Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 27, 2015

S. Litkowski, Ed.
B. Decraene
Orange
C. Filsfils
K. Raza
Cisco Systems
M. Horneffer
Deutsche Telekom
P. Sarkar
Juniper Networks
June 25, 2015

Operational management of Loop Free Alternates
draft-ietf-rtgwg-lfa-manageability-11

Abstract

Loop Free Alternates (LFA), as defined in RFC 5286 is an IP Fast ReRoute (IP FRR) mechanism enabling traffic protection for IP traffic (and MPLS LDP traffic by extension). Following first deployment experiences, this document provides operational feedback on LFA, 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 solution.

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].

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 http://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 December 27, 2015.

Copyright Notice

Table of Contents

1. Introduction

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

Section 3 provides real uses 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 express requirements for the operational management of LFA and especially a policy framework to manage alternates.

Section 7 details some operational considerations of LFA like IS-IS overload bit management or troubleshooting informations.

2. Definitions

- Per-prefix LFA : LFA computation, and best alternate evaluation is done for each destination prefix, as opposed to "Per-next hop" simplification also proposed in [RFC5286] Section 3.8.
- PE router : Provider Edge router. These routers are connecting customers
- 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 by P routers and links between them.
- Core link : network link part of the core network i.e. a P router to P router link.
- Link-protecting LFA : alternate providing protection against link failure.
- Node-protecting LFA : alternate providing protection against node failure.

- Connected alternate : alternate adjacent (at IGP level) to the point of local repair (i.e. an IGP neighbor).

- Remote alternate : alternate which is does not share an IGP adjacency with the point of local repair.

3. Operational issues with default LFA tie breakers

[RFC5286] introduces the notion of tie breakers when selecting the LFA among multiple candidate alternate next-hops. When multiple LFA exist, RFC 5286 has favored the selection of the LFA providing the best coverage of the failure cases. While this is indeed a goal, this is one among multiple and in some deployment this lead to the selection of a suboptimal LFA. The following sections details real use cases of such limitations.

Note that the use case of 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 failures within core network

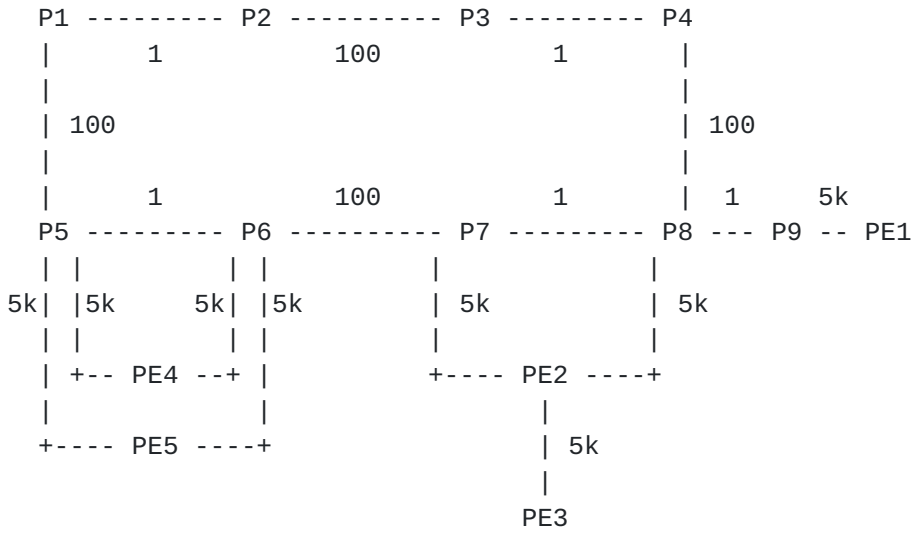


Figure 1

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

In figure 1, let us consider the traffic flowing from PE1 to PE4. The nominal path is P9-P8-P7-P6-PE4. Let us consider the failure of link P7-P8. As P4 primary path to PE4 is P8-P7-P6-PE4, P4 is not an LFA for P8 (because P4 will loop back traffic 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 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 was not directly affected by the failure, its links become congested and its traffic will suffer from the congestion.

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

- From PE2 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 situation, traffic not affected by the failure may be affected by the congestion.
- From P8 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 an Edge router as an alternate to protect a core failure, a service provider may consider this as a bad routing design and would like to prevent it.

3.2. Case 2: PE router choosen to protect core failures while P router LFA exists

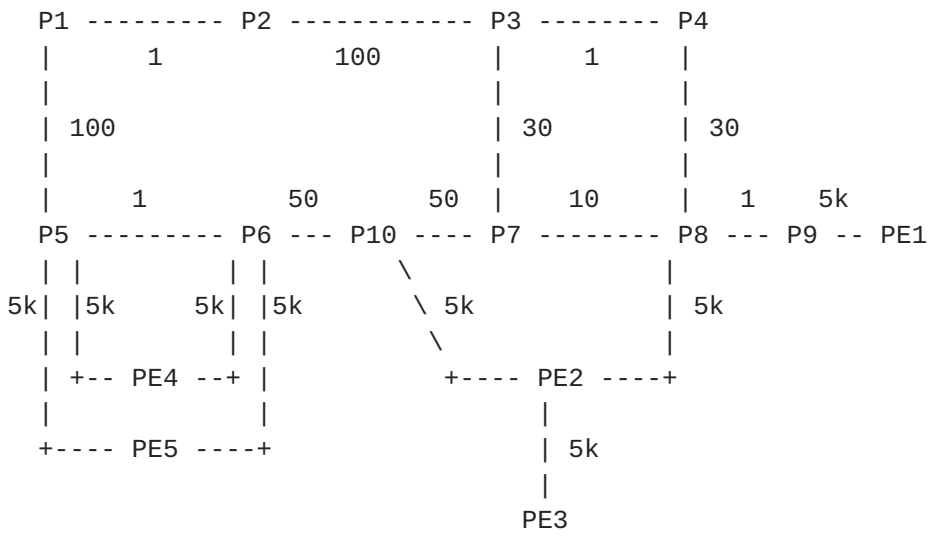


Figure 2

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

In the figure 2, let us consider the traffic coming from PE1 to PE4. Nominal path is P9-P8-P7-P10-P6-PE4. Let us 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 best LFA due to its better protection type. Just like in case 1, this may lead to congestion on PE2 links upon LFA activation.

3.3. Case 3: suboptimal P router alternate choice

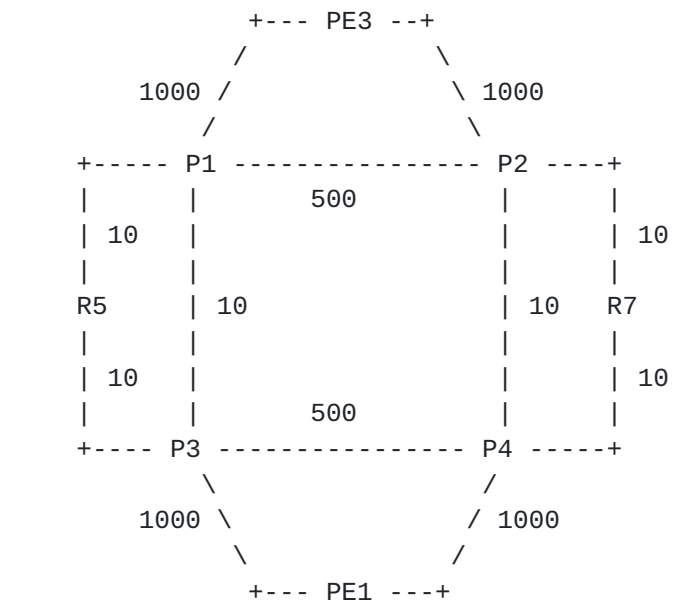


Figure 3

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

In the figure above, 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 best LFA due to its better protection type. However, it may not be desirable to use P4 for bandwidth capacity reason. A service provider may prefer to use high bandwidth links as prefered LFA. In this example, prefering shortest path over protection type may achieve the expected behavior, but in cases where metric are not reflecting bandwidth, it 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

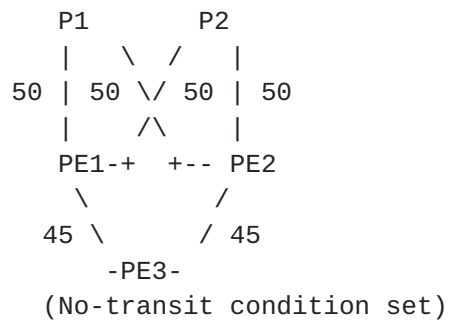


Figure 4

IS-IS and OSPF protocols define some way to prevent a router to be used as transit.

IS-IS overload bit is defined in [ISO10589] and OSPF R-bit is defined in [RFC5340]. OSPF Stub Router is also defined in [RFC6987] as a method to prevent transit on a node by advertising MaxLinkMetric on all non stub links.

In the figure above, PE3 has its no-transit condition set (permanently, for design reason) and wants to protect traffic using 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 case of no-transit condition set on a node, LFA behavior must be clarified.

4. Need for coverage monitoring

As per [RFC6571], LFA coverage highly depends on the used network topology. Even if remote LFA ([RFC7490]) extends significantly the coverage of the basic LFA specification, there is still some cases where protection would not be available. As network topologies are constantly evolving (network extension, capacity addings, latency optimization etc.), the protection coverage may change. Fast reroute functionality may be critical for some services supported by the network, a service provider must constantly know what protection coverage is currently available on the network. Moreover, predicting the protection coverage in case of network topology change is mandatory.

Today network simulation tool associated with whatif scenarios functionality are often used by service providers for the overall

network design (capacity, path optimization etc.). Section 7.5, Section 7.4 and Section 7.3 of this document propose to add LFA informations into such tool and within routers, so a service provider may be able :

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

Documentation of LFA selection algorithms by implementers (default and tuning options) is important in order to leave possibility for 3rd party modules to model these policy-LFA expressions.

5. Need for LFA activation granularity

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

Moreover a service provider may implement multiple different FRR mechanism in its networks for different usages (MRT, TE FRR). In this scenario, an implementation MAY allow to compute alternates for a specific destination even if the destination is already protected by another mechanism. This will bring redundancy and let the ability for the operator to select the best option for FRR using a policy language.

Section 6 of this document propose some implementation guidelines.

6. Configuration requirements

Controlling best alternate and LFA activation granularity is a requirement for Service Providers. This section defines configuration requirements for LFA.

6.1. LFA enabling/disabling scope

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

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

- Per routing context: VRF, virtual/logical router, global routing table, etc.
- Per interface
- Per protocol instance, topology, area
- Per prefixes: 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, LFA MUST be computed and installed for this prefix even if the primary outgoing interface is not configured for protection.

An implementation of 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 decision from the IGP routing protocol, MPLS dataplane may be protected by LFA. The implementation may allow operator to control this inheritance of protection from the IP prefix to the MPLS label bound to this prefix. The protection inheritance will concern : IP to MPLS, MPLS to MPLS, and MPLS to IP entries. As example, LDP and segment-routing extensions for ISIS and OSPF are control plane eligible to this inheritance of protection.

6.2. Policy based LFA selection

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

1. An implementation of LFA SHOULD support policy-based decision for determining the best LFA.
2. Policy based decision SHOULD be based on multiple criterions, with each criteria 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. An implementation SHOULD also support election of multiple LFAs, for loadbalancing purposes.

4. Policy SHOULD be applicable to a protected interface or to a specific set of destinations. In case of application on the protected interface, all destinations primarily routed on this interface SHOULD use the interface policy.

5. It is an implementation choice to reevaluate policy dynamically or not (in case of policy change). If a dynamic approach is chosen, the implementation SHOULD recompute the best LFAs and reinstall them in 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 vs remote alternates

In addition to connected LFAs, tunnels (e.g. IP, LDP, RSVP-TE or 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 usage of tunnels to extend LFA ([RFC5286]) coverage is described in either [RFC7490] or [I-D.francois-segment-routing-ti-lfa]. These documents present some use cases of LDP tunnels ([RFC7490]) or Segment Routing tunnels ([I-D.francois-segment-routing-ti-lfa]). This document considers any type of tunneling techniques to reach remote alternates (IP, GRE, LDP, RSVP-TE, L2TP, Segment Routing etc.) and does not restrict the remote alternates to the usage presented in the referenced document.

In figure 1, there is no P router alternate for P8 to reach PE4 or PE5 , so P8 is using PE2 as alternate, which 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 shortest path) can be setup and P8 would be able to use P3 as 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 LFA, computation of PQ set ([RFC7490]) SHOULD be performed before best alternate selection.

6.2.2. Mandatory criteria

An implementation of LFA MUST support the following criteria:

- Non candidate link: A link marked as "non candidate" will never be used as 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, node protection. In case of node protection preference, an implementation SHOULD support fall back to link protection if node protection is not available.
- Shortest path: lowest IGP metric used to reach the destination.
- SRLG (as defined in [RFC5286] Section 3, see also Section 6.2.4.1 for more details).

6.2.3. Additional criteria

An implementation of LFA SHOULD support the following criteria:

- Downstreamness of an alternate : preference of a downstream path over a non downstream path SHOULD be configurable.
- Link coloring with : include, exclude and preference based system (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. Criteria evaluation

6.2.4.1. SRLG

[RFC5286] Section 3. proposes to reuse GMPLS IGP extensions to encode Shared Risk Link Groups ([RFC4205] and [RFC4203]). The section is also describing the algorithm to compute SRLG protection.

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

- Exclusion alternates violating SRLG.

- Maintenance of a preference system between alternates based on SRLG violations. How the preference system is implemented is out of scope of this document but here are few examples :

  * Preference based on number of violations. In this case : the more violations = the less preferred.

  * Preference based on violation cost. In this case, each SRLG violation has an associated cost. The lower violation cost sum is preferred.

When applying SRLG criteria, the SRLG violation check SHOULD be performed on source to alternate as well as alternate to destination paths based on the SRLG set of the primary path. In the case of remote LFA, PQ to destination path attributes would be retrieved from SPT rooted at 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 to encode 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 signalled in the IGP and admin-groups IGP extensions ([RFC5305] and [RFC3630]) that are already standardized, implemented and widely-used, SHOULD be used for encoding and signalling link colors.

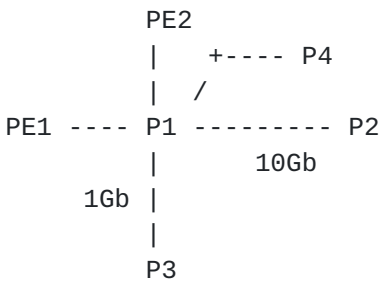


Figure 8

Example : 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 LFA. 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 prefered.

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

- Marking PEs links with color RED
- Marking 10Gb CORE link with color BLUE
- Marking 1Gb CORE link with color YELLOW
- Configured the protected interface P1->P4 with :
  - 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 10Gb link will be 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 color with a 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 multiple colors configuration on a single protecting interface.

6.2.4.3. Bandwidth

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

- if a router S has a set of X destinations primarly forwarded to N, using per prefix LFA may lead to have a subset of X protected by a neighbor N1, another subset by N2, another subset by Nx etc.

- S is not aware about traffic flows to each destination and is not able to evaluate how much traffic will be sent to N1,N2, etc. Nx in case of FRR activation.

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 with a small cost as information is easily available.

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

- PRUNE : exclude a LFA if link speed to reach it is lower than the link speed of the primary next hop interface.

- PREFER : prefer a 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 interface on each node (using link color) to identify alternate node type (as example), it would be helpful if routers could be identified in the IGP. This would allow a grouped processing on multiple nodes. As an implementation need to exclude some specific alternates (see Section 6.2.3), an implementation :

- SHOULD be able to give a preference to specific alternate.

- SHOULD be able to give a preference to a group of alternate.

- SHOULD be able to exclude a specific alternate.

- SHOULD be able to exclude a group of alternate.

A specific alternate may be identified by its interface, IP address or router ID and group of alternates may be identified by a marker (tag) advertised in IGP. The IGP encoding and signalling for marking group of alternates SHOULD be done using [I-D.ietf-isis-node-admin-tag], [I-D.ietf-ospf-node-admin-tag]. Using a tag/marker is referred as Node coloring in comparison to link coloring option presented in Section 6.2.4.2.

Consider the following network:

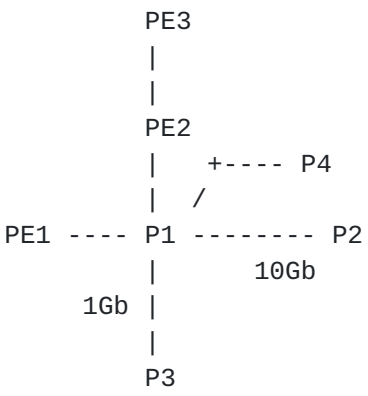


Figure 9

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 router function/role as follows :

- criteria 1 -> alternate preference: exclude tag 100 and 200.
- criteria 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.

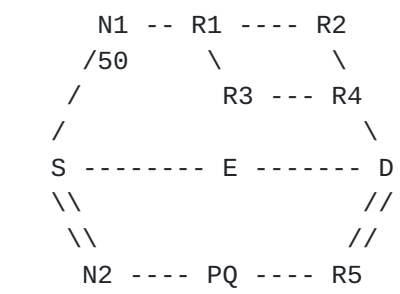


Figure 5

In the figure above, we consider a primary path from S to D, S using E as primary nexthop. All metrics are 1 except {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 PLR to the alternate.
  - {N1,R1,R2|R3,R4,D}: path from alternate to destination.
- {S,N2,PQ,R5,D} where PQ is a remote alternate. Again the path consists of two sub-paths:
  - {S,N2,PQ}: path from PLR to the alternate.
  - {PQ,R5,D}: path from alternate to destination.

As displayed in the figure, some part of the alternate path may fanout in multipath due to ECMP.

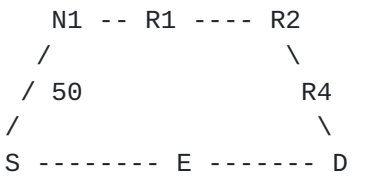
6.2.5.2. Alternate path attributes

Some criterions listed in the previous sections are requiring to retrieve some characteristic 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.

o 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.



In the figure above, N1 is a connected alternate to each D from S. We consider that all links have a RED color except {R1,R2} which is BLUE. We consider all links to be 10Gbps, except {N1,R1} which is 2.5Gbps. The bandwidth attribute collected for the alternate path will be 10Gbps. 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 alternate path using a connected alternate:

o attributes from PLR to alternate are retrieved from the interface connected to the alternate. In case 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).

o path attributes from alternate to destination are retrieved from SPF rooted at the alternate. As the alternate is a connected alternate, the SPF has already been computed to find the alternate, so there is no need of additional computation.

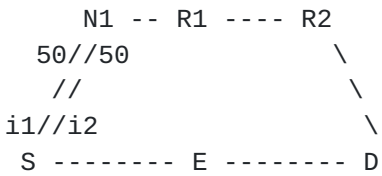


Figure 6

In the figure above, we consider a primary path from S to D, S using E as primary nexthop. All metrics are considered as 1 expect {S,N1}

links which are using metric of 50. We consider the following SRLG groups 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 primary path and may be depreferred or pruned by 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. Alternate path is sharing risks with primary path and may be depreferred or pruned by 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 alternate are retrieved using the PLR's primary SPF (when using a PQ-node from P-Space) or the immediate neighbor's SPF (when using a PQ from extended P-Space). These are then combined with the attributes of the

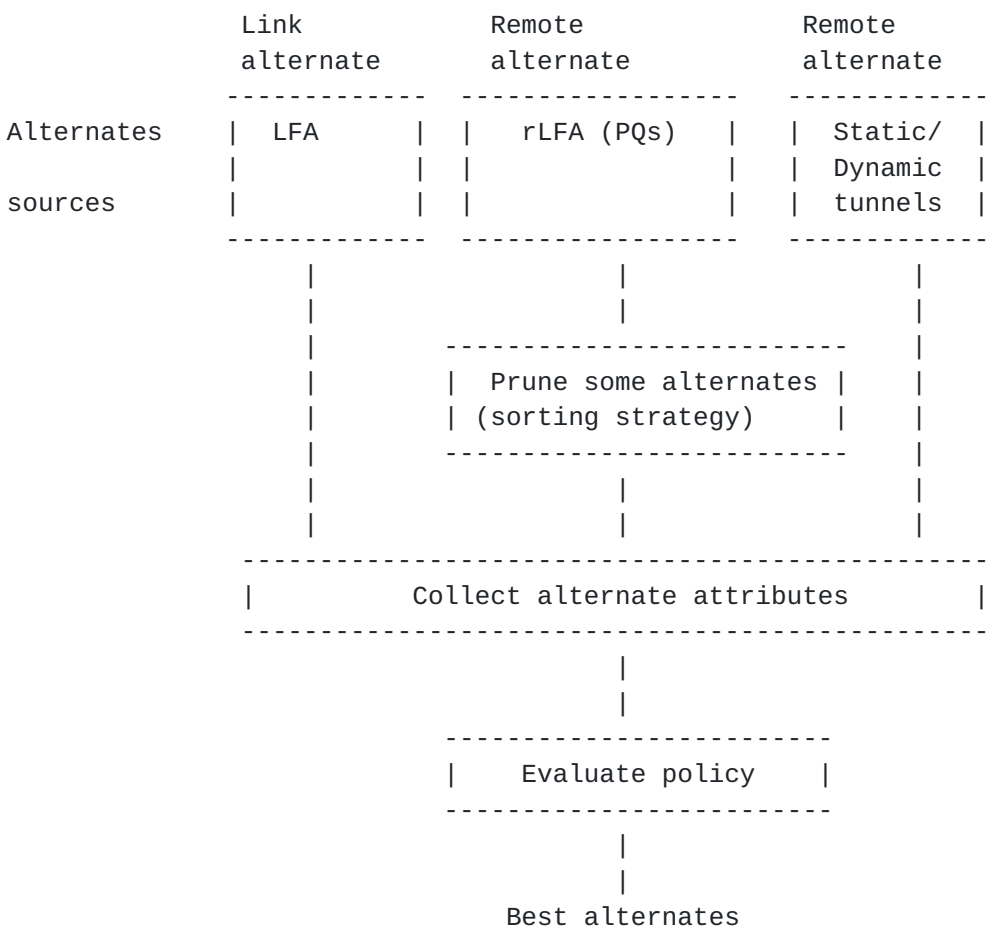
link(s) to reach the immediate neighbor. In both cases, no additional SPF is required.

- Attributes from remote alternate to destination path may be retrieved from SPF rooted at the remote alternate. An additional forward SPF is required for each remote alternate (PQ-node) as indicated in [I-D.ietf-rtgwg-rlfa-node-protection] section 3.2 . In some remote alternate scenarios, like [I-D.francois-segment-routing-ti-lfa], alternate to destination path attributes may be obtained using a different technique.

The number of remote alternates may be very high. . In case of remote LFA, simulations of real-world network topologies have shown that order of hundreths of PQ may be possible. The computational overhead to collect all path attributes of all PQ to destination paths may grow beyond practical reason.

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. [I-D.ietf-rtgwg-rlfa-node-protection] Section 2.3.3 provides a way to reduce the number of PQ to be evaluated.

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



6.2.5.5. Collecting attributes in case of multipath

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

In the figure 5 (in Section 6.2.5), S has two alternates paths to reach D. Each alternate path fans out into multipath due to ECMP. Considering 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

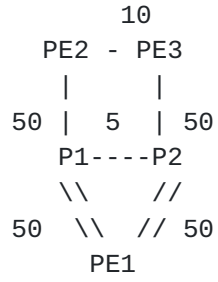


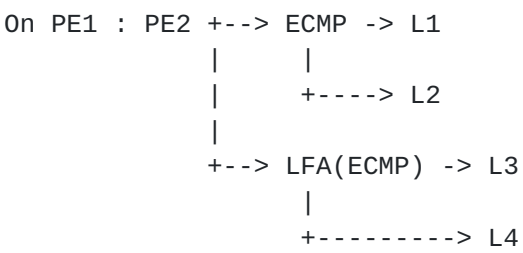
Figure 7

Links between P1 and PE1 are L1 and L2, links between P2 and PE1 are L3 and L4

In the figure above, primary path from PE1 to PE2 is through P1 using ECMP on two parallel links L1 and L2. In case of standard ECMP behavior, if L1 is failing, postconvergence next hop would become L2 and there would be no longer ECMP. If LFA is activated, as stated in [RFC5286] Section 3.4., "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, LFA will so prefer L2 as alternate to protect L1 which makes sense compared to postconvergence behavior.

Considering a different scenario using figure 7, where L1 and L2 are configured as a layer 3 bundle using a local feature, as well as L3/ L4 being 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 nominal situation, ECMP is still available from PE1 to PE2, but if L1 is failing, postconvergence next hop would become 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 :



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 [RFC5286] Section 3.4., protecting a link within an ECMP by another primary next hop is not a MUST. Moreover, we already presented in this document, that maximizing the coverage of the failure case may not be the right approach and policy based choice of alternate may be preferred.

An implementation SHOULD allow to prefer to protect a primary next hop by another primary next hop. An implementation SHOULD allow to prefer to protect a primary next hop by a NON primary next hop. An implementation SHOULD allow to use an ECMP bundle as a LFA.

7. Operational aspects

7.1. No-transit condition on LFA computing node

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

In addition to RFC 5286 inequality 1 Loop-Free Criterion (Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D)), the IS-IS overload bit or 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 OSPF R-bit clear, no neighbor will loop back to traffic to itself.

An OSPF router acting as a stub router [RFC 6987] SHOULD behave as if R-bit was clear regarding LFA computation.

7.2. Manual triggering of FRR

Service providers often perform manual link shutdown (using router CLI) to perform some network changes/tests. A manual link shutdown may be done at multiple level : physical interface, logical interface, IGP interface, BFD session etc. Especially testing or troubleshooting FRR requires to perform the manual shutdown on the remote end of the link as generally a local shutdown would not trigger FRR.

To enhance such situation, an implementation SHOULD support triggering/activating LFA Fast Reroute 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 need to be controlled by a timer in case the operator forgot to revert traffic on primary path. When the timer expires, the traffic is automatically reverted to the primary path. This will make easier tests of fast-reroute path and then revert back to the primary path without causing a global network convergence.

For example :

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

- if an implementation supports FRR activation on physical link down event (e.g. Rx laser Off detection, or error threshold raised etc.), this implementation SHOULD support FRR activation when a manual shutdown at physical interface is done. But if an implementation does not support FRR activation on physical link down event, there is no need for this implementation to support FRR activation on manual physical link shutdown.

- A CLI command may allow to switch from primary path to FRR path for testing FRR path for a specific. There is no impact on controlplane, only dataplane of the local node could be changed. A similar command may allow to switch back traffic from FRR path to primary path.

7.3. Required local information

LFA introduction requires some enhancement in standard routing information provided by implementations. Moreover, due to the non 100% coverage, coverage informations 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 activation domain of LFA (area, level, topology, instance, virtual router, address family etc.).
- MUST provide number of protected prefixes as well as non protected prefixes globally.
- SHOULD provide number of protected prefixes as well as non protected prefixes per link.
- MAY provide number of protected prefixes as well as non protected prefixes per priority if implementation supports prefix-priority insertion in 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 (no protection required or 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 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 get alerted about changes of coverage.

An implementation SHOULD :

- provide an alert system if total coverage (for a node) is below a defined threshold or comes back to a normal situation.
- provide an alert system if coverage of a specific link is below a defined threshold or comes back to a normal situation.

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. LFA and network planning

The operator may choose to run simulations in order to ensure full coverage of a certain type 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 profiles described in [RFC6571], that is, he seeks to design and engineer the network topology in a way that a certain 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 if the algorithm used is available to the simulation tool. Thus, an implementation SHOULD:

- Behave deterministic in its selection LFA process. I.e. 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 of its behavior that allows 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 to tune the selection of the alternate. This is not seen as a security threat as:

- all candidates are already eligible as per [RFC5286] and considered useable.

- the policy is based on information from the router's own configuration and from the IGP which are both considered trusted.

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

This document does not introduce any change in security consideration compared to [RFC5286]. The policy mechanism introduced in this document allow to tune the best alternate choice but does not change the list of alternates that are eligible. As defined in [RFC5286] Section 7., 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. IANA Considerations

This document has no action for IANA.

10. Contributors

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

11. References

11.1. Normative References

[I-D.ietf-isis-node-admin-tag]
Sarkar, P., Gredler, H., Hegde, S., Litkowski, S., Decraene, B., Li, Z., Aries, E., Rodriguez, R., and H. Raghuveer, "Advertising Per-node Admin Tags in IS-IS", draft-ietf-isis-node-admin-tag-02 (work in progress), June 2015.

[I-D.ietf-ospf-node-admin-tag]
Hegde, S., Raghuveer, H., Gredler, H., Shakir, R., Smirnov, A., Li, Z., and B. Decraene, "Advertising per-node administrative tags in OSPF", draft-ietf-ospf-node-admin-tag-02 (work in progress), June 2015.

[ISO10589]
"Intermediate system to Intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.

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

[RFC3137] Retana, A., Nguyen, L., White, R., Zinin, A., and D. McPherson, "OSPF Stub Router Advertisement", RFC 3137, June 2001.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

[RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

[RFC4205] Kompella, K. and Y. Rekhter, "Intermediate System to Intermediate System (IS-IS) Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4205, October 2005.

[RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

[RFC5307] Kompella, K. and Y. Rekhter, "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, October 2008.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

[RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

[RFC6987] Retana, A., Nguyen, L., Zinin, A., White, R., and D. McPherson, "OSPF Stub Router Advertisement", RFC 6987, September 2013.

[RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, April 2015.

11.2. Informative References

[I-D.francois-segment-routing-ti-lfa]
Francois, P., Filsfils, C., Bashandy, A., and B. Decraene, "Topology Independent Fast Reroute using Segment Routing", draft-francois-segment-routing-ti-lfa-00 (work in progress), November 2013.

[I-D.ietf-rtgwg-rlfa-node-protection]
Sarkar, P., Gredler, H., Hegde, S., Bowers, C., Litkowski, S., and H. Raghuveer, "Remote-LFA Node Protection and Manageability", draft-ietf-rtgwg-rlfa-node-protection-02 (work in progress), June 2015.

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
Juniper Networks

Email: psarkar@juniper.net