Network Working Group
Internet-Draft
Intended status: Informational
Expires: March 22, 2010

M. Shand
S. Bryant
Cisco Systems
September 18, 2009

IP Fast Reroute Framework
draft-ietf-rtgwg-ipfrr-framework-12

Status of this Memo

Copyright Notice

Abstract

This document provides a framework for the development of IP fast-reroute mechanisms which provide protection against link or router failure by invoking locally determined repair paths. Unlike MPLS

fast-reroute, the mechanisms are applicable to a network employing conventional IP routing and forwarding.

Table of Contents

1. Terminology

This section defines words and acronyms used in this draft and other drafts discussing IP fast-reroute.

D	Used to denote the destination router under discussion.
Distance_opt(A,B)	The distance of the shortest path from A to B.
Downstream Path	This is a subset of the loop-free alternates where the neighbor N meets the following condition:- Distance_opt(N, D) < Distance_opt(S,D)
E	Used to denote the router which is the primary next-hop neighbor to get from S to the destination D. Where there is an ECMP set for the shortest path from S to D, these are referred to as E_1, E_2, etc.
ECMP	Equal cost multi-path: Where, for a particular destination D, multiple primary next-hops are used to forward traffic because there exist multiple shortest paths from S via different output layer-3 interfaces.
FIB	Forwarding Information Base. The database used by the packet forwarder to determine what actions to perform on a packet.
IPFRR	IP fast-reroute.
Link(A->B)	A link connecting router A to router B.

LFA Loop Free Alternate. A neighbor N, that is not a primary next-hop neighbor E, whose shortest path to the destination D does not go back through the router S. The neighbor N must meet the following condition:-

Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D)

Loop Free Neighbor A neighbor N_i, which is not the particular primary neighbor E_k under discussion, and whose shortest path to D does not traverse S. For example, if there are two primary neighbors E_1 and E_2, E_1 is a loop-free neighbor with regard to E_2 and vice versa.

Loop Free Link Protecting Alternate
A path via a Loop-Free Neighbor N_i that reaches destination D without going through the particular link of S that is being protected. In some cases the path to D may go through the primary neighbor E.

Loop Free Node-protecting Alternate
A path via a Loop-Free Neighbor N_i that reaches destination D without going through the particular primary neighbor (E) of S which is being protected.

N_i The ith neighbor of S.

Primary Neighbor A neighbor N_i of S which is one of the next hops for destination D in S's FIB prior to any failure.

R_i_j The jth neighbor of N_i.

Routing Transition The process whereby routers converge on a new

topology. In conventional networks this process frequently causes some disruption to packet delivery.

RPF — Reverse Path Forwarding. I.e. checking that a packet is received over the interface which would be used to send packets addressed to the source address of the packet.

S — Used to denote a router that is the source of a repair that is computed in anticipation of the failure of a neighboring router denoted as E, or of the link between S and E. It is the viewpoint from which IP fast-reroute is described.

SPF — Shortest Path First, e.g. Dijkstra's algorithm.

SPT — Shortest path tree

Upstream Forwarding Loop
A forwarding loop that involves a set of routers, none of which are directly connected to the link that has caused the topology change that triggered a new SPF in any of the routers.

2. Introduction

When a link or node failure occurs in a routed network, there is inevitably a period of disruption to the delivery of traffic until the network re-converges on the new topology. Packets for destinations which were previously reached by traversing the failed component may be dropped or may suffer looping. Traditionally such disruptions have lasted for periods of at least several seconds, and most applications have been constructed to tolerate such a quality of service.

Recent advances in routers have reduced this interval to under a second for carefully configured networks using link state IGPs.

However, new Internet services are emerging which may be sensitive to periods of traffic loss which are orders of magnitude shorter than this.

Addressing these issues is difficult because the distributed nature of the network imposes an intrinsic limit on the minimum convergence time which can be achieved.

However, there is an alternative approach, which is to compute backup routes that allow the failure to be repaired locally by the router(s) detecting the failure without the immediate need to inform other routers of the failure. In this case, the disruption time can be limited to the small time taken to detect the adjacent failure and invoke the backup routes. This is analogous to the technique employed by MPLS fast-reroute [RFC4090], but the mechanisms employed for the backup routes in pure IP networks are necessarily very different.

This document provides a framework for the development of this approach.

3. Problem Analysis

The duration of the packet delivery disruption caused by a conventional routing transition is determined by a number of factors:

1. The time taken to detect the failure. This may be of the order of a few milliseconds when it can be detected at the physical layer, up to several tens of seconds when a routing protocol Hello is employed. During this period packets will be unavoidably lost.

2. The time taken for the local router to react to the failure. This will typically involve generating and flooding new routing updates, perhaps after some hold-down delay, and re-computing the router's FIB.

3. The time taken to pass the information about the failure to other routers in the network. In the absence of routing protocol packet loss, this is typically between 10 milliseconds and 100 milliseconds per hop.

4. The time taken to re-compute the forwarding tables. This is typically a few milliseconds for a link state protocol using Dijkstra's algorithm.

5. The time taken to load the revised forwarding tables into the forwarding hardware. This time is very implementation dependant and also depends on the number of prefixes affected by the failure, but may be several hundred milliseconds.

The disruption will last until the routers adjacent to the failure have completed steps 1 and 2, and then all the routers in the network whose paths are affected by the failure have completed the remaining steps.

The initial packet loss is caused by the router(s) adjacent to the failure continuing to attempt to transmit packets across the failure until it is detected. This loss is unavoidable, but the detection time can be reduced to a few tens of milliseconds as described in Section 4.1.

In some topologies subsequent packet loss may be caused by the "micro-loops" which may form as a result of temporary inconsistencies between routers' forwarding tables[I-D.ietf-rtgwg-lf-conv-frmwk]. These inconsistencies are caused by steps 3, 4 and 5 above and in many routers it is step 5 which is both the largest factor and which has the greatest variance between routers. The large variance arises from implementation differences and from the differing impact that a failure has on each individual router. For example, the number of prefixes affected by the failure may vary dramatically from one router to another.

In order to achieve packet disruption times which are commensurate

with the failure detection times two mechanisms may be required:-

1. A mechanism for the router(s) adjacent to the failure to rapidly invoke a repair path, which is unaffected by any subsequent re-convergence.

2. In topologies that are susceptible to micro-loops, a mechanism to prevent the effects of any micro-loops during subsequent re-

convergence.

Performing the first task without the second may result in the repair path being starved of traffic and hence being redundant. Performing the second without the first will result in traffic being discarded by the router(s) adjacent to the failure.

Repair paths may always be used in isolation where the failure is short-lived. In this case, the repair paths can be kept in place until the failure is repaired in which case there is no need to advertise the failure to other routers.

Similarly, micro-loop avoidance may be used in isolation to prevent loops arising from pre-planned management action. In which case the link or node being shut down can remain in service for a short time after its removal has been announced into the network, and hence it can function as its own "repair path".

Note that micro-loops may also occur when a link or node is restored to service and thus a micro-loop avoidance mechanism may be required for both link up and link down cases.

4. Mechanisms for IP Fast-reroute

The set of mechanisms required for an effective solution to the problem can be broken down into the sub-problems described in this section.

4.1. Mechanisms for fast failure detection

It is critical that the failure detection time is minimized. A number of well documented approaches are possible, such as:

1. Physical detection; for example, loss of light.

2. Routing protocol independent protocol detection; for example, The Bidirectional Failure Detection protocol [I-D.ietf-bfd-base].

3. Routing protocol detection; for example, use of "fast Hellos".

4.2. Mechanisms for repair paths

Once a failure has been detected by one of the above mechanisms, traffic which previously traversed the failure is transmitted over one or more repair paths. The design of the repair paths should be such that they can be pre-calculated in anticipation of each local failure and made available for invocation with minimal delay. There are three basic categories of repair paths:

1. Equal cost multi-paths (ECMP). Where such paths exist, and one or more of the alternate paths do not traverse the failure, they may trivially be used as repair paths.

2. Loop free alternate paths. Such a path exists when a direct neighbor of the router adjacent to the failure has a path to the destination which can be guaranteed not to traverse the failure.

3. Multi-hop repair paths. When there is no feasible loop free alternate path it may still be possible to locate a router, which is more than one hop away from the router adjacent to the failure, from which traffic will be forwarded to the destination without traversing the failure.

ECMP and loop free alternate paths (as described in [RFC5286]) offer the simplest repair paths and would normally be used when they are available. It is anticipated that around 80% of failures (see Section 4.2.2) can be repaired using these basic methods alone.

Multi-hop repair paths are more complex, both in the computations required to determine their existence, and in the mechanisms required to invoke them. They can be further classified as:

1. Mechanisms where one or more alternate FIBs are pre-computed in all routers and the repaired packet is instructed to be forwarded using a "repair FIB" by some method of per packet signaling such as detecting a "U-turn" [I-D.atlas-ip-local-protect-uturn] , [FIFR] or by marking the packet [SIMULA].

2. Mechanisms functionally equivalent to a loose source route which is invoked using the normal FIB. These include tunnels [I-D.bryant-ipfrr-tunnels], alternative shortest paths [I-D.tian-frr-alt-shortest-path] and label based mechanisms.

3. Mechanisms employing special addresses or labels which are installed in the FIBs of all routers with routes pre-computed to avoid certain components of the network. For example

[I-D.ietf-rtgwg-ipfrr-notvia-addresses].

In many cases a repair path which reaches two hops away from the router detecting the failure will suffice, and it is anticipated that around 98% of failures (see Section 4.2.2) can be repaired by this method. However, to provide complete repair coverage some use of longer multi-hop repair paths is generally necessary.

4.2.1. Scope of repair paths

A particular repair path may be valid for all destinations which require repair or may only be valid for a subset of destinations. If a repair path is valid for a node immediately downstream of the failure, then it will be valid for all destinations previously reachable by traversing the failure. However, in cases where such a repair path is difficult to achieve because it requires a high order multi-hop repair path, it may still be possible to identify lower order repair paths (possibly even loop free alternate paths) which allow the majority of destinations to be repaired. When IPFRR is unable to provide complete repair, it is desirable that the extent of the repair coverage can be determined and reported via network management.

There is a trade-off to be achieved between minimizing the number of repair paths to be computed, and minimizing the overheads incurred in using higher order multi-hop repair paths for destinations for which they are not strictly necessary. However, the computational cost of determining repair paths on an individual destination basis can be very high.

It will frequently be the case that the majority of destinations may be repaired using only the "basic" repair mechanism, leaving a smaller subset of the destinations to be repaired using one of the more complex multi-hop methods. Such a hybrid approach may go some way to resolving the conflict between completeness and complexity.

The use of repair paths may result in excessive traffic passing over a link, resulting in congestion discard. This reduces the effectiveness of IPFRR. Mechanisms to influence the distribution of repaired traffic to minimize this effect are therefore desirable.

4.2.2. Analysis of repair coverage

The repair coverage obtained is dependent on the repair strategy and highly dependent on the detailed topology and metrics. Estimates of the repair coverage quoted in this document are for illustrative

purposes only and may not be always be achievable.

In some cases the repair strategy will permit the repair of all single link or node failures in the network for all possible destinations. This can be defined as 100% coverage. However, where the coverage is less than 100% it is important for the purposes of comparisons between different proposed repair strategies to define what is meant by such a percentage. There are four possibilities:

1. The percentage of links (or nodes) which can be fully protected for all destinations. This is appropriate where the requirement is to protect all traffic, but some percentage of the possible failures may be identified as being un-protectable.

2. The percentage of destinations which can be fully protected for all link (or node) failures. This is appropriate where the requirement is to protect against all possible failures, but some percentage of destinations may be identified as being un-protectable.

3. For all destinations (d) and for all failures (f), the percentage of the total potential failure cases (d*f) which are protected. This is appropriate where the requirement is an overall "best effort" protection.

4. The percentage of packets normally passing though the network that will continue to reach their destination. This requires a traffic matrix for the network as part of the analysis.

4.2.3. Link or node repair

A repair path may be computed to protect against failure of an adjacent link, or failure of an adjacent node. In general, link protection is simpler to achieve. A repair which protects against node failure will also protect against link failure for all destinations except those for which the adjacent node is a single point of failure.

In some cases it may be necessary to distinguish between a link or node failure in order that the optimal repair strategy is invoked. Methods for link/node failure determination may be based on

techniques such as BFD[I-D.ietf-bfd-base]. This determination may be made prior to invoking any repairs, but this will increase the period of packet loss following a failure unless the determination can be performed as part of the failure detection mechanism itself. Alternatively, a subsequent determination can be used to optimise an already invoked default strategy.

4.2.4. Maintenance of Repair paths

In order to meet the response time goals, it is expected (though not required) that repair paths, and their associated FIB entries, will be pre-computed and installed ready for invocation when a failure is detected. Following invocation the repair paths remain in effect until they are no longer required. This will normally be when the routing protocol has re-converged on the new topology taking into account the failure, and traffic will no longer be using the repair paths.

The repair paths have the property that they are unaffected by any topology changes resulting from the failure which caused their instantiation. Therefore there is no need to re-compute them during the convergence period. They may be affected by an unrelated simultaneous topology change, but such events are out of scope of this work (see Section 4.2.5).

Once the routing protocol has re-converged it is necessary for all repair paths to take account of the new topology. Various optimizations may permit the efficient identification of repair paths which are unaffected by the change, and hence do not require full re-computation. Since the new repair paths will not be required until the next failure occurs, the re-computation may be performed as a background task and be subject to a hold-down, but excessive delay in completing this operation will increase the risk of a new failure occurring before the repair paths are in place.

4.2.5. Multiple failures and Shared Risk Link Groups

Complete protection against multiple unrelated failures is out of scope of this work. However, it is important that the occurrence of

a second failure while one failure is undergoing repair should not result in a level of service which is significantly worse than that which would have been achieved in the absence of any repair strategy.

Shared Risk Link Groups (SRLGs) are an example of multiple related failures, and the more complex aspects of their protection is a matter for further study.

One specific example of an SRLG which is clearly within the scope of this work is a node failure. This causes the simultaneous failure of multiple links, but their closely defined topological relationship makes the problem more tractable.

4.3. Local Area Networks

Protection against partial or complete failure of LANs is more complex than the point to point case. In general there is a trade-off between the simplicity of the repair and the ability to provide complete and optimal repair coverage.

4.4. Mechanisms for micro-loop prevention

Ensuring the absence of micro-loops is important not only because they can cause packet loss in traffic which is affected by the failure, but because by saturating a link with looping packets they can also cause congestion loss of traffic flowing over that link which would otherwise be unaffected by the failure.

A number of solutions to the problem of micro-loop formation have been proposed and are summarized in [I-D.ietf-rtgwg-lf-conv-frmwk]. The following factors are significant in their classification:

1. Partial or complete protection against micro-loops.

2. Delay imposed upon convergence.

3. Tolerance of multiple failures (from node failures, and in general).

4. Computational complexity (pre-computed or real time).

5. Applicability to scheduled events.

6. Applicability to link/node reinstatement.

7. Topological constraints.

5. Management Considerations

While many of the management requirements will be specific to particular IPFRR solutions, the following general aspects need to be addressed:

1. Configuration

    A. Enabling/disabling IPFRR support.

    B. Enabling/disabling protection on a per link/node basis.

    C. Expressing preferences regarding the links/nodes used for repair paths.

    D. Configuration of failure detection mechanisms.

    E. Configuration of loop avoidance strategies

2. Monitoring and operational support

    A. Notification of links/nodes/destinations which cannot be protected.

    B. Notification of pre-computed repair paths, and anticipated traffic patterns.

    C. Counts of failure detections, protection invocations and packets forwarded over repair paths.

D. Testing repairs.

6. Scope and applicability

The initial scope of this work is in the context of link state IGPs. Link state protocols provide ubiquitous topology information, which facilitates the computation of repairs paths.

Provision of similar facilities in non-link state IGPs and BGP is a matter for further study, but the correct operation of the repair mechanisms for traffic with a destination outside the IGP domain is an important consideration for solutions based on this framework

7. IANA Considerations

There are no IANA considerations that arise from this framework document.

8. Security Considerations

This framework document does not itself introduce any security issues, but attention must be paid to the security implications of any proposed solutions to the problem.

Where the chosen solution uses tunnels it is necessary to ensure that the tunnel is not used as an attack vector. One method of addressing this is to use a set of tunnel endpoint addresses that are excluded

from use by user traffic.

There is a compatibility issue between IPFRR and reverse path forwarding (RPF) checking. Many of the solutions described in this document result in traffic arriving from a direction inconsistent with a standard RPF check. When a network relies on RPF checking for security purposes, an alternative security mechanism will need to be deployed in order to permit IPFRR to used.

Because the repair path will often be of a different length to the pre-failure path, security mechanisms which rely on specific TTL

values will be adversely affected.

9. Acknowledgements

The authors would like to acknowledge contributions made by Alia Atlas, Clarence Filsfils, Pierre Francois, Joel Halpern, Stefano Previdi and Alex Zinin.

10. Informative References

[FIFR] Nelakuditi, S., Lee, S., Lu, Y., Zhang, Z., and C. Chuah, "Fast local rerouting for handling transient link failures."", Tech. Rep. TR-2004-004, 2004.

[I-D.atlas-ip-local-protect-uturn] Atlas, A., "U-turn Alternates for IP/LDP Fast-Reroute", draft-atlas-ip-local-protect-uturn-03 (work in progress), March 2006.

[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.ietf-bfd-base] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", draft-ietf-bfd-base-09 (work in progress), February 2009.

[I-D.ietf-rtgwg-ipfrr-notvia-addresses] Shand, M., Bryant, S., and S. Previdi, "IP Fast Reroute Using Not-via Addresses", draft-ietf-rtgwg-ipfrr-notvia-addresses-04 (work in progress), July 2009.

[I-D.ietf-rtgwg-lf-conv-frmwk] Shand, M. and S. Bryant, "A Framework for Loop-free Convergence", draft-ietf-rtgwg-lf-conv-frmwk-05 (work in progress), June 2009.

[I-D.tian-frr-alt-shortest-path]
Tian, A., "Fast Reroute using Alternative Shortest Paths", draft-tian-frr-alt-shortest-path-01 (work in progress), July 2004.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

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

[SIMULA] Lysne, O., Kvalbein, A., Cicic, T., Gjessing, S., and A. Hansen, "Fast IP Network Recovery using Multiple Routing Configurations."", Infocom 10.1109/INFOCOM.2006.227, 2006, <http://folk.uio.no/amundk/infocom06.pdf>.

Authors' Addresses

Mike Shand
Cisco Systems
250, Longwater Avenue.
Reading, Berks RG2 6GB
UK

Email: mshand@cisco.com

Stewart Bryant
Cisco Systems
250, Longwater Avenue.
Reading, Berks RG2 6GB
UK

Email: stbryant@cisco.com