Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 24, 2013

S. Bryant
C. Filsfils
S. Previdi
Cisco Systems
M. Shand
Independent Contributor
N. So
Tata Communications

May 23, 2013

Remote LFA FRR
draft-ietf-rtgwg-remote-lfa-02

Abstract

This draft describes an extension to the basic IP fast re-route mechanism described in RFC5286 that provides additional backup connectivity for link failures when none can be provided by the basic mechanisms.

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 [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 November 24, 2013.

Copyright Notice

1. Terminology

This draft uses the terms defined in [RFC5714]. This section defines additional terms used in this draft.

Extended P-space

: The union of the P-space of the neighbours of a specific router with respect to the protected link.

P-space

: P-space is the set of routers reachable from a specific router without any path (including equal cost path splits) transiting the protected link.

  For example, the P-space of S, is the set of routers that S can reach without using the protected link S-E.

PQ node

: A node which is a member of both the extended P-space and the Q-space.

Q-space

: Q-space is the set of routers from which a specific router can be reached without any path (including equal cost path splits) transiting the protected link.

Repair tunnel

: A tunnel established for the purpose of providing a virtual neighbor which is a Loop Free Alternate.

Remote LFA

: The tail-end of a repair tunnel. This tail-end is a member of both the extended-P space the Q space. It is also termed a "PQ" node.

In this document we use the notation X-Y to mean the path from X to Y over the link directly connecting X and Y, whilst the notation X->Y refers to the shortest path from X to Y via some set of unspecified nodes including the null set (i.e. including over a link directly connecting X and Y).

2. Introduction

RFC 5714 [RFC5714] describes a framework for IP Fast Re-route and provides a summary of various proposed IPFRR solutions. A basic mechanism using loop-free alternates (LFAs) is described in [RFC5286] that provides good repair coverage in many topologies[I-D.filsfils-rtgwg-lfa-applicability], especially those that are highly meshed. However, some topologies, notably ring based topologies are not well protected by LFAs alone. This is illustrated in Figure 1 below.

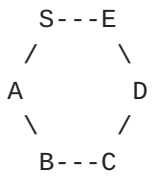

Figure 1: A simple ring topology

If all link costs are equal, the link S-E cannot be fully protected by LFAs. The destination C is an ECMP from S, and so can be protected when S-E fails, but D and E are not protectable using LFAs

This draft describes extensions to the basic repair mechanism in which tunnels are used to provide additional logical links which can then be used as loop free alternates where none exist in the original topology. For example if a tunnel is provided between S and C as shown in Figure 2 then C, now being a direct neighbor of S would become an LFA for D and E. The non-failure traffic distribution is not disrupted by the provision of such a tunnel since it is only used for repair traffic and MUST NOT be used for normal traffic.

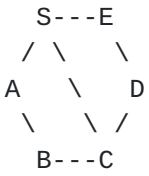

Figure 2: The addition of a tunnel

The use of this technique is not restricted to ring based topologies, but is a general mechanism which can be used to enhance the protection provided by LFAs.

This technique describes in this document is directed at providing repairs in the case of link failures. Considerations regarding node failures are discussed in Section 6.

3. Repair Paths

As with LFA FRR, when a router detects an adjacent link failure, it uses one or more repair paths in place of the failed link. Repair paths are pre-computed in anticipation of later failures so they can be promptly activated when a failure is detected.

A tunneled repair path tunnels traffic to some staging point in the network from which it is assumed that, in the absence of multiple failures, it will travel to its destination using normal forwarding without looping back. This is equivalent to providing a virtual loop-free alternate to supplement the physical loop-free alternates. Hence the name "Remote LFA FRR". When a link cannot be entirely protected with local LFA neighbors, the protecting router seeks the help of a remote LFA staging point.

3.1. Tunnels as Repair Paths

Consider an arbitrary protected link S-E. In LFA FRR, if a path to the destination from a neighbor N of S does not cause a packet to loop back over the link S-E (i.e. N is a loop-free alternate), then S can send the packet to N and the packet will be delivered to the destination using the pre-failure forwarding information. If there is no such LFA neighbor, then S may be able to create a virtual LFA by using a tunnel to carry the packet to a point in the network which is not a direct neighbor of S from which the packet will be delivered to the destination without looping back to S. In this document such a tunnel is termed a repair tunnel. The tail-end of this tunnel is called a "remote LFA" or a "PQ node".

Note that the repair tunnel terminates at some intermediate router between S and E, and not E itself. This is clearly the case, since if it were possible to construct a tunnel from S to E then a conventional LFA would have been sufficient to effect the repair.

3.2. Tunnel Requirements

There are a number of IP in IP tunnel mechanisms that may be used to fulfil the requirements of this design, such as IP-in-IP [RFC1853] and GRE[RFC1701] .

In an MPLS enabled network using LDP[RFC5036], a simple label stack[RFC3032] may be used to provide the required repair tunnel. In this case the outer label is S's neighbor's label for the repair tunnel end point, and the inner label is the repair tunnel end point's label for the packet destination. In order for S to obtain the correct inner label it is necessary to establish a directed LDP session[RFC5036] to the tunnel end point.

The selection of the specific tunnelling mechanism (and any necessary enhancements) used to provide a repair path is outside the scope of this document. The authors simply note that deployment in an MPLS/LDP environment is extremely simple and straight-forward as an LDP LSP from S to the PQ node is readily available, and hence does not require any new protocol extension or design change. This LSP is automatically established as a basic property of LDP behavior. The performance of the encapsulation and decapsulation is also excellent as encapsulation is just a push of one label (like conventional MPLS TE FRR) and the decapsulation occurs naturally at the penultimate hop before the PQ node.

When a failure is detected, it is necessary to immediately redirect traffic to the repair path. Consequently, the repair tunnel used must be provisioned beforehand in anticipation of the failure. Since the location of the repair tunnels is dynamically determined it is necessary to establish the repair tunnels without management action. Multiple repairs may share a tunnel end point.

4. Construction of Repair Paths

4.1. Identifying Required Tunneled Repair Paths

Not all links will require protection using a tunneled repair path. Referring to Figure 1, if E can already be protected via an LFA, S-E does not need to be protected using a repair tunnel, since all destinations normally reachable through E must therefore also be protectable by an LFA. Such an LFA is frequently termed a "link LFA". Tunneled repair paths are only required for links which do not have a link LFA.

4.2. Determining Tunnel End Points

The repair tunnel endpoint needs to be a node in the network reachable from S without traversing S-E. In addition, the repair tunnel end point needs to be a node from which packets will normally flow towards their destination without being attracted back to the failed link S-E.

Note that once released from the tunnel, the packet will be forwarded, as normal, on the shortest path from the release point to its destination. This may result in the packet traversing the router E at the far end of the protected link S-E., but this is obviously not required.

The properties that are required of repair tunnel end points are therefore:

- The repair tunneled point MUST be reachable from the tunnel source without traversing the failed link; and

- When released, tunneled packets MUST proceed towards their destination without being attracted back over the failed link.

Provided both these requirements are met, packets forwarded over the repair tunnel will reach their destination and will not loop.

In some topologies it will not be possible to find a repair tunnel endpoint that exhibits both the required properties. For example if the ring topology illustrated in Figure 1 had a cost of 4 for the link B-C, while the remaining links were cost 1, then it would not be possible to establish a tunnel from S to C (without resorting to some form of source routing).

4.2.1. Computing Repair Paths

The set of routers which can be reached from S without traversing S-E is termed the P-space of S with respect to the link S-E. The P-space can be obtained by computing a shortest path tree (SPT) rooted at S and excising the sub-tree reached via the link S-E (including those which are members of an ECMP). In the case of Figure 1 the P-space comprises nodes A and B only. Expressed in cost terms the set of routers {P} are those for which the shortest path cost S->P is strictly less than the shortest path cost S->E->P.

The set of routers from which the node E can be reached, by normal forwarding, without traversing the link S-E is termed the Q-space of E with respect to the link S-E. The Q-space can be obtained by computing a reverse shortest path tree (rSPT) rooted at E, with the sub-tree which traverses the failed link excised (including those which are members of an ECMP). The rSPT uses the cost towards the root rather than from it and yields the best paths towards the root from other nodes in the network. In the case of Figure 1 the Q-space comprises nodes C and D only. Expressed in cost terms the set of routers {Q} are those for which the shortest path cost E->Q is strictly less than the shortest path cost E->S->Q. In Figure 1 the intersection of the E's Q-space with S's P-space defines the set of viable repair tunnel end-points, known as "PQ nodes". As can be seen, for the case of Figure 1 there is no common node and hence no viable repair tunnel end-point.

Note that the Q-space calculation could be conducted for each individual destination and a per-destination repair tunnel end point determined. However this would, in the worst case, require an SPF computation per destination which is not currently considered to be scalable. We therefore use the Q-space of E as a proxy for the

Q-space of each destination. This approximation is obviously correct since the repair is only used for the set of destinations which were, prior to the failure, routed through node E. This is analogous to the use of link-LFAs rather than per-prefix LFAs.

4.2.2. Extended P-space

The description in Section 4.2.1 calculated router S's P-space rooted at S itself. However, since router S will only use a repair path when it has detected the failure of the link S-E, the initial hop of the repair path need not be subject to S's normal forwarding decision process. Thus we introduce the concept of extended P-space. Router S's extended P-space is the union of the P-spaces of each of S's neighbours. This may be calculated by computing the an SPT at each of S's neighbors (N) (excluding E) and excising the subtree reached via the path N->S->E. The use of extended P-space may allow router S to reach potential repair tunnel end points that were otherwise unreachable. In cost terms a router is in extended P-space if the shortest path cost S-N->P is strictly less than the shortest path cost S-E->P.

Another way to describe extended P-space is that it is the union of (un-extended) P-space and the set of destinations for which S has a per-prefix LFA protecting the link S-E. i.e. the repair tunnel end point can be reached either directly or using a per-prefix LFA.

Since in the case of Figure 1 node A is a per-prefix LFA for the destination node C, the set of extended P-space nodes comprises nodes A, B and C. Since node C is also in E's Q-space, there is now a node common to both extended P-space and Q-space which can be used as a repair tunnel end-point to protect the link S-E.

4.2.3. Selecting Repair Paths

The mechanisms described above will identify all the possible repair tunnel end points that can be used to protect a particular link. In a well-connected network there are likely to be multiple possible release points for each protected link. All will deliver the packets correctly so, arguably, it does not matter which is chosen. However, one repair tunnel end point may be preferred over the others on the basis of path cost or some other selection criteria.

There is no technical requirement for the selection criteria to be consistent across all routers, but such consistency may be desirable from an operational point of view. In general there are advantages in choosing the repair tunnel end point closest (shortest metric) to S. Choosing the closest maximises the opportunity for the traffic to be load balanced once it has been released from the tunnel. For

consistency in behavior is RECOMMENDED that member of the set of routers {P} with the lowest cost S->P be the default choice for P. In the event of a tie the router with the lowest node identifier SHOULD be selected.

5. Example Application of Remote LFAs

An example of a commonly deployed topology which is not fully protected by LFAs alone is shown in Figure 3. PE1 and PE2 are connected in the same site. P1 and P2 may be geographically separated (inter-site). In order to guarantee the lowest latency path from/to all other remote PEs, normally the shortest path follows the geographical distance of the site locations. Therefore, to ensure this, a lower IGP metric (5) is assigned between PE1 and PE2. A high metric (1000) is set on the P-PE links to prevent the PEs being used for transit traffic. The PEs are not individually dual-homed in order to reduce costs.

This is a common topology in SP networks.

When a failure occurs on the link between PE1 and P2, PE1 does not have an LFA for traffic reachable via P1. Similarly, by symmetry, if the link between PE2 and P1 fails, PE2 does not have an LFA for traffic reachable via P2.

Increasing the metric between PE1 and PE2 to allow the LFA would impact the normal traffic performance by potentially increasing the latency.

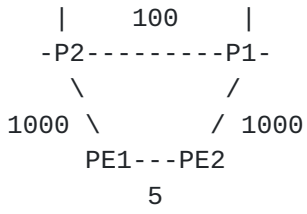

Figure 3: Example SP topology

Clearly, full protection can be provided, using the techniques described in this draft, by PE1 choosing P2 as a PQ node, and PE2 choosing P1 as a PQ node.

6. Node Failures

When the failure is a node failure rather than a link failure there is a danger that the RLFA repair will loop. This is discussed in detail in [I-D.bryant-ipfrr-tunnels]. In summary problem is that two of more of E's neighbors each with E as the next hop to some

destination D may attempt to repair a packet addressed to destination D via the other neighbor and then E, thus causing a loop to form. As will be noted from [I-D.bryant-ipfrr-tunnels], this can rapidly become a complex problem to address.

There are a number of ways to minimize the probability of a loop forming when a node failure occurs and there exists the possibility that two of E's neighbors may form a mutual repair.

1. Detect when a packet has arrived on some interface I that is also the interface used to reach the first hop on the RLFA path to PQ, and drop the packet. This is useful in the case of a ring topology.

2. Require that the path from PQ to destination D never passes through E (including in the ECMP case), i.e. only use node protecting paths in which the cost PQ to D is strictly less than the cost PQ to E plus the cost E to D.

3. Require that where the packet may pass through another neighbor of E, that node is down stream (i.e. strictly closer to D than the repairing node). This means that some neighbor of E (X) can repair via some other neighbor of E (Y), but Y cannot repair via X.

Case 1 accepts that loops may form and suppresses them by dropping packets. Dropping packets may be considered less detrimental than looping packets. Cases 2 and 3 above prevent the formation of a loop, but at the expense of a reduced repair coverage and at the cost of additional complexity in the algorithm to compute the repair path.

The probability of a node failure and the consequences of node failure in any particular topology will depend on the node design, the particular topology in use, and node failure strategy (including the null strategy). It is recommended that a network operator perform an analysis of the consequences and probability of node failure in their network, and determine whether the incidence and consequence of occurrence are acceptable.

7. Operation in an LDP environment

Where this technique is used in an MPLS network using LDP [RFC5036], S will need to push two labels onto the repair packet. First it needs to push PQ's label to the destination, and then it needs to push its own label for PQ. In the example Section 3.1 S already has the first hop (B) label for the PQ node (C) as a result of the ordinary operation of LDP. To get the PQ node (C) label for the destination (D), S needs to establish a targeted LDP session with C.

The label stack for normal operation and RLFA operation is shown below in Figure 4.

X = Normal label stack packet arriving at S
Y = Normal label stack packet leaving S
Z = RLFA label stack to D via C as PQ node

Figure 4

To establish an targeted LDP session with a candidate PQ node the repairing node (S) needs to know what IP address PQ is willing to use for targeted LDP sessions. This in turn requires PQ to advertise this address in the IGP in use. What address is used, how this is advertised in the IGP, and whether this is a special IP address or an IP address also used for some other purpose is out of scope for this document and must be specified in an IGP specific RFC.

8. Historical Note

The basic concepts behind Remote LFA were invented in 2002 and were later included in [I-D.bryant-ipfrr-tunnels], submitted in 2004.

[I-D.bryant-ipfrr-tunnels], targeted a 100% protection coverage and hence included additional mechanisms on top of the Remote LFA concept. The addition of these mechanisms made the proposal very complex and computationally intensive and it was therefore not pursued as a working group item.

As explained in [I-D.filsfils-rtgwg-lfa-applicability], the purpose of the LFA FRR technology is not to provide coverage at any cost. A solution for this already exists with MPLS TE FRR. MPLS TE FRR is a mature technology which is able to provide protection in any topology thanks to the explicit routing capability of MPLS TE.

The purpose of LFA FRR technology is to provide for a simple FRR solution when such a solution is possible. The first step along this simplicity approach was "local" LFA [RFC5286]. We propose "Remote LFA" as a natural second step. The following section motivates its benefits in terms of simplicity, incremental deployment and significant coverage increase.

9. Benefits

Remote LFAs preserve the benefits of RFC5286: simplicity, incremental deployment and good protection coverage.

9.1. Simplicity

The remote LFA algorithm is simple to compute.

- The extended P space does not require any new computation (it is known once per-prefix LFA computation is completed).
- The Q-space is a single reverse SPF rooted at the neighbor.
- The directed LDP session is automatically computed and established.

In edge topologies (square, ring), the directed LDP session position and number is deterministic and hence troubleshooting is simple.

In core topologies, our simulation indicates that the 90th percentile number of LDP sessions per node to achieve the significant Remote LFA coverage observed in section 7.3 is <= 6. This is insignificant compared to the number of LDP sessions commonly deployed per router which is frequently is in the several hundreds.

9.2. Incremental Deployment

The establishment of the directed LDP session to the PQ node does not require any new technology on the PQ node. Indeed, routers commonly support the ability to accept a remote request to open a directed LDP session. The new capability is restricted to the Remote-LFA computing node (the originator of the LDP session).

9.3. Significant Coverage Extension

The previous sections have already explained how Remote LFAs provide protection for frequently occurring edge topologies: square and rings. In the core, we extend the analysis framework in section 4.3 of [I-D.filsfils-rtgwg-lfa-applicability]and provide hereafter the Remote LFA coverage results for the 11 topologies:

Topology	Per-link LFA	Per-prefix LFA	Remote LFA
T1	45%	77%	78%
T2	49%	99%	100%
T3	88%	99%	99%
T4	68%	84%	92%
T5	75%	94%	99%
T6	87%	99%	100%
T7	16%	67%	96%
T8	87%	100%	100%
T9	67%	80%	98%
T10	98%	100%	100%
T11	59%	77%	95%
Average	67%	89%	96%
Median	68%	94%	99%

Another study[ISOCORE2010] confirms the significant coverage increase provided by Remote LFAs.

10. Complete Protection

As shown in the previous table, Remote LFA provides for 96% average (99% median) protection in the 11 analyzed SP topologies.

In an MPLS network, this is achieved without any scalability impact as the tunnels to the PQ nodes are always present as a property of an LDP-based deployment.

In the very few cases where P and Q spaces have an empty intersection, one could select the closest node in the Q space and signal an explicitely-routed RSVP TE LSP to that Q node. A directed LDP session is then established with the selected Q node and the rest of the solution is identical to that described elsewhere in this document.

The drawbacks of this solution are:

1. only available for MPLS network;

2. the addition of LSPs in the SP infrastructure.

This extension is described for exhaustivity. In practice, the "Remote LFA" solution should be preferred for three reasons: its simplicity, its excellent coverage in the analyzed backbones and its

complete coverage in the most frequent access/aggregation topologies (box or ring).

11. IANA Considerations

There are no IANA considerations that arise from this architectural description of IPFRR. The RFC Editor may remove this section on publication.

12. Security Considerations

The security considerations of RFC 5286 also apply.

To prevent their use as an attack vector the repair tunnel endpoints SHOULD be assigned from a set of addresses that are not reachable from outside the routing domain.

13. Acknowledgments

The authors acknowledge the technical contributions made to this work by Stefano Previdi.

14. Informative References

[I-D.bryant-ipfrr-tunnels]
Bryant, S., Filsfils, C., Previdi, S., and M. Shand, "IP Fast Reroute using tunnels", draft-bryant-ipfrr-tunnels-03 (work in progress), November 2007.

[I-D.filsfils-rtgwg-lfa-applicability]
Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "LFA applicability in SP networks", draft-filsfils-rtgwg-lfa-applicability-00 (work in progress), March 2010.

[ISOCORE2010]
So, N., Lin, T., and C. Chen, "LFA (Loop Free Alternates) Case Studies in Verizon's LDP Network", 2010.

[RFC1701] Hanks, S., Li, T., Farinacci, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 1701, October 1994.

[RFC1853] Simpson, W., "IP in IP Tunneling", RFC 1853, October 1995.

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

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

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

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

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

Authors' Addresses

Stewart Bryant
Cisco Systems
250, Longwater, Green Park,
Reading RG2 6GB, UK
UK

Email: stbryant@cisco.com

Clarence Filsfils
Cisco Systems
De Kleetlaan 6a
1831 Diegem
Belgium

Email: cfilsfil@cisco.com

Stefano Previdi
Cisco Systems

Email: sprevidi@cisco.com

Mike Shand
Independent Contributor

Email: imc.shand@gmail.com

Ning So
Tata Communications
Mobile Broadband Services

Email: Ning.So@tatacommunications.com