Network Working Group
Request for Comments: 5286
Category: Standards Track

A. Atlas, Ed.
BT
A. Zinin, Ed.
Alcatel-Lucent

September 2008

Basic Specification for IP Fast Reroute: Loop-Free Alternates

Status of This Memo

Abstract

This document describes the use of loop-free alternates to provide local protection for unicast traffic in pure IP and MPLS/LDP networks in the event of a single failure, whether link, node, or shared risk link group (SRLG). The goal of this technology is to reduce the packet loss that happens while routers converge after a topology change due to a failure. Rapid failure repair is achieved through use of precalculated backup next-hops that are loop-free and safe to use until the distributed network convergence process completes. This simple approach does not require any support from other routers. The extent to which this goal can be met by this specification is dependent on the topology of the network.

Table of Contents

1. Introduction

Applications for interactive multimedia services such as Voice over IP (VoIP) and pseudowires can be very sensitive to traffic loss, such as occurs when a link or router in the network fails. A router's convergence time is generally on the order of hundreds of milliseconds; the application traffic may be sensitive to losses greater than tens of milliseconds.

As discussed in [FRAMEWORK], minimizing traffic loss requires a mechanism for the router adjacent to a failure to rapidly invoke a repair path, which is minimally affected by any subsequent re-convergence. This specification describes such a mechanism that allows a router whose local link has failed to forward traffic to a pre-computed alternate until the router installs the new primary next-hops based upon the changed network topology. The terminology used in this specification is given in [FRAMEWORK]. The described mechanism assumes that routing in the network is performed using a link-state routing protocol -- OSPF [RFC2328] [RFC2740] [RFC5340] or IS-IS [RFC1195] [RFC2966] (for IPv4 or IPv6). The mechanism also assumes that both the primary path and the alternate path are in the same routing area.

When a local link fails, a router currently must signal the event to its neighbors via the IGP, recompute new primary next-hops for all affected prefixes, and only then install those new primary next-hops into the forwarding plane. Until the new primary next-hops are installed, traffic directed towards the affected prefixes is discarded. This process can take hundreds of milliseconds.

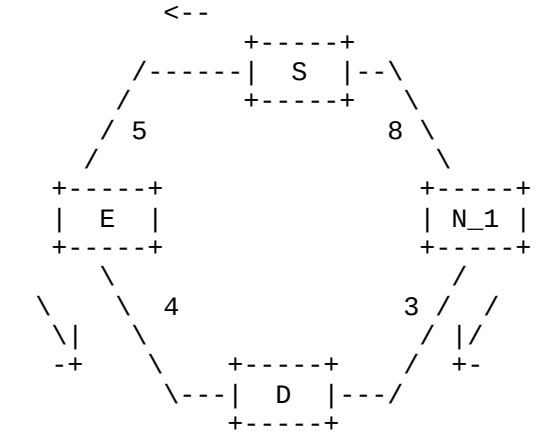

Figure 1: Basic Topology

The goal of IP Fast Reroute (IPFRR) is to reduce failure reaction time to 10s of milliseconds by using a pre-computed alternate next-hop, in the event that the currently selected primary next-hop fails, so that the alternate can be rapidly used when the failure is detected. A network with this feature experiences less traffic loss and less micro-looping of packets than a network without IPFRR. There are cases where traffic loss is still a possibility since IPFRR coverage varies, but in the worst possible situation a network with IPFRR is equivalent with respect to traffic convergence to a network without IPFRR.

To clarify the behavior of IP Fast Reroute, consider the simple topology in Figure 1. When router S computes its shortest path to router D, router S determines to use the link to router E as its primary next-hop. Without IP Fast Reroute, that link is the only next-hop that router S computes to reach D. With IP Fast Reroute, S also looks for an alternate next-hop to use. In this example, S would determine that it could send traffic destined to D by using the link to router N_1 and therefore S would install the link to N_1 as its alternate next-hop. At some later time, the link between router S and router E could fail. When that link fails, S and E will be the first to detect it. On detecting the failure, S will stop sending traffic destined for D towards E via the failed link, and instead send the traffic to S's pre-computed alternate next-hop, which is the link to N_1, until a new SPF is run and its results are installed. As with the primary next-hop, an alternate next-hop is computed for each destination. The process of computing an alternate next-hop does not alter the primary next-hop computed via a standard SPF.

If in the example of Figure 1, the link cost from N_1 to D increased to 30 from 3, then N_1 would not be a loop-free alternate, because the cost of the path from N_1 to D via S would be 17 while the cost from N_1 directly to D would be 30. In real networks, we may often face this situation. The existence of a suitable loop-free alternate next-hop is dependent on the topology and the nature of the failure for which the alternate is calculated.

This specification uses the terminology introduced in [FRAMEWORK]. In particular, it uses Distance_opt(X,Y), abbreviated to D_opt(X,Y), to indicate the shortest distance from X to Y. S is used to indicate the calculating router. N_i is a neighbor of S; N is used as an abbreviation when only one neighbor is being discussed. D is the destination under consideration.

A neighbor N can provide a loop-free alternate (LFA) if and only if

$$Distance\_opt(N, D) < Distance\_opt(N, S) + Distance\_opt(S, D)$$

Inequality 1: Loop-Free Criterion

A subset of loop-free alternates are downstream paths that must meet a more restrictive condition that is applicable to more complex failure scenarios:

$$Distance\_opt(N, D) < Distance\_opt(S, D)$$

Inequality 2: Downstream Path Criterion

1.1. Failure Scenarios

The alternate next-hop can protect against a single link failure, a single node failure, failure of one or more links within a shared risk link group, or a combination of these. Whenever a failure occurs that is more extensive than what the alternate was intended to protect, there is the possibility of temporarily looping traffic (note again, that such a loop would only last until the next complete SPF calculation). The example where a node fails when the alternate provided only link protection is illustrated below. If unexpected simultaneous failures occur, then micro-looping may occur since the alternates are not pre-computed to avoid the set of failed links.

If only link protection is provided and the node fails, it is possible for traffic using the alternates to experience micro-looping. This issue is illustrated in Figure 2. If Link(S->E) fails, then the link-protecting alternate via N will work correctly. However, if router E fails, then both S and N will detect a failure and switch to their alternates. In this example, that would cause S to redirect the traffic to N and N to redirect the traffic to S and thus causing a forwarding loop. Such a scenario can arise because the key assumption, that all other routers in the network are forwarding based upon the shortest path, is violated because of a second simultaneous correlated failure -- another link connected to the same primary neighbor. If there are not other protection mechanisms to handle node failure, a node failure is still a concern when only using link-protecting LFAs.

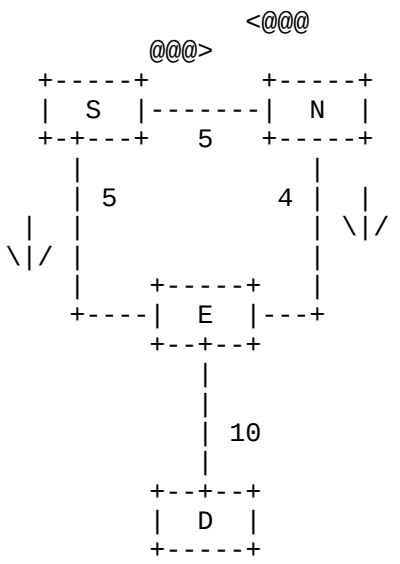

Figure 2: Link-Protecting Alternates Causing Loop on Node Failure

Micro-looping of traffic via the alternates caused when a more extensive failure than planned for occurs can be prevented via selection of only downstream paths as alternates. A micro-loop due to the use of alternates can be avoided by using downstream paths because each succeeding router in the path to the destination must be closer to the destination than its predecessor (according to the topology prior to the failures). Although use of downstream paths ensures that the micro-looping via alternates does not occur, such a restriction can severely limit the coverage of alternates. In Figure 2, S would be able to use N as a downstream alternate, but N could not use S; therefore, N would have no alternate and would discard the traffic, thus avoiding the micro-loop.

As shown above, the use of either a node-protecting LFA (described in Section 3.2) or a downstream path provides protection against micro-looping in the event of node failure. There are topologies where there may be either a node-protecting LFA, a downstream path, both, or neither. A node may select either a node-protecting LFA or a downstream path without risk of causing micro-loops in the event of neighbor node failure. While a link-and-node-protecting LFA guarantees protection against either link or node failure, a downstream path provides protection only against a link failure and may or may not provide protection against a node failure depending on the protection available at the downstream node, but it cannot cause a micro-loop. For example, in Figure 2, if S uses N as a downstream path, although no looping can occur, the traffic will not be

protected in the event of the failure of node E because N has no viable repair path, and it will simply discard the packet. However, if N had a link-and-node-protecting LFA or downstream path via some other path (not shown), then the repair may succeed.

Since the functionality of link-and-node-protecting LFAs is greater than that of link-protecting downstream paths, a router SHOULD select a link-and-node-protecting LFA over a link-protecting downstream path. If there are any destinations for which a link-and-node-protecting LFA is not available, then by definition the path to all of those destinations from any neighbor of the computing router (S) must be through the node (E) being protected (otherwise there would be a node protecting LFA for that destination). Consequently, if there exists a downstream path to the protected node as destination, then that downstream path may be used for all those destinations for which a link-and-node-protecting LFA is not available; the existence of a downstream path can be determined by a single check of the condition Distance_opt(N, E) < Distance_opt(S, E).

It may be desirable to find an alternate that can protect against other correlated failures (of which node failure is a specific instance). In the general case, these are handled by shared risk link groups (SRLGs) where any links in the network can belong to the SRLG. General SRLGs may add unacceptably to the computational complexity of finding a loop-free alternate.

However, a sub-category of SRLGs is of interest and can be applied only during the selection of an acceptable alternate. This sub-category is to express correlated failures of links that are connected to the same router, for example, if there are multiple logical sub-interfaces on the same physical interface, such as VLANs on an Ethernet interface, if multiple interfaces use the same physical port because of channelization, or if multiple interfaces share a correlated failure because they are on the same line-card. This sub-category of SRLGs will be referred to as local-SRLGs. A local-SRLG has all of its member links with one end connected to the same router. Thus, router S could select a loop-free alternate that does not use a link in the same local-SRLG as the primary next-hop. The failure of local-SRLGs belonging to E can be protected against via node protection, i.e., picking a loop-free node-protecting alternate.

Where SRLG protection is provided, it is in the context of the particular OSPF or IS-IS area, whose topology is used in the SPF computations to compute the loop-free alternates. If an SRLG contains links in multiple areas, then separate SRLG-protecting alternates would be required in each area that is traversed by the affected traffic.

1.2. Requirement 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].

2. Applicability of Described Mechanisms

IP Fast Reroute mechanisms described in this memo cover intra-domain routing only, with OSPF [RFC2328] [RFC2740] [RFC5340] or IS-IS [RFC1195] [RFC2966] as the IGP. Specifically, Fast Reroute for BGP inter-domain routing is not part of this specification.

Certain aspects of OSPF inter-area routing behavior explained in Section 6.3 and Appendix A impact the ability of the router calculating the backup next-hops to assess traffic trajectories. In order to avoid micro-looping and ensure required coverage, certain constraints are applied to multi-area OSPF networks:

a. Loop-free alternates should not be used in the backbone area if there are any virtual links configured unless, for each transit area, there is a full mesh of virtual links between all Area Border Routers (ABRs) in that area. Loop-free alternates may be used in non-backbone areas regardless of whether there are virtual links configured.

b. Loop-free alternates should not be used for inter-area routes in an area that contains more than one alternate ABR [RFC3509].

c. Loop-free alternates should not be used for AS External routes or Autonomous System Border Router (ASBR) routes in a non-backbone area of a network where there exists an ABR that is announced as an ASBR in multiple non-backbone areas and there exists another ABR that is in at least two of the same non-backbone areas.

d. Loop-free alternates should not be used in a non-backbone area of a network for AS External routes where an AS External prefix is advertised with the same type of external metric by multiple ASBRs, which are in different non-backbone areas, with a forwarding address of 0.0.0.0 or by one or more ASBRs with forwarding addresses in multiple non-backbone areas when an ABR exists simultaneously in two or more of those non-backbone areas.

3. Alternate Next-Hop Calculation

In addition to the set of primary next-hops obtained through a shortest path tree (SPT) computation that is part of standard link-state routing functionality, routers supporting IP Fast Reroute also calculate a set of backup next-hops that are engaged when a local failure occurs. These backup next-hops are calculated to provide the required type of protection (i.e., link-protecting and/or node-protecting) and to guarantee that when the expected failure occurs, forwarding traffic through them will not result in a loop. Such next-hops are called loop-free alternates or LFAs throughout this specification.

In general, to be able to calculate the set of LFAs for a specific destination D, a router needs to know the following basic pieces of information:

- Shortest-path distance from the calculating router to the destination (Distance_opt(S, D))
- Shortest-path distance from the router's IGP neighbors to the destination (Distance_opt(N, D))
- Shortest path distance from the router's IGP neighbors to itself (Distance_opt(N, S))
- Distance_opt(S, D) is normally available from the regular SPF calculation performed by the link-state routing protocols. Distance_opt(N, D) and Distance_opt(N, S) can be obtained by performing additional SPF calculations from the perspective of each IGP neighbor (i.e., considering the neighbor's vertex as the root of the SPT--called SPT(N) hereafter--rather than the calculating router's one, called SPT(S)).

This specification defines a form of SRLG protection limited to those SRLGs that include a link to which the calculating router is directly connected. Only that set of SRLGs could cause a local failure; the calculating router only computes alternates to handle a local failure. Information about local link SRLG membership is manually configured. Information about remote link SRLG membership may be dynamically obtained using [RFC4205] or [RFC4203]. Define SRLG_local(S) to be the set of SRLGs that include a link to which the calculating router S is directly connected Only SRLG_local(S) is of interest during the calculation, but the calculating router must correctly handle changes to SRLG_local(S) triggered by local link SRLG membership changes.

In order to choose among all available LFAs that provide required SRLG protection for a given destination, the calculating router needs to track the set of SRLGs in SRLG_local(S) that the path through a specific IGP neighbor involves. To do so, each node D in the network topology is associated with SRLG_set(N, D), which is the set of SRLGs that would be crossed if traffic to D was forwarded through N. To calculate this set, the router initializes SRLG_set(N, N) for each of its IGP neighbors to be empty. During the SPT(N) calculation, when a new vertex V is added to the SPT, its SRLG_set(N, V) is set to the union of SRLG sets associated with its parents, and the SRLG sets in SRLG_local(S) that are associated with the links from V's parents to V. The union of the set of SRLGs associated with a candidate alternate next-hop and the SRLG_set(N, D) for the neighbor reached via that candidate next-hop is used to determine SRLG protection.

The following sections provide information required for calculation of LFAs. Sections 3.1 through 3.4 define different types of LFA conditions. Section 3.5 describes constraints imposed by the IS-IS overload and OSPF stub router functionality. Section 3.6 defines the summarized algorithm for LFA calculation using the definitions in the previous sections.

3.1. Basic Loop-Free Condition

Alternate next hops used by implementations following this specification MUST conform to at least the loop-freeness condition stated above in Inequality 1. This condition guarantees that forwarding traffic to an LFA will not result in a loop after a link failure.

Further conditions may be applied when determining link-protecting and/or node-protecting alternate next-hops as described in Sections 3.2 and 3.3.

3.2. Node-Protecting Alternate Next-Hops

For an alternate next-hop N to protect against node failure of a primary neighbor E for destination D, N must be loop-free with respect to both E and D. In other words, N's path to D must not go through E. This is the case if Inequality 3 is true, where N is the neighbor providing a loop-free alternate.

$$Distance\_opt(N, D) < Distance\_opt(N, E) + Distance\_opt(E, D)$$

Inequality 3: Criteria for a Node-Protecting Loop-Free Alternate

If Distance_opt(N,D) = Distance_opt(N, E) + Distance_opt(E, D), it is possible that N has equal-cost paths and one of those could provide protection against E's node failure. However, it is equally possible that one of N's paths goes through E, and the calculating router has no way to influence N's decision to use it. Therefore, it SHOULD be assumed that an alternate next-hop does not offer node protection if Inequality 3 is not met.

3.3. Broadcast and Non-Broadcast Multi-Access (NBMA) Links

Verification of the link-protection property of a next-hop in the case of a broadcast link is more elaborate than for a point-to-point link. This is because a broadcast link is represented as a pseudo-node with zero-cost links connecting it to other nodes.

Because failure of an interface attached to a broadcast segment may mean loss of connectivity of the whole segment, the condition described for broadcast link protection is pessimistic and requires that the alternate is loop-free with regard to the pseudo-node. Consider the example in Figure 3.

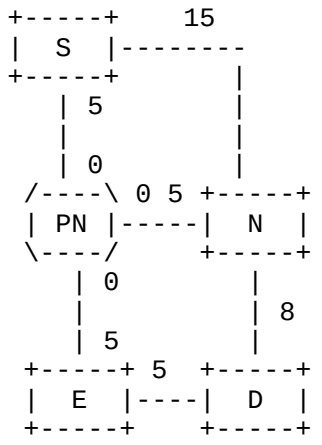

Figure 3: Loop-Free Alternate That Is Link-Protecting

In Figure 3, N offers a loop-free alternate that is link-protecting. If the primary next-hop uses a broadcast link, then an alternate SHOULD be loop-free with respect to that link's pseudo-node (PN) to provide link protection. This requirement is described in Inequality 4 below.

$$D\_opt(N, D) < D\_opt(N, PN) + D\_opt(PN, D)$$

Inequality 4: Loop-Free Link-Protecting Criterion for Broadcast Links

Because the shortest path from the pseudo-node goes through E, if a loop-free alternate from a neighbor N is node-protecting, the alternate will also be link-protecting unless the router S can only reach the alternate neighbor N via the same pseudo-node. Since this is the only case for which a node-protecting LFA is not link-protecting, this implies that for point-to-point interfaces, an LFA that is node-protecting is always link-protecting. Because S can direct the traffic away from the shortest path to use the alternate N, traffic might pass through the same broadcast link as it would when S sent the traffic to the primary E. Thus, an LFA from N that is node-protecting is not automatically link-protecting for a broadcast or NBMA link.

To obtain link protection, it is necessary both that the path from the selected alternate next-hop does not traverse the link of interest and that the link used from S to reach that alternate next-hop is not the link of interest. The latter can only occur with non-point-to-point links. Therefore, if the primary next-hop is across a broadcast or NBMA interface, it is necessary to consider link protection during the alternate selection. To clarify, consider the topology in Figure 3. For N to provide link protection, it is first necessary that N's shortest path to D does not traverse the pseudo-node PN. Second, it is necessary that the alternate next-hop selected by S does not traverse PN. In this example, S's shortest path to N is via the pseudo-node. Thus, to obtain link protection, S must find a next-hop to N (the point-to-point link from S to N in this example) that avoids the pseudo-node PN.

Similar consideration of the link from S to the selected alternate next-hop as well as the path from the selected alternate next-hop is also necessary for SRLG protection. S's shortest path to the selected neighbor N may not be acceptable as an alternate next-hop to provide SRLG protection, even if the path from N to D can provide SRLG protection.

3.4. ECMP and Alternates

With Equal-Cost Multi-Path (ECMP), a prefix may have multiple primary next-hops that are used to forward traffic. When a particular primary next-hop fails, alternate next-hops should be used to preserve the traffic. These alternate next-hops may themselves also be primary next-hops, but need not be. Other primary next-hops are not guaranteed to provide protection against the failure scenarios of concern.

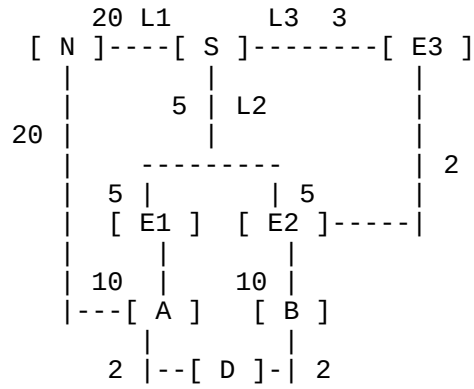

Figure 4: ECMP Where Primary Next-Hops Provide Limited Protection

In Figure 4 S has three primary next-hops to reach D; these are L2 to E1, L2 to E2, and L3 to E3. The primary next-hop L2 to E1 can obtain link and node protection from L3 to E3, which is one of the other primary next-hops; L2 to E1 cannot obtain link protection from the other primary next-hop L2 to E2. Similarly, the primary next-hop L2 to E2 can only get node protection from L2 to E1 and can only get link protection from L3 to E3. The third primary next-hop L3 to E3 can obtain link and node protection from L2 to E1 and from L2 to E2. It is possible for both the primary next-hop L2 to E2 and the primary next-hop L2 to E1 to obtain an alternate next-hop that provides both link and node protection by using L1.

Alternate next-hops are determined for each primary next-hop separately. As with alternate selection in the non-ECMP case, these alternate next-hops should maximize the coverage of the failure cases.

3.5. Interactions with IS-IS Overload, RFC 3137, and Costed Out Links

As described in [RFC3137], there are cases where it is desirable not to have a router used as a transit node. For those cases, it is also desirable not to have the router used on an alternate path.

For computing an alternate, a router MUST NOT use an alternate next-hop that is along a link whose cost or reverse cost is LSInfinity (for OSPF) or the maximum cost (for IS-IS) or that has the overload bit set (for IS-IS). For a broadcast link, the reverse cost associated with a potential alternate next-hop is the cost towards the pseudo-node advertised by the next-hop router. For point-to-point links, if a specific link from the next-hop router cannot be associated with a particular link, then the reverse cost considered is that of the minimum cost link from the next-hop router back to S.

In the case of OSPF, if all links from router S to a neighbor N_i have a reverse cost of LSInfinity, then router S MUST NOT use N_i as an alternate.

Similarly in the case of IS-IS, if N_i has the overload bit set, then S MUST NOT consider using N_i as an alternate.

This preserves the desired behavior of diverting traffic away from a router that is following [RFC3137], and it also preserves the desired behavior when an operator sets the cost of a link to LSInfinity for maintenance that is not permitting traffic across that link unless there is no other path.

If a link or router that is costed out was the only possible alternate to protect traffic from a particular router S to a particular destination, then there should be no alternate provided for protection.

3.5.1. Interactions with IS-IS Link Attributes

[RFC5029] describes several flags whose interactions with LFAs need to be defined. A router SHOULD NOT specify the "local protection available" flag as a result of having LFAs. A router SHOULD NOT use an alternate next-hop that is along a link for which the link has been advertised with the attribute "link excluded from local protection path" or with the attribute "local maintenance required".

3.6. Selection Procedure

A router supporting this specification SHOULD attempt to select at least one loop-free alternate next-hop for each primary next-hop used for a given prefix. A router MAY decide to not use an available loop-free alternate next-hop. A reason for such a decision might be that the loop-free alternate next-hop does not provide protection for the failure scenario of interest.

The alternate selection should maximize the coverage of the failure cases.

When calculating alternate next-hops, the calculating router S applies the following rules.

1. S SHOULD select a loop-free node-protecting alternate next-hop, if one is available. If no loop-free node-protecting alternate is available, then S MAY select a loop-free link-protecting alternate.

2. If S has a choice between a loop-free link-and-node-protecting alternate and a loop-free node-protecting alternate that is not link-protecting, S SHOULD select a loop-free link-and-node-protecting alternate. This can occur as explained in Section 3.3.

3. If S has multiple primary next-hops, then S SHOULD select as a loop-free alternate either one of the other primary next-hops or a loop-free node-protecting alternate if available. If no loop-free node-protecting alternate is available and no other primary next-hop can provide link-protection, then S SHOULD select a loop-free link-protecting alternate.

4. Implementations SHOULD support a mode where other primary next-hops satisfying the basic loop-free condition and providing at least link or node protection are preferred over any non-primary alternates. This mode is provided to allow the administrator to preserve traffic patterns based on regular ECMP behavior.

5. Implementations considering SRLGs MAY use SRLG protection to determine that a node-protecting or link-protecting alternate is not available for use.

Following the above rules maximizes the level of protection and use of primary (ECMP) next-hops.

Each next-hop is associated with a set of non-mutually-exclusive characteristics based on whether it is used as a primary next-hop to a particular destination D, and the type of protection it can provide relative to a specific primary next-hop E:

Primary Path - The next-hop is used by S as primary.

Loop-Free Node-Protecting Alternate - This next-hop satisfies Inequality 1 and Inequality 3. The path avoids S, S’s primary neighbor E, and the link from S to E.

Loop-Free Link-Protecting Alternate - This next-hop satisfies Inequality 1 but not Inequality 3. If the primary next-hop uses a broadcast link, then this next-hop satisfies Inequality 4.

An alternate path may also provide none, some, or complete SRLG protection as well as node and link or link protection. For instance, a link may belong to two SRLGs G1 and G2. The alternate path might avoid other links in G1 but not G2, in which case the alternate would only provide partial SRLG protection.

Below is an algorithm that can be used to calculate loop-free alternate next-hops. The algorithm is given for informational purposes, and implementations are free to use any other algorithm as long as it satisfies the rules described above.

The following procedure describes how to select an alternate next-hop. The procedure is described to determine alternate next-hops to use to reach each router in the topology. Prefixes that are advertised by a single router can use the alternate next-hop computed for the router to which they are attached. The same procedure can be used to reach a prefix that is advertised by more than one router when the logical topological transformation described in Section 6.1 is used.

S is the computing router. S has neighbors N_1 to N_j. A candidate next-hop is indicated by (outgoing link, neighbor) and the outgoing link must be bidirectionally connected, as is determined by the IGP. The candidate next-hops of S are enumerated as H_1 through H_k. Recall that S may have multiple next-hops over different interfaces to a neighbor. H_i.link refers to the outgoing link of that next-hop and H_i.neighbor refers to the neighbor of that next-hop.

For a particular destination router D, let S have already computed D_opt(S, D), and for each neighbor N_i, D_opt(N_i, D), D_opt(N_i, S), and D_opt(N_i, N_j), the distance from N_i to each other neighbor N_j, and the set of SRLGs traversed by the path D_opt(N_i, D). S should follow the below procedure for every primary next-hop selected to reach D. This set of primary next-hops is represented P_1 to P_p. This procedure finds the alternate next-hop(s) for P_i.

First, initialize the alternate information for P_i as follows:

```
P_i.alt_next_hops = {}
P_i.alt_type = NONE
P_i.alt_link-protect = FALSE
P_i.alt_node-protect = FALSE
P_i.alt_srlg-protect = {}
```

For each candidate next-hop H_h,

1. Initialize variables as follows:

```
cand_type = NONE
cand_link-protect = FALSE
cand_node-protect = FALSE
cand_srlg-protect = {}
```

2. If H_h is P_i, skip it and continue to the next candidate next-hop.

3. If H_h.link is administratively allowed to be used as an alternate,

   and the cost of H_h.link is less than the maximum,
   and the reverse cost of H_h is less than the maximum,
   and H_h.neighbor is not overloaded (for IS-IS),
   and H_h.link is bidirectional,

   then H_h can be considered as an alternate. Otherwise, skip it and continue to the next candidate next-hop.

4. If D_opt(H_h.neighbor, D) >= D_opt(H_h.neighbor, S) + D_opt(S, D), then H_h is not loop-free. Skip it and continue to the next candidate next-hop.

5. cand_type = LOOP-FREE.

6. If H_h is a primary next-hop, set cand_type to PRIMARY.

7. If H_h.link is not P_i.link, set cand_link-protect to TRUE.

8. If D_opt(H_h.neighbor, D) < D_opt(H_h.neighbor, P_i.neighbor) + D_opt(P_i.neighbor, D), set cand_node-protect to TRUE.

9. If the router considers SRLGs, then set the cand_srlg-protect to the set of SRLGs traversed on the path from S via P_i.link to P_i.neighbor. Remove the set of SRLGs to which H_h belongs from cand_srlg-protect. Remove from cand_srlg-protect the set of SRLGs traversed on the path from H_h.neighbor to D. Now cand_srlg-protect holds the set of SRLGs to which P_i belongs and that are not traversed on the path from S via H_h to D.

10. If cand_type is PRIMARY, the router prefers other primary next-hops for use as the alternate, and the P_i.alt_type is not PRIMARY, goto Step 20.

11. If cand_type is not PRIMARY, P_i.alt_type is PRIMARY, and the router prefers other primary next-hops for use as the alternate, then continue to the next candidate next-hop

12. If cand_node-protect is TRUE and P_i.alt_node-protect is FALSE, goto Paragraph 20.

13. If cand_link-protect is TRUE and P_i.alt_link-protect is FALSE, goto Step 20.

14. If cand_srlg-protect has a better set of SRLGs than P_i.alt_srlg-protect, goto Step 20.

15. If cand_srlg-protect is different from P_i.alt_srlg-protect, then select between H_h and P_i.alt_next_hops based upon distance, IP addresses, or any router-local tie-breaker. If H_h is preferred, then goto Step 20. If P_i.alt_next_hops is preferred, skip H_h and continue to the next candidate next-hop.

16. If D_opt(H_h.neighbor, D) < D_opt(P_i.neighbor, D) and D_opt(P_i.alt_next_hops, D) >= D_opt(P_i.neighbor, D), then H_h is a downstream alternate and P_i.alt_next_hops is simply an LFA. Prefer H_h and goto Step 20.

17. Based upon the alternate types, the alternate distances, IP addresses, or other tie-breakers, decide if H_h is preferred to P_i.alt_next_hops. If so, goto Step 20.

18. Decide if P_i.alt_next_hops is preferred to H_h. If so, then skip H_h and continue to the next candidate next-hop.

19. Add H_h into P_i.alt_next_hops. Set P_i.alt_type to the better type of H_h.alt_type and P_i.alt_type. Continue to the next candidate next-hop.

20. Replace the P_i alternate next-hop set with H_h as follows:

```
P_i.alt_next_hops = {H_h}
P_i.alt_type = cand_type
P_i.alt_link-protect = cand_link-protect
P_i.alt_node-protect = cand_node-protect
P_i.alt_srlg-protect = cand_srlg-protect
```

    Continue to the next candidate next-hop.

3.7. LFA Types and Trade-Offs

LFAs can provide different amounts of protection, and the decision about which type to prefer is dependent upon network topology and other techniques in use in the network. This section describes the different protection levels and the trade-offs associated with each.

1. Primary Next-hop: When there are equal-cost primary next-hops, using one as an alternate is guaranteed not to cause micro-loops involving S. Traffic flows across the paths that the network will converge to, but congestion may be experienced on the primary paths since traffic is sent across fewer. All primary next-hops are downstream paths.

2. Downstream Paths: A downstream path, unlike an LFA, is guaranteed not to cause a micro-loop involving S regardless of the actual failure detected. However, the expected coverage of such alternates in a network is expected to be poor. All downstream paths are LFAs.

3. LFA: An LFA can have good coverage of a network, depending on topology. However, it is possible to get micro-loops involving S if an unprotected failure occurs (e.g., a node fails when the LFA only was link-protecting).

The different types of protection are abbreviated as LP (link-protecting), NP (node-protecting), and SP (SRLG-protecting).

a. LP, NP, and SP: If such an alternate exists, it gives protection against all failures.

b. LP and NP only: Many networks may handle SRLG failures via another method or may focus on node and link failures as being more common.

c. LP only: A network may handle node failures via a high-availability technique and be concerned primarily about protecting the more common link failure case.

d. NP only: These only exist on interfaces that aren't point-to-point. If link protection is handled in a different layer, then an NP alternate may be acceptable.

3.8. A Simplification: Per-Next-Hop LFAs

It is possible to simplify the computation and use of LFAs when solely link protection is desired by considering and computing only one link-protecting LFA for each next-hop connected to the router. All prefixes that use that next-hop as a primary will use the LFA computed for that next-hop as its LFA.

Even a prefix with multiple primary next-hops will have each primary next-hop protected individually by the primary next-hop's associated LFA. That associated LFA might or might not be another of the primary next-hops of the prefix.

This simplification may reduce coverage in a network. In addition to limiting protection for multi-homed prefixes (see Section 6.1), the computation per next-hop may also not find an LFA when one could be found for some of the prefixes that use that next-hop.

For example, consider Figure 4 where S has three ECMP next-hops, E1, E2, and E3 to reach D. For the prefix D, E3 can give link protection for the next-hops E1 and E2; E1 and E2 can give link protection for the next-hops E3. However, if one uses this simplification to compute LFAs for E1, E2, and E3 individually, there is no link-protecting LFA for E1. E3 and E2 can protect each other.

4. Using an Alternate

If an alternate next-hop is available, the router redirects traffic to the alternate next-hop in case of a primary next-hop failure as follows.

When a next-hop failure is detected via a local interface failure or other failure detection mechanisms (see [FRAMEWORK]), the router SHOULD:

1. Remove the primary next-hop associated with the failure.

2. Install the loop-free alternate calculated for the failed next-hop if it is not already installed (e.g., the alternate is also a primary next-hop).

Note that the router MAY remove other next-hops if it believes (via SRLG analysis) that they may have been affected by the same failure, even if it is not visible at the time of failure detection.

The alternate next-hop MUST be used only for traffic types that are routed according to the shortest path. Multicast traffic is specifically out of scope for this specification.

4.1. Terminating Use of Alternate

A router MUST limit the amount of time an alternate next-hop is used after the primary next-hop has become unavailable. This ensures that the router will start using the new primary next-hops. It ensures that all possible transient conditions are removed and the network converges according to the deployed routing protocol.

There are techniques available to handle the micro-forwarding loops that can occur in a networking during convergence.

A router that implements [MICROLOOP] SHOULD follow the rules given there for terminating the use of an alternate.

A router that implements [ORDERED-FIB] SHOULD follow the rules given there for terminating the use of an alternate.

It is desirable to avoid micro-forwarding loops involving S. An example illustrating the problem is given in Figure 5. If the link from S to E fails, S will use N1 as an alternate and S will compute N2 as the new primary next-hop to reach D. If S starts using N2 as soon as S can compute and install its new primary, it is probable that N2 will not have yet installed its new primary next-hop. This would cause traffic to loop and be dropped until N2 has installed the new topology. This can be avoided by S delaying its installation and leaving traffic on the alternate next-hop.

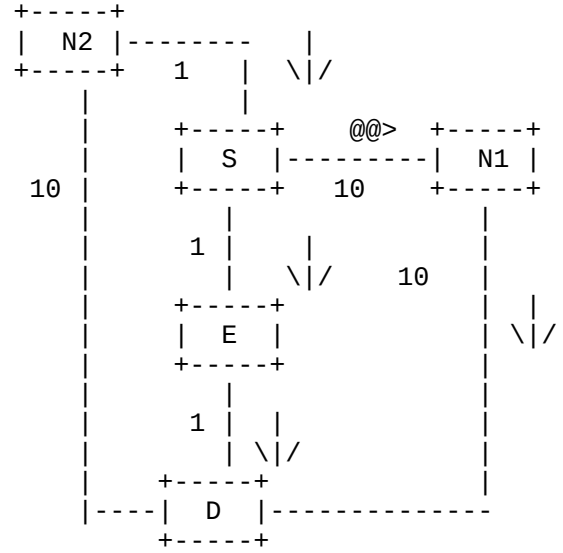

Figure 5: Example Where Continued Use of Alternate Is Desirable

This is an example of a case where the new primary is not a loop-free alternate before the failure and therefore may have been forwarding traffic through S. This will occur when the path via a previously upstream node is shorter than the path via a loop-free alternate neighbor. In these cases, it is useful to give sufficient time to ensure that the new primary neighbor and other nodes on the new primary path have switched to the new route.

If the newly selected primary was loop-free before the failure, then it is safe to switch to that new primary immediately; the new primary wasn't dependent on the failure and therefore its path will not have changed.

Given that there is an alternate providing appropriate protection and while the assumption of a single failure holds, it is safe to delay the installation of the new primaries; this will not create

forwarding loops because the alternate's path to the destination is known to not go via S or the failed element and will therefore not be affected by the failure.

An implementation SHOULD continue to use the alternate next-hops for packet forwarding even after the new routing information is available based on the new network topology. The use of the alternate next-hops for packet forwarding SHOULD terminate:

a. if the new primary next-hop was loop-free prior to the topology change, or

b. if a configured hold-down, which represents a worst-case bound on the length of the network convergence transition, has expired, or

c. if notification of an unrelated topological change in the network is received.

5. Requirements on LDP Mode

Since LDP [RFC5036] traffic will follow the path specified by the IGP, it is also possible for the LDP traffic to follow the loop-free alternates indicated by the IGP. To do so, it is necessary for LDP to have the appropriate labels available for the alternate so that the appropriate out-segments can be installed in the forwarding plane before the failure occurs.

This means that a Label Switching Router (LSR) running LDP must distribute its labels for the Forwarding Equivalence Classes (FECs) it can provide to all its neighbors, regardless of whether or not they are upstream. Additionally, LDP must be acting in liberal label retention mode so that the labels that correspond to neighbors that aren't currently the primary neighbor are stored. Similarly, LDP should be in downstream unsolicited mode, so that the labels for the FEC are distributed other than along the SPT.

If these requirements are met, then LDP can use the loop-free alternates without requiring any targeted sessions or signaling extensions for this purpose.

6. Routing Aspects

6.1. Multi-Homed Prefixes

An SPF-like computation is run for each topology, which corresponds to a particular OSPF area or IS-IS level. The IGP needs to determine loop-free alternates to multi-homed routes. Multi-homed routes occur for routes obtained from outside the routing domain by multiple

routers, for subnets on links where the subnet is announced from multiple ends of the link, and for routes advertised by multiple routers to provide resiliency.

Figure 6 demonstrates such a topology. In this example, the shortest path to reach the prefix p is via E. The prefix p will have the link to E as its primary next-hop. If the alternate next-hop for the prefix p is simply inherited from the router advertising it on the shortest path to p, then the prefix p's alternate next-hop would be the link to C. This would provide link protection, but not the node protection that is possible via A.

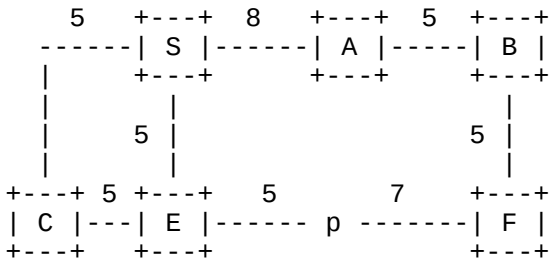

Figure 6: Multi-Homed Prefix

To determine the best protection possible, the prefix p can be treated in the SPF computations as a node with unidirectional links to it from those routers that have advertised the prefix. Such a node need never have its links explored, as it has no out-going links.

If there exist multiple multi-homed prefixes that share the same connectivity and the difference in metrics to those routers, then a single node can be used to represent the set. For instance, if in Figure 6 there were another prefix X that was connected to E with a metric of 1 and to F with a metric of 3, then that prefix X could use the same alternate next-hop as was computed for prefix p.

A router SHOULD compute the alternate next-hop for an IGP multi-homed prefix by considering alternate paths via all routers that have announced that prefix.

In all cases, a router MAY safely simplify the multi-homed prefix (MHP) calculation by assuming that the MHP is solely attached to the router that was its pre-failure optimal point of attachment. However, this may result in a prefix not being considered repairable, when the full computation would show that a repair was possible.

6.2. IS-IS

The applicability and interactions of LFAs with multi-topology IS-IS [RFC5120] is out of scope for this specification.

6.3. OSPF

OSPF introduces certain complications because it is possible for the traffic path to exit an area and then re-enter that area. This can occur whenever a router considers the same route from multiple areas. There are several cases where issues such as this can occur. They happen when another area permits a shorter path to connect two ABRs than is available in the area where the LFA has been computed. To clarify, an example topology is given in Appendix A.

a. Virtual Links: These allow paths to leave the backbone area and traverse the transit area. The path provided via the transit area can exit via any ABR. The path taken is not the shortest path determined by doing an SPF in the backbone area.

b. Alternate ABR [RFC3509]: When an ABR is not connected to the backbone, it considers the inter-area summaries from multiple areas. The ABR A may determine to use area 2 but that path could traverse another alternate ABR B that determines to use area 1. This can lead to scenarios similar to that illustrated in Figure 7.

c. ASBR Summaries: An ASBR may itself be an ABR and can be announced into multiple areas. This presents other ABRs with a decision as to which area to use. This is the example illustrated in Figure 7.

d. AS External Prefixes: A prefix may be advertised by multiple ASBRs in different areas and/or with multiple forwarding addresses that are in different areas, which are connected via at least one common ABR. This presents such ABRs with a decision as to which area to use to reach the prefix.

Loop-free alternates should not be used in an area where one of the above issues affects that area.

6.3.1. OSPF External Routing

When a forwarding address is set in an OSPF AS-external Link State Advertisement (LSA), all routers in the network calculate their next-hops for the external prefix by doing a lookup for the forwarding address in the routing table, rather than using the next-hops

calculated for the ASBR. In this case, the alternate next-hops SHOULD be computed by selecting among the alternate paths to the forwarding link(s) instead of among alternate paths to the ASBR.

6.3.2. OSPF Multi-Topology

The applicability and interactions of LFAs with multi-topology OSPF [RFC4915] [MT-OSPFv3] is out of scope for this specification.

6.4. BGP Next-Hop Synchronization

Typically, BGP prefixes are advertised with the AS exit router's router-id as the BGP next-hop, and AS exit routers are reached by means of IGP routes. BGP resolves its advertised next-hop to the immediate next-hop by potential recursive lookups in the routing database. IP Fast Reroute computes the alternate next-hops to all IGP destinations, which include alternate next-hops to the AS exit router's router-id. BGP simply inherits the alternate next-hop from IGP. The BGP decision process is unaltered; BGP continues to use the IGP optimal distance to find the nearest exit router. Multicast BGP (MBGP) routes do not need to copy the alternate next-hops.

It is possible to provide ASBR protection if BGP selected a set of BGP next-hops and allowed the IGP to determine the primary and alternate next-hops as if the BGP route were a multi-homed prefix. This is for future study.

6.5. Multicast Considerations

Multicast traffic is out of scope for this specification of IP Fast Reroute. The alternate next-hops SHOULD NOT be used for multicast Reverse Path Forwarding (RPF) checks.

7. Security Considerations

The mechanism described in this document does not modify any routing protocol messages, and hence no new threats related to packet modifications or replay attacks are introduced. Traffic to certain destinations can be temporarily routed via next-hop routers that would not be used with the same topology change if this mechanism wasn't employed. However, these next-hop routers can be used anyway when a different topological change occurs, and hence this can't be viewed as a new security threat.

In LDP, the wider distribution of FEC label information is still to neighbors with whom a trusted LDP session has been established. This wider distribution and the recommendation of using liberal label retention mode are believed to have no significant security impact.

8. Acknowledgements

The authors would like to thank Joel Halpern, Mike Shand, Stewart Bryant, and Stefano Previdi for their assistance and useful review.

9. References

9.1. Normative References

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

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC2740] Coltun, R., Ferguson, D., and J. Moy, "OSPF for IPv6", RFC 2740, December 1999.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

9.2. Informative References

[FRAMEWORK] Shand, M. and S. Bryant, "IP Fast Reroute Framework", Work in Progress, February 2008.

[MICROLOOP] Zinin, A., "Analysis and Minimization of Microloops in Link-state Routing Protocols", Work in Progress, October 2005.

[MT-OSPFv3] Mirtorabi, S. and A. Roy, "Multi-topology routing in OSPFv3 (MT-OSPFv3)", Work in Progress, July 2007.

[ORDERED-FIB] Francois, P., "Loop-free convergence using oFIB", Work in Progress, February 2008.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[RFC2966] Li, T., Przygienda, T., and H. Smit, "Domain-wide Prefix Distribution with Two-Level IS-IS", RFC 2966, October 2000.

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

[RFC3509] Zinin, A., Lindem, A., and D. Yeung, "Alternative Implementations of OSPF Area Border Routers", RFC 3509, April 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.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.

[RFC5029] Vasseur, JP. and S. Previdi, "Definition of an IS-IS Link Attribute Sub-TLV", RFC 5029, September 2007.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.

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

Appendix A. OSPF Example Where LFA Based on Local Area Topology Is Insufficient

This appendix provides an example scenario where the local area topology does not suffice to determine that an LFA is available. As described in Section 6.3, one problem scenario is for ASBR summaries where the ASBR is available in two areas via intra-area routes and there is at least one ABR or alternate ABR that is in both areas. The following Figure 7 illustrates this case.

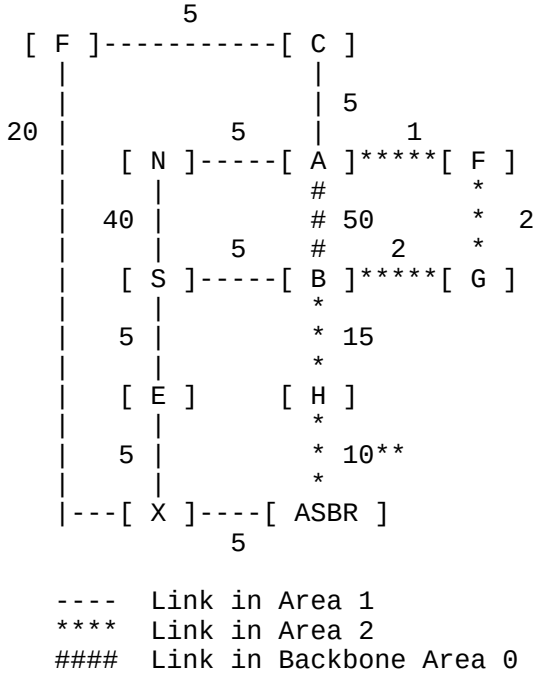

Figure 7: Topology with Multi-Area ASBR Causing Area Transiting

In Figure 7, the ASBR is also an ABR and is announced into both area 1 and area 2. A and B are both ABRs that are also connected to the backbone area. S determines that N can provide a loop-free alternate to reach the ASBR. N's path goes via A. A also sees an intra-area route to ASBR via area 2; the cost of the path in area 2 is 30, which is less than 35, the cost of the path in area 1. Therefore, A uses the path from area 2 and directs traffic to F. The path from F in area 2 goes to B. B is also an ABR and learns the ASBR from both areas 1 and area 2; B's path via area 1 is shorter (cost 20) than B's path via area 2 (cost 25). Therefore, B uses the path from area 1 that connects to S.

Authors' Addresses

Alia K. Atlas (editor)
BT

EMail: alia.atlas@bt.com

Alex Zinin (editor)
Alcatel-Lucent
750D Chai Chee Rd, #06-06
Technopark@ChaiChee
Singapore 469004

EMail: alex.zinin@alcatel-lucent.com

Raveendra Torvi
FutureWei Technologies Inc.
1700 Alma Dr. Suite 100
Plano, TX 75075
USA

EMail: traveendra@huawei.com

Gagan Choudhury
AT&T
200 Laurel Avenue, Room D5-3C21
Middletown, NJ 07748
USA

Phone: +1 732 420-3721
EMail: gchoudhury@att.com

Christian Martin
iPath Technologies

EMail: chris@ipath.net

Brent Imhoff
Juniper Networks
1194 North Mathilda
Sunnyvale, CA 94089
USA

Phone: +1 314 378 2571
EMail: bimhoff@planetspork.com

Don Fedyk
Nortel Networks
600 Technology Park
Billerica, MA 01821
USA

Phone: +1 978 288 3041
EMail: dwfedyk@nortelnetworks.com

Full Copyright Statement

Intellectual Property