Network Working Group M. Shand
Internet Draft
Expiration Date: Dec 2004 Cisco Systems

June 2004

IP Fast Reroute Framework

draft-ietf-rtgwg-ipfrr-framework-00.txt

Status of this Memo

Abstract

This document provides a framework for the development of IP fast re-route 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. An essential part of such mechanisms is the prevention of packet loss caused by the loops which normally occur during the re-convergence of the network following a failure.

Terminology

This section defines words, acronyms, and actions used in this draft.

1. 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 [MPLSFRR], 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.

2. 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 mS 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, 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 10mS and 100mS per hop in a well designed router.

4. The time taken to re-compute the forwarding tables. This is typically a few mS 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 mS.

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 mS as described in section 3.1.

Subsequent packet loss is caused by the "micro-loops" which form because of temporary inconsistencies between routers' forwarding tables. These occur as a result of the different times at which routers update their forwarding tables to reflect the failure. These variable delays 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 it is necessary to perform two distinct tasks:

1. Provide 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. Provide a mechanism to prevent the effects of micro loops during subsequent re-convergence.

Performing the first task without the second will 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. Both tasks are necessary for an effective solution to the problem.

However, repair paths can be used in isolation where the failure is short-lived. The repair paths can be kept in place until the failure is repaired and there is no need to advertise the failure to other routers.

Similarly, micro loop avoidance can be used in isolation to prevent loops arising from pre-planned management action.

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

3. Mechanisms for IP Fast-route

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

3.1. Mechanisms for fast failure detection

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

1. Physical detection, such as loss of light.

2. The Bidirectional Failure Detection protocol [BFD]

3. Other forms of "fast hellos"

3.2. 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 multiple 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. Downstream paths. (Also known as "loop free feasible alternates".) Such a path exists when a direct neighbor of the router adjacent to the failure has a path to the destination which cannot traverse the failure.

3. Multihop repair paths. When there is no feasible downstream 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 downstream paths offer the simplest repair paths and would normally be used when they are available. It is anticipated that around 80% of failures can be repaired using these alone.

Multi-hop repair paths are considerably 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 signaling such as detecting a "U-turn" or marking the packet.

2. Mechanisms functionally equivalent to a loose source route which is invoked using the normal FIB. These include tunnels and label based mechanisms.

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 95% of failures can be repaired by this method. However, to effect complete repair coverage some use of longer multi-hop repair paths is generally necessary.

3.2.1. 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 downstream 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 tradeoff 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.

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.

3.2.2. 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. 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.

3.2.3. Multiple failures and Shared Risk 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 Groups are an example of multiple related failures, and 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.

3.3. Mechanisms for micro-loop prevention

Control of micro-loops is important not only because they can cause packet loss in traffic which is affected by the failure, but because they can also cause congestion loss of traffic which would otherwise be unaffected by the failure.

A number of solutions to the problem of micro-loop formation have been proposed. 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.

4. Scope and applicability

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

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

5. IANA considerations

There are no IANA considerations that arise from this description of IPFRR. However there may be changes to the IGPs to support IPFRR in which there will be IANA considerations.

6. 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.

Acknowledgments

Normative References

Internet-drafts are works in progress available from http://www.ietf.org/internet-drafts/

Informative References

Internet-drafts are works in progress available from
http://www.ietf.org/internet-drafts/

BFD Katz, D., and Ward, D., "Bidirectional Forwarding Detection", draft-katz-ward-bfd-01.txt, August 2003 (work in progress).

MPLSFRR Pan, P. et al, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", draft-ietf-mpls-rsvp-lsp-fastreroute-05.txt

Author's Address

Mike Shand
Cisco Systems,
250, Longwater Avenue,
Green Park,
Reading, RG2 6GB,
United Kingdom. Email: mshand@cisco.com

Full copyright statement