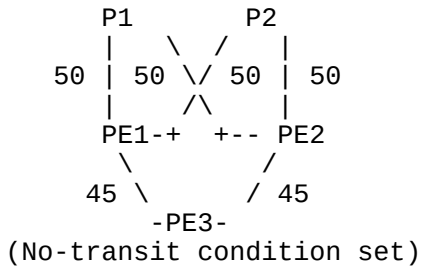


Figure 4

The IS-IS and OSPF protocols define some way to prevent a router from being used for transit.

The IS-IS overload bit is defined in [ISO10589], and the OSPF R-bit is defined in [RFC5340]. Also, the OSPF stub router is defined in [RFC6987] as a method to prevent transit on a node by advertising MaxLinkMetric on all non-stub links.

In Figure 4, PE3 has its no-transit condition set (permanently, for design reasons) and wants to protect traffic using an LFA for destination PE2.

On PE3, the loop-free condition is not satisfied: 100 !< 45 + 45. PE1 is thus not considered as an LFA. However, thanks to the no-transit condition on PE3, we know that PE1 will not loop the traffic back to PE3. So, PE1 is an LFA to reach PE2.

In the case of a no-transit condition set on a node, LFA behavior must be clarified.

4. Need for Coverage Monitoring

As per [RFC6571], LFA coverage depends strongly on the network topology that is in use. Even if the remote-LFA mechanism [RFC7490] significantly extends the coverage of the basic LFA specification, there are still some cases where protection would not be available. As network topologies are constantly evolving (network extension, additional capacity, latency optimization, etc.), the protection coverage may change. Fast Reroute (FRR) functionality may be critical for some services supported by the network; a service provider must always know what type of protection coverage is currently available on the network. Moreover, predicting protection coverage in the event of network topology changes is mandatory.

Today, network simulation tools associated with "what if" scenarios are often used by service providers for the overall network design (capacity, path optimization, etc.). Sections 7.3, 7.4, and 7.5 of this document propose the addition of LFA information into such tools and within routers, so that a service provider may be able to:

- evaluate protection coverage after a topology change.
- adjust the topology change to cover the primary need (e.g., latency optimization, bandwidth increase) as well as LFA protection.
- constantly monitor the LFA coverage in the live network and receive alerts.

Documentation of LFA selection algorithms by implementers (default and tuning options) is important in order to make it possible for third-party modules to model these policy-based LFA selection algorithms.

5. Need for LFA Activation Granularity

As in all FRR mechanisms, an LFA installs backup paths in the Forwarding Information Base (FIB). Depending on the hardware used by a service provider, FIB resources may be critical. Activating LFAs by default on all available components (IGP topologies, interfaces, address families, etc.) may lead to a waste of FIB resources, as generally only a few destinations in a network should be protected (e.g., loopback addresses supporting MPLS services) compared to the number of destinations in the Routing Information Base (RIB).

Moreover, a service provider may implement multiple different FRR mechanisms in its networks for different applications (e.g., Maximally Redundant Trees (MRTs), TE FRR). In this scenario, an implementation MAY allow the computation of alternates for a specific destination even if the destination is already protected by another mechanism. This will provide redundancy and permit the operator to select the best option for FRR, using a policy language.

Section 6 provides some implementation guidelines.

6. Configuration Requirements

Controlling the selection of the best alternate and the granularity of LFA activation is a requirement for service providers. This section defines configuration requirements for LFAs.

6.1. LFA Enabling/Disabling Scope

The granularity of LFA activation SHOULD be controlled (as alternate next hops consume memory in the forwarding plane).

An implementation of an LFA SHOULD allow its activation, with the following granularities:

- Per routing context: Virtual Routing and Forwarding (VRF), virtual/logical router, global routing table, etc.
- Per interface.
- Per protocol instance, topology, area.
- Per prefix: Prefix protection SHOULD have a higher priority compared to interface protection. This means that if a specific prefix must be protected due to a configuration request, an LFA MUST be computed and installed for that prefix even if the primary outgoing interface is not configured for protection.

An implementation of an LFA MAY allow its activation, with the following criteria:

- Per address family: IPv4 unicast, IPv6 unicast.
- Per MPLS control plane: For MPLS control planes that inherit routing decisions from the IGP routing protocol, the MPLS data plane may be protected by an LFA. The implementation may allow an operator to control this inheritance of protection from the IP prefix to the MPLS label bound to this prefix. The inheritance of protection will concern IP-to-MPLS, MPLS-to-MPLS, and MPLS-to-IP entries. As an example, LDP and Segment Routing extensions [SEG-RTG-ARCH] for IS-IS and OSPF are control-plane eligible for this inheritance of protection.

6.2. Policy-Based LFA Selection

When multiple alternates exist, the LFA selection algorithm is based on tiebreakers. Current tiebreakers do not provide sufficient control regarding how the best alternate is chosen. This document proposes an enhanced tiebreaker allowing service providers to manage all specific cases:

1. An LFA implementation SHOULD support policy-based decisions for determining the best LFA.

2. Policy-based decisions SHOULD be based on multiple criteria, with each criterion having a level of preference.

3. If the defined policy does not allow the determination of a unique best LFA, an implementation SHOULD pick only one based on its own decision. For load-balancing purposes, an implementation SHOULD also support the election of multiple LFAs.

4. The policy SHOULD be applicable to a protected interface or a specific set of destinations. In the case of applicability to the protected interface, all destinations primarily routed on that interface SHOULD use the policy for that interface.

5. The choice of whether or not to dynamically re-evaluate policy (in the event of a policy change) is left to the implementation. If a dynamic approach is chosen, the implementation SHOULD recompute the best LFAs and reinstall them in the FIB without service disruption. If a non-dynamic approach is chosen, the policy would be taken into account upon the next IGP event. In this case, the implementation SHOULD support a command to manually force the recomputation/reinstallation of LFAs.

6.2.1. Connected versus Remote Alternates

In addition to connected LFAs, tunnels (e.g., IP, LDP, RSVP-TE, Segment Routing) to distant routers may be used to complement LFA coverage (tunnel tail used as virtual neighbor). When a router has multiple alternate candidates for a specific destination, it may have connected alternates and remote alternates (reachable via a tunnel). Connected alternates may not always provide an optimal routing path, and it may be preferable to select a remote alternate over a connected alternate. Some uses of tunnels to extend LFA [RFC5286] coverage are described in [RFC7490] and [TI-LFA]. [RFC7490] and [TI-LFA] present some use cases for LDP tunnels and Segment Routing tunnels, respectively. This document considers any type of tunneling techniques to reach remote alternates (IP, Generic Routing

Encapsulation (GRE), LDP, RSVP-TE, the Layer 2 Tunneling Protocol (L2TP), Segment Routing, etc.) and does not restrict the remote alternates to the uses presented in these other documents.

In Figure 1, there is no P router alternate for P8 to reach PE4 or PE5, so P8 is using PE2 as an alternate; this may generate congestion when FRR is activated. Instead, we could have a remote alternate for P8 to protect traffic to PE4 and PE5. For example, a tunnel from P8 to P3 (following the shortest path) can be set up, and P8 would be able to use P3 as a remote alternate to protect traffic to PE4 and PE5. In this scenario, traffic will not use a PE link during FRR activation.

When selecting the best alternate, the selection algorithm MUST consider all available alternates (connected or tunnel). For example, with remote LFAs, computation of PQ sets [RFC7490] SHOULD be performed before the selection of the best alternate.

6.2.2. Mandatory Criteria

An LFA implementation MUST support the following criteria:

- Non-candidate link: A link marked as "non-candidate" will never be used as an LFA.
- A primary next hop being protected by another primary next hop of the same prefix (ECMP case).
- Type of protection provided by the alternate: link protection or node protection. In the case of preference for node protection, an implementation SHOULD support fallback to link protection if node protection is not available.
- Shortest path: lowest IGP metric used to reach the destination.
- Shared Risk Link Groups (SRLGs) (as defined in Section 3 of [RFC5286]; see also Section 6.2.4.1 for more details).

6.2.3. Additional Criteria

An LFA implementation SHOULD support the following criteria:

- A downstream alternate: Preference for a downstream path over a non-downstream path SHOULD be configurable.
- Link coloring with "include", "exclude", and preference-based systems (see Section 6.2.4.2).
- Link bandwidth (see Section 6.2.4.3).
- Alternate preference / node coloring (see Section 6.2.4.4).

6.2.4. Evaluation of Criteria

6.2.4.1. SRLGs

Section 3 of [RFC5286] proposes the reuse of GMPLS IGP extensions to encode SRLGs [RFC5307] [RFC4203]. Section 3 of [RFC5286] also describes the algorithm to compute SRLG protection.

When SRLG protection is computed, an implementation SHOULD allow the following:

- Exclusion of alternates in violation of SRLGs.
- Maintenance of a preference system between alternates based on SRLG violations. How the preference system is implemented is out of scope for this document, but here are two examples:
  - Preference based on the number of violations. In this case, more violations = less preferred.
  - Preference based on violation cost. In this case, each SRLG violation has an associated cost. The lower violation costs are preferred.

When applying SRLG criteria, the SRLG violation check SHOULD be performed on sources to alternates as well as alternates to destination paths, based on the SRLG set of the primary path. In the case of remote LFAs, PQ-to-destination path attributes would be retrieved from the Shortest Path Tree (SPT) rooted at the PQ.

6.2.4.2. Link Coloring

Link coloring is a powerful system to control the choice of alternates. Link colors are markers that will allow the encoding of properties of a particular link. Protecting interfaces are tagged with colors. Protected interfaces are configured to include some colors with a preference level and exclude others.

Link color information SHOULD be signaled in the IGP, and administrative-group IGP extensions [RFC5305] [RFC3630] that are already standardized, implemented, and widely used SHOULD be used for encoding and signaling link colors.

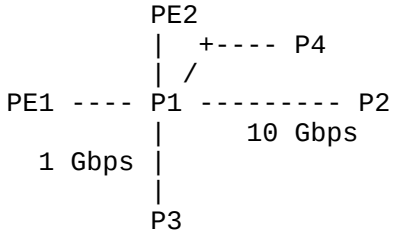


Figure 5

In the example in Figure 5, the P1 router is connected to three P routers and two PEs. P1 is configured to protect the P1-P4 link. We assume that, given the topology, all neighbors are candidate LFAs. We would like to enforce a policy in the network where only a core router may protect against the failure of a core link and where high-capacity links are preferred.

In this example, we can use the proposed link coloring by:

- Marking the PE links with the color RED.
- Marking the 10 Gbps core link with the color BLUE.
- Marking the 1 Gbps core link with the color YELLOW.
- Configuring the protected interface P1->P4 as follows:
  - Include BLUE, preference 200.
  - Include YELLOW, preference 100.
  - Exclude RED.

Using this, PE links will never be used to protect against P1-P4 link failure, and the 10 Gbps link will be preferred.

The main advantage of this solution is that it can easily be duplicated on other interfaces and other nodes without change. A service provider has only to define the color system (associate a color with a level of significance), as it is done already for TE affinities or BGP communities.

An implementation of link coloring:

- SHOULD support multiple "include" and "exclude" colors on a single protected interface.
- SHOULD provide a level of preference between included colors.
- SHOULD support the configuration of multiple colors on a single protecting interface.

6.2.4.3. Bandwidth

As mentioned in previous sections, not taking into account the bandwidth of an alternate could lead to congestion during FRR activation. We propose that the bandwidth criteria be based on the link speed information, for the following reasons:

- If a router S has a set of X destinations primarily forwarded to N, using per-prefix LFAs may lead to having a subset of X protected by a neighbor N1, another subset by N2, another subset by Nx, etc.
- S is not aware of traffic flows to each destination, so in the case of FRR activation, S is not able to evaluate how much traffic will be sent to N1, N2, Nx, etc.

Based on this, it is not useful to gather available bandwidth on alternate paths, as the router does not know how much bandwidth it requires for protection. The proposed link speed approach provides a good approximation at low cost, as information is easily available.

The bandwidth criteria of the policy framework SHOULD work in at least the following two ways:

- Prune: Exclude an LFA if the link speed to reach it is lower than the link speed of the primary next-hop interface.
- Prefer: Prefer an LFA based on its bandwidth to reach it compared to the link speed of the primary next-hop interface.

6.2.4.4. Alternate Preference / Node Coloring

Rather than tagging interfaces on each node (using link colors) to identify the types of alternate nodes (as an example), it would be helpful if routers could be identified in the IGP. This would allow grouped processing on multiple nodes. As an implementation needs to exclude some specific alternates (see Section 6.2.3), an implementation SHOULD be able to:

- give preference to a specific alternate.
- give preference to a group of alternates.
- exclude a specific alternate.
- exclude a group of alternates.

A specific alternate may be identified by its interface, IP address, or router ID, and a group of alternates may be identified by a marker (tag) advertised in IGP. The IGP encoding and signaling for marking groups of alternates SHOULD be done according to [RFC7917] and [RFC7777]. Using a tag/marker is referred to as "node coloring", as compared to the link coloring option presented in Section 6.2.4.2.

Consider the following network:

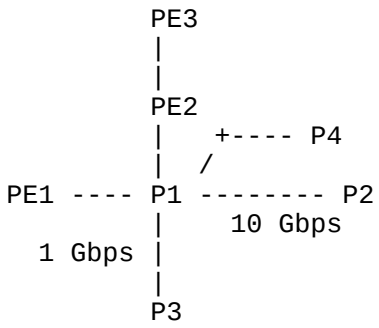


Figure 6

In the example above, each node is configured with a specific tag flooded through the IGP.

- PE1,PE3: 200 (non-candidate).
- PE2: 100 (edge/core).
- P1,P2,P3: 50 (core).

A simple policy could be configured on P1 to choose the best alternate for P1->P4 based on the function or role of the router, as follows:

- criterion 1 -> alternate preference: exclude tags 100 and 200.
- criterion 2 -> bandwidth.

6.2.5. Retrieving Alternate Path Attributes

6.2.5.1. Alternate Path

The alternate path is composed of two distinct parts: PLR to alternate and alternate to destination.

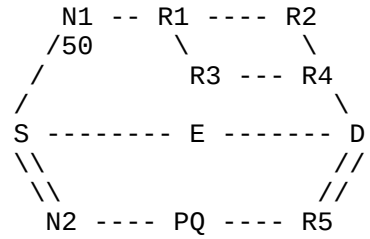


Figure 7

In Figure 7, we consider a primary path from S to D, with S using E as the primary next hop. All metrics are 1, except that {S,N1} = 50. Two alternate paths are available:

- {S,N1,R1,R2|R3,R4,D}, where N1 is a connected alternate. This consists of two sub-paths:
  - {S,N1}: path from the PLR to the alternate.
  - {N1,R1,R2|R3,R4,D}: path from the alternate to the destination.
- {S,N2,PQ,R5,D}, where the PQ is a remote alternate. Again, the path consists of two sub-paths:
  - {S,N2,PQ}: path from the PLR to the alternate.
  - {PQ,R5,D}: path from the alternate to the destination.

As displayed in Figure 7, some parts of the alternate path may fan out to multiple paths due to ECMP.

6.2.5.2. Alternate Path Attributes

Some criteria listed in the previous sections require the retrieval of some characteristics of the alternate path (SRLG, bandwidth, color, tag, etc.). We call these characteristics "path attributes". A path attribute can record a list of node properties (e.g., node tag) or link properties (e.g., link color).

This document defines two types of path attributes:

- Cumulative attribute: When a path attribute is cumulative, the implementation SHOULD record the value of the attribute on each element (link and node) along the alternate path. SRLG, link color, and node color are cumulative attributes.

- Unitary attribute: When a path attribute is unitary, the implementation SHOULD record the value of the attribute only on the first element along the alternate path (first node, or first link). Bandwidth is a unitary attribute.

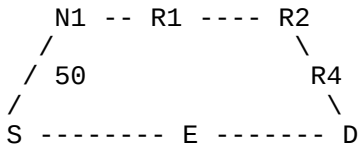


Figure 8

In Figure 8, N1 is a connected alternate to reach D from S. We consider that all links have a RED color except {R1,R2}, which is BLUE. We consider all links to be 10 Gbps except {N1,R1}, which is 2.5 Gbps. The bandwidth attribute collected for the alternate path will be 10 Gbps. As the attribute is unitary, only the link speed of the first link {S,N1} is recorded. The link color attribute collected for the alternate path will be {RED,RED,BLUE,RED,RED}. As the attribute is cumulative, the value of the attribute on each link along the path is recorded.

6.2.5.3. Connected Alternate

For an alternate path using a connected alternate:

- Attributes from the PLR to the alternate are retrieved from the interface connected to the alternate. If the alternate is connected through multiple interfaces, the evaluation of attributes SHOULD be done once per interface (each interface is considered as a separate alternate) and once per ECMP group of interfaces (Layer 3 bundle).

- Path attributes from the alternate to the destination are retrieved from the SPT rooted at the alternate. As the alternate is a connected alternate, the SPT has already been computed to find the alternate, so there is no need for additional computation.

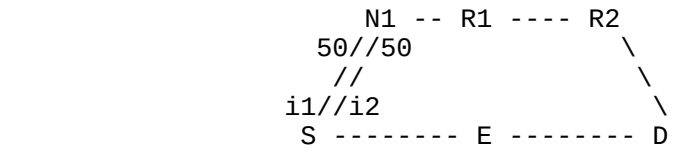


Figure 9

In Figure 9, we consider a primary path from S to D, with S using E as the primary next hop. All metrics are considered as 1 except {S,N1} links, which are using a metric of 50. We consider the following SRLGs on links:

- {S,N1} using i1: SRLG1,SRLG10.
- {S,N1} using i2: SRLG2,SRLG20.
- {N1,R1}: SRLG3.
- {R1,R2}: SRLG4.
- {R2,D}: SRLG5.
- {S,E}: SRLG10.
- {E,D}: SRLG6.

S is connected to the alternate using two interfaces: i1 and i2.

If i1 and i2 are not part of an ECMP group, the evaluation of attributes is done once per interface, and each interface is considered as a separate alternate path. Two alternate paths will be available with the associated SRLG attributes:

- Alternate path #1: {S,N1 using if1,R1,R2,D}: SRLG1,SRLG10,SRLG3,SRLG4,SRLG5.
- Alternate path #2: {S,N1 using if2,R1,R2,D}: SRLG2,SRLG20,SRLG3,SRLG4,SRLG5.

Alternate path #1 is sharing risks with the primary path and may be pruned, or its preference may be revoked, per user-defined policy.

If i1 and i2 are part of an ECMP group, the evaluation of attributes is done once per ECMP group, and the implementation considers a single alternate path {S,N1 using if1|if2,R1,R2,D} with the following SRLG attributes: SRLG1,SRLG10,SRLG2,SRLG20,SRLG3,SRLG4,SRLG5. The alternate path is sharing risks with the primary path and may be pruned, or its preference may be revoked, per user-defined policy.

6.2.5.4. Remote Alternate

For alternate path using a remote alternate (tunnel):

- Attributes on the path from the PLR to the alternate are retrieved using the PLR's primary SPT (when using a PQ node from the P-space) or the immediate neighbor's SPT (when using a PQ from the extended P-space). These are then combined with the attributes of the link(s) to reach the immediate neighbor. In both cases, no additional SPT is required.

- Attributes from the remote alternate to the destination path may be retrieved from the SPT rooted at the remote alternate. An additional forward SPT is required for each remote alternate (PQ node), as indicated in Section 2.3.2 of [REMOTE-LFA-NODE]. In some remote-alternate scenarios, like [TI-LFA], alternate-to-destination path attributes may be obtained using a different technique.

The number of remote alternates may be very high. In the case of remote LFAs, simulations of real-world network topologies have shown that as many as hundreds of PQs are possible. The computational overhead of collecting all path attributes of all such PQs to destination paths could grow beyond reasonable levels.

To handle this situation, implementations need to limit the number of remote alternates to be evaluated to a finite number before collecting alternate path attributes and running the policy evaluation. Section 2.3.3 of [REMOTE-LFA-NODE] provides a way to reduce the number of PQs to be evaluated.

Some other remote alternate techniques using static or dynamic tunnels may not require this pruning.

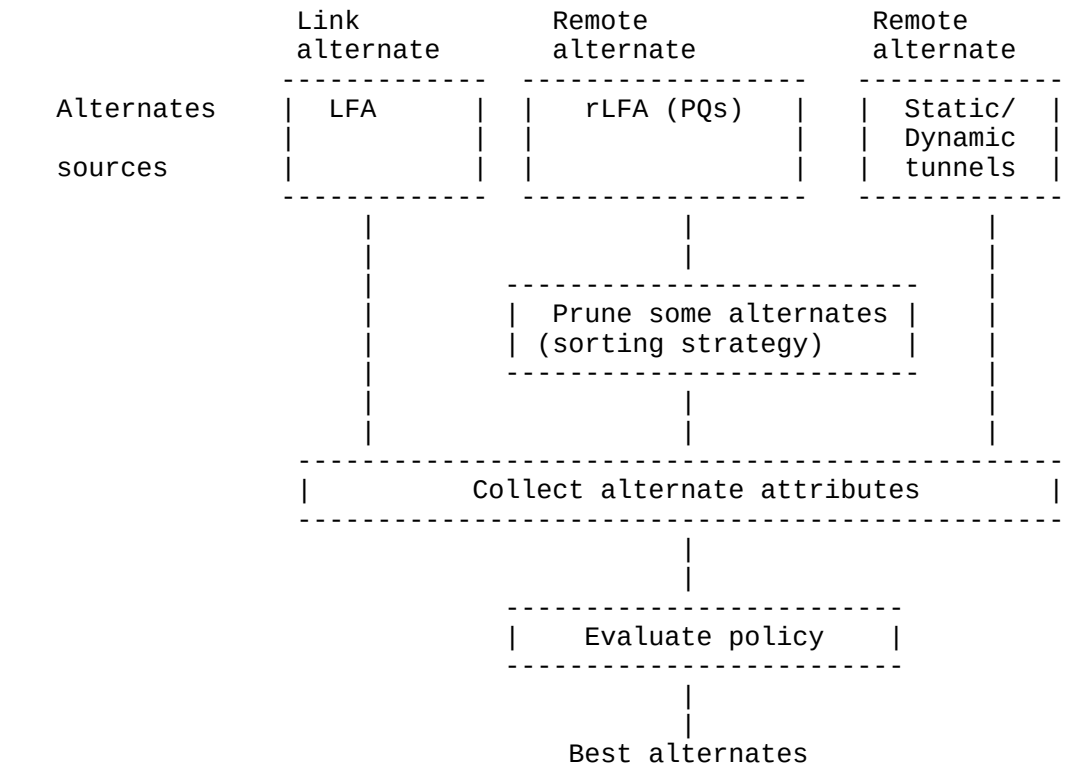


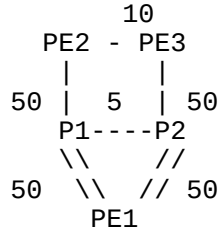
Figure 10

6.2.5.5. Collecting Attributes in the Case of Multiple Paths

As described in Section 6.2.5, there may be some situations where an alternate path or part of an alternate path fans out to multiple paths (e.g., ECMP). When collecting path attributes in such a case, an implementation SHOULD consider the union of attributes of each sub-path.

In Figure 7 (in Section 6.2.5.1), S has two alternate paths to reach D. Each alternate path fans out to multiple paths due to ECMP. Consider the following link color attributes: all links are RED except {R1,R3}, which is BLUE. The user wants to use an alternate path with only RED links. The first alternate path {S,N1,R1,R2|R3,R4,D} does not fit the constraint, as {R1,R3} is BLUE. The second alternate path {S,N2,PQ,R5,D} fits the constraint and will be preferred, as it uses only RED links.

6.2.6. ECMP LFAs



Links between P1 and PE1 are L1 and L2.
Links between P2 and PE1 are L3 and L4.

Figure 11

In Figure 11, the primary path from PE1 to PE2 is through P1, using ECMP on two parallel links -- L1 and L2. In the case of standard ECMP behavior, if L1 is failing, the post-convergence next hop would become L2 and ECMP would no longer be in use. If an LFA is activated, as stated in Section 3.4 of [RFC5286], "alternate next-hops may themselves also be primary next-hops, but need not be" and "alternate next-hops should maximize the coverage of the failure cases." In this scenario, there is no alternate providing node protection, so PE1 will prefer L2 as the alternate to protect L1; this makes sense compared to post-convergence behavior.

Consider a different scenario, again referring to Figure 11, where L1 and L2 are configured as a Layer 3 bundle using a local feature and L3/L4 comprise a second Layer 3 bundle. Layer 3 bundles are configured as if a link in the bundle is failing; the traffic must be rerouted out of the bundle. Layer 3 bundles are generally introduced to increase bandwidth between nodes. In a nominal situation, ECMP is still available from PE1 to PE2, but if L1 is failing, the post-convergence next hop would become the ECMP on L3 and L4. In this case, LFA behavior SHOULD be adapted in order to reflect the bandwidth requirement.

We would expect the following FIB entry on PE1:

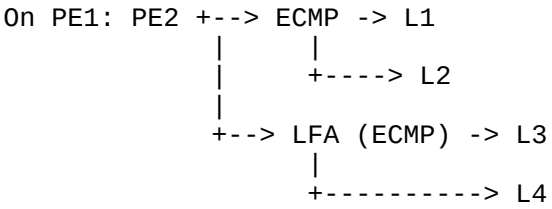


Figure 12

If L1 or L2 is failing, traffic must be switched on the LFA ECMP bundle rather than using the other primary next hop.

As mentioned in Section 3.4 of [RFC5286], protecting a link within an ECMP by another primary next hop is not a MUST. Moreover, as already discussed in this document, maximizing coverage against the failure cases may not be the right approach, and a policy-based choice of an alternate may be preferred.

An implementation SHOULD allow setting a preference to protect a primary next hop with another primary next hop. An implementation SHOULD also allow setting a preference to protect a primary next hop with a NON-primary next hop. An implementation SHOULD allow the use of an ECMP bundle as an LFA.

7. Operational Aspects

7.1. No-Transit Condition on LFA Computing Node

In Section 3.5 of [RFC5286], the setting of the no-transit condition (through the IS-IS overload bit or the OSPF R-bit) in an LFA computation is only taken into account for the case where a neighbor has the no-transit condition set.

In addition to Inequality 1 (Loop-Free Criterion) (Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D)) [RFC5286], the IS-IS overload bit or the OSPF R-bit of the LFA calculating neighbor (S) SHOULD be taken into account. Indeed, if it has the IS-IS overload bit set or the OSPF R-bit clear, no neighbor will loop traffic back to itself.

An OSPF router acting as a stub router [RFC6987] SHOULD behave as if the R-bit was clear regarding the LFA computation.

7.2. Manual Triggering of FRR

Service providers often perform manual link shutdown (using a router's command-line interface (CLI)) to perform network changes/tests. A manual link shutdown may be done at multiple levels: physical interface, logical interface, IGP interface, Bidirectional Forwarding Detection (BFD) session, etc. In particular, testing or troubleshooting FRR requires that manual shutdown be performed on the remote end of the link, as a local shutdown would not generally trigger FRR.

To permit such a situation, an implementation SHOULD support triggering/activating LFA FRR for a given link when a manual shutdown is done on a component that currently supports FRR activation.

An implementation MAY also support FRR activation for a specific interface or a specific prefix on a primary next-hop interface and revert without any action on any running component of the node (links or protocols). In this use case, the FRR activation time needs to be controlled by a timer in case the operator forgot to revert the traffic to the primary path. When the timer expires, the traffic is automatically reverted to the primary path. This will simplify the testing of the FRR path; traffic can then be reverted back to the primary path without causing a global network convergence.

For example:

- If an implementation supports FRR activation upon a BFD session-down event, that implementation SHOULD support FRR activation when a manual shutdown is done on the BFD session. But if an implementation does not support FRR activation upon a BFD session-down event, there is no need for that implementation to support FRR activation upon manual shutdown of a BFD session.

- If an implementation supports FRR activation upon a physical link-down event (e.g., Rx laser "off" detection, error threshold raised), that implementation SHOULD support FRR activation when a manual shutdown of a physical interface is done. But if an implementation does not support FRR activation upon a physical link-down event, there is no need for that implementation to support FRR activation upon manual shutdown of a physical link.

- A CLI command may allow switching from the primary path to the FRR path to test the FRR path for a specific interface or prefix. There is no impact on the control plane; only the data plane of the local node may be changed. A similar command may allow switching traffic back from the FRR path to the primary path.

7.3. Required Local Information

The introduction of LFAs in a network requires some enhancements to standard routing information provided by implementations. Moreover, due to "non-100%" coverage, coverage information is also required.

Hence, an implementation:

- MUST be able to display, for every prefix, the primary next hop as well as the alternate next-hop information.
- MUST provide coverage information per LFA activation domain (area, level, topology, instance, virtual router, address family, etc.).
- MUST provide the number of protected prefixes as well as non-protected prefixes globally.
- SHOULD provide the number of protected prefixes as well as non-protected prefixes per link.
- MAY provide the number of protected prefixes as well as non-protected prefixes per priority if the implementation supports prefix-priority insertion in the RIB/FIB.
- SHOULD provide a reason for choosing an alternate (policy and criteria) and for excluding an alternate.
- SHOULD provide the list of non-protected prefixes and the reason why they are not protected (e.g., no protection required, no alternate available).

7.4. Coverage Monitoring

It is pretty easy to evaluate the coverage of a network in a nominal situation, but topology changes may change the level of coverage. In some situations, the network may no longer be able to provide the required level of protection. Hence, it becomes very important for service providers to receive alerts regarding changes in coverage.

An implementation SHOULD:

- provide an alert system if total coverage (for a node) is below a defined threshold or when coverage returns to normal.
- provide an alert system if coverage for a specific link is below a defined threshold or when coverage returns to normal.

An implementation MAY:

- trigger an alert if a specific destination is not protected anymore or when protection comes back up for this destination.

Although the procedures for providing alerts are beyond the scope of this document, we recommend that implementations consider standard and well-used mechanisms like syslog or SNMP traps.

7.5. LFAs and Network Planning

The operator may choose to run simulations in order to ensure a certain type of full coverage for the whole network or a given subset of the network. This is particularly likely if he operates the network in the sense of the third backbone profile described in Section 4 of [RFC6571]; that is, he seeks to design and engineer the network topology in such a way that a certain level of coverage is always achieved. Obviously, a complete and exact simulation of the IP FRR coverage can only be achieved if the behavior is deterministic and the algorithm used is available to the simulation tool. Thus, an implementation SHOULD:

- Behave deterministically in its LFA selection process. That is, in the same topology and with the same policy configuration, the implementation MUST always choose the same alternate for a given prefix.

- Document its behavior. The implementation SHOULD provide enough documentation regarding its behavior to allow an implementer of a simulation tool to foresee the exact choice of the LFA implementation for every prefix in a given topology. This SHOULD take into account all possible policy configuration options. One possible way to document this behavior is to disclose the algorithm used to choose alternates.

8. Security Considerations

The policy mechanism introduced in this document allows the tuning of the selection of the alternate. This is not seen as a security threat, because:

- all candidates are already eligible as per [RFC5286] and considered usable.
- the policy is based on information from the router's own configuration and from the IGP, both of which are considered trusted.

Hence, this document does not introduce any new security considerations as compared to [RFC5286].

As noted above, the policy mechanism introduced in this document allows the tuning of the selection of the best alternate but does not change the list of alternates that are eligible. As described in Section 7 of [RFC5286], this best alternate "can be used anyway when a different topological change occurs, and hence this can't be viewed as a new security threat."

9. References

9.1. Normative References

[ISO10589] International Organization for Standardization, "Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO Standard 10589, 2002.

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

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <http://www.rfc-editor.org/info/rfc3630>.

[RFC4203] Kompella, K., Ed., and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <http://www.rfc-editor.org/info/rfc4203>.

[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, <http://www.rfc-editor.org/info/rfc5286>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <http://www.rfc-editor.org/info/rfc5305>.

[RFC5307] Kompella, K., Ed., and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <http://www.rfc-editor.org/info/rfc5307>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <http://www.rfc-editor.org/info/rfc5340>.

[RFC6571] Filsfils, C., Ed., Francois, P., Ed., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, DOI 10.17487/RFC6571, June 2012, <http://www.rfc-editor.org/info/rfc6571>.

[RFC6987] Retana, A., Nguyen, L., Zinin, A., White, R., and D. McPherson, "OSPF Stub Router Advertisement", RFC 6987, DOI 10.17487/RFC6987, September 2013, <http://www.rfc-editor.org/info/rfc6987>.

[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, <http://www.rfc-editor.org/info/rfc7490>.

[RFC7777] Hegde, S., Shakir, R., Smirnov, A., Li, Z., and B. Decraene, "Advertising Node Administrative Tags in OSPF", RFC 7777, DOI 10.17487/RFC7777, March 2016, <http://www.rfc-editor.org/info/rfc7777>.

[RFC7917] Sarkar, P., Ed., Gredler, H., Hegde, S., Litkowski, S., and B. Decraene, "Advertising Node Administrative Tags in IS-IS", RFC 7917, DOI 10.17487/RFC7917, July 2016, <http://www.rfc-editor.org/info/rfc7917>.

9.2. Informative References

[REMOTE-LFA-NODE]
Sarkar, P., Ed., Hegde, S., Bowers, C., Gredler, H., and S. Litkowski, "Remote-LFA Node Protection and Manageability", Work in Progress, draft-ietf-rtgwg-rlfa-node-protection-05, December 2015.

[SEG-RTG-ARCH]
Filsfils, C., Ed., Previdi, S., Ed., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", Work in Progress, draft-ietf-spring-segment-routing-09, July 2016.

[TI-LFA] Francois, P., Filsfils, C., Bashandy, A., Decraene, B., and S. Litkowski, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, draft-francois-segment-routing-ti-lfa-00, November 2013.

Contributors

Significant contributions were made by Pierre Francois, Hannes Gredler, Chris Bowers, Jeff Tantsura, Uma Chunduri, Acee Lindem, and Mustapha Aissaoui, whom the authors would like to acknowledge.

Authors' Addresses

Stephane Litkowski (editor)
Orange

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Clarence Filsfils
Cisco Systems

Email: cfilsfil@cisco.com

Kamran Raza
Cisco Systems

Email: skraza@cisco.com

Martin Horneffer
Deutsche Telekom

Email: Martin.Horneffer@telekom.de

Pushpasis Sarkar
Individual Contributor

Email: pushpasis.ietf@gmail.com