Network Working Group
Internet Draft
Expiration Date: Nov 2004

S. Bryant
C. Filsfils
S. Previdi
M. Shand
Cisco Systems

May 2004

IP Fast Reroute using tunnels

draft-bryant-ipfrr-tunnels-00.txt

Status of this Memo

Abstract

This draft describes an IP fast re-route mechanism that provides backup connectivity in the event of a link or router failure. In the absence of single points of failure and asymmetric costs, the mechanism provides complete protection against any single failure. If perfect repair is not possible, the identity of all the unprotected links and routers is known in advance. The draft also describes the mechanisms needed to prevent the packet loss caused by loops which normally occur during the reconvergence of the network following a failure.

Table of Contents

Terminology

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

A	Frequently used to denote a router that is the source of a repair path computed in anticipation of the failure of a neighboring router denoted as B.
B	Frequently used to denote a router whose anticipated failure is the subject of repair path computations.
Directed forwarding	The ability of the repairing router (A) to specify the next hop (Q) on exit from a tunnel end-point (P)
Extended P-space	The union of the p-space of the neighbors of a specific router with respect to a common component. Extended p-space does not include the additional space reachable though directed forwarding.
FIB	Forwarding Information Base. The database used by the packet forwarder to determine what actions to perform on a packet
IPFRR	IP fast re-route

P	The router in P-space to which a packet is tunneled for repair.
PQ	A router that is in both P and Q space and hence does not need directed forwarding.
P-space	P-space is the set of routers reachable from a specific router without any path (including equal cost path splits) transiting a specified component. For example, the P-space of A, is the set of routers that A can reach without using B (router failure case) or the A-B link failure case).
Q	The router in Q space, to which the packet is directed by router P on exit from the repair tunnel. Q will always be adjacent to P, or P itself.

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 a specified component.
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.
SPF	Shortest Path First, e.g. Dijkstra's algorithm.
SPT	Shortest path tree

1. Introduction

When the topology of a network changes (due to link or router failure, recovery or management action), the routers need to converge on a common view of the new topology. During this process, referred to as a routing transition, packet delivery between certain source/destination pairs may be disrupted. This occurs due to the time it takes for the topology change to be propagated around the network plus the time it takes each individual router to determine

and then update the forwarding information base (FIB) for the affected destinations. During this transition, packets are lost due to the continuing attempts to use of the failed component, and due to forwarding loops. Forwarding loops arise due to the inconsistent FIBs that occur as a result of the difference in time taken by routers to execute the transition process.

The service failures caused by routing transitions are largely hidden by higher-level protocols that retransmit the lost data. However new Internet services are emerging which are more sensitive to the packet disruption that occurs during a transition. To make the transition transparent to their users, these services require a short routing transition. Ideally, routing transitions would be completed in zero time with no packet loss.

Regardless of how optimally the mechanisms involved have been designed and implemented, it is inevitable that a routing transition will take some minimum interval that is greater than zero. The solution described here uses pre-computed backup routes and controlled notification of network changes. A set of repair paths temporarily provides substitute connectivity in place of a link, or router that has failed. Once the set of repair paths has been activated, there should be no further packet loss as a result of the associated failure. To achieve the maximum benefit from repair paths, they must be activated immediately a failure has been detected, and a controlled transition to normal operation invoked to prevent packet loss due to micro-looping. The packet loss attributable to the failure will then be confined to the unavoidable loss that occurs as a result of the latency of the failure detection mechanism itself.

The mechanisms described here have been designed for use with any link-state routing protocol.

2. Goals, non-goals, limitations and constraints

2.1. Goals

The following are the goals of IPFRR:

- Protect against any link or router failure in the network.
- No constraints on the network topology or link costs.

- Never worse than the existing routing convergence mechanism.
- Co-existence with non-IP fast-reroute capable routers in the network.

2.2. Non-Goals

The following are non-goals of IPFRR:

- Protection of a single point of failure.
- To provide protection in the presence of multiple concurrent failures other than those that occur due to the failure of a single router.
- Shared risk group protection.
- Complete fault coverage in networks that make use of asymmetric costs.

2.3. Limitations

The following limitations apply to IPFRR:

- Because the mechanisms described here rely on complete topological information from the link state routing protocol, they will only work within a single link state flooding domain.
- Reverse Path Forwarding (RPF) checks cannot be used in conjunction with IPFRR. This is because the use of tunnels may result in packets arriving over different interfaces than expected.

2.4. Constraints

The following constraints are assumed:

- Following a failure, only the routers adjacent to the failure have any knowledge of the failure.
- There is insufficient time following a failure to compute a repair strategy based on knowledge of the specific failure that has occurred.
- Multiple concurrent failures may not be protected.

3. Repair Paths

When a router detects an adjacent failure, it uses a set of repair paths in place of the failed component, and continues to use this until the completion of the routing transition. Only routers adjacent to the failed component are aware of the nature of the failure. Once the routing transition has been completed, the router will have no further use for the repair paths since all routers in the network will have revised their forwarding data and the failed link will have been eliminated from this computation.

Repair paths are pre-computed in anticipation of later failures so they can be promptly activated when a failure is detected.

Three types of repair path are considered here.

1. Equal cost path-split.

   Where a link is being used as a member of an equal cost path-split set for some destination, the other members of the set may be used to provide an alternative path, provided that they avoid the network component being protected.

2. Downstream Path.

   A 'downstream path' is a next hop that will get a packet nearer to its destination. It does not necessarily represent the shortest path to the destination but has the property that a packet sent on it will not loop back because, having traversed this hop, it is then closer to its destination.

3. Tunnel.

   A tunneled repair path tunnels traffic to some staging point from which it will travel to its destination using normal forwarding without looping back. The repair path can be thought of as providing a virtual link, originating at a router adjacent to a failure, and diverting traffic around the failure.

3.1. Tunnels as Repair Paths

The repair strategies described in this draft operate on the basis that if a packet can somehow be sent to the other side of the failure, it will subsequently proceed towards its destination exactly as if it had traversed the failed component. See Figure 1.

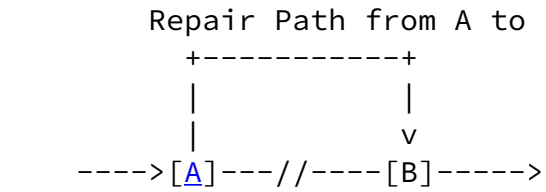


Figure 1 Simple Link Repair

Creating a repair path from A to B may require a packet to traverse an unnatural route. If a suitable natural path starts at a neighbor (i.e. it is a downstream path), then A can force the packet directly there. If this is not the case, then A must use a tunnel to force the packet down the repair path. Note that the tunnel does not have to go from A to B. The tunnel can terminate at any router in the network, provided that A can be sure that the packet will proceed correctly to its destination from that router.

A repair path computed for a link failure may not however work satisfactorily when the neighboring router has, itself, failed. This is illustrated in Figure 2.

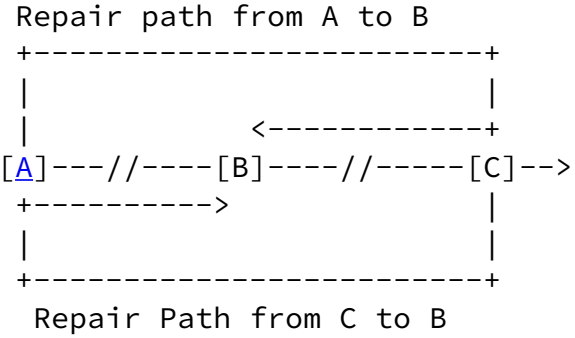


Figure 2 Looping Link Repair when Router Fails

Consider the case of a router B with just two neighbors A and C. When router B fails, both A and C will observe the failure of their local link to B, but will have no immediate knowledge that B itself has failed. If they were both to attempt to repair traffic around their local link, they would invoke mutual repairs which would loop.

Since it is not easy for a router to immediately distinguish between a link failure and the failure of its neighbor, repair paths are calculated in anticipation of adjacent router failure. Thus, for each of its protected links, router A (Figure 3) pre-computes a set of tunneled repair paths, one for each of the neighbors (C,D,E) of its neighbor B on the A-B link. The set of destinations that are normally assigned to link A-B will be assigned to a repair path based on the neighbor of B through which router B would have forwarded traffic to them.

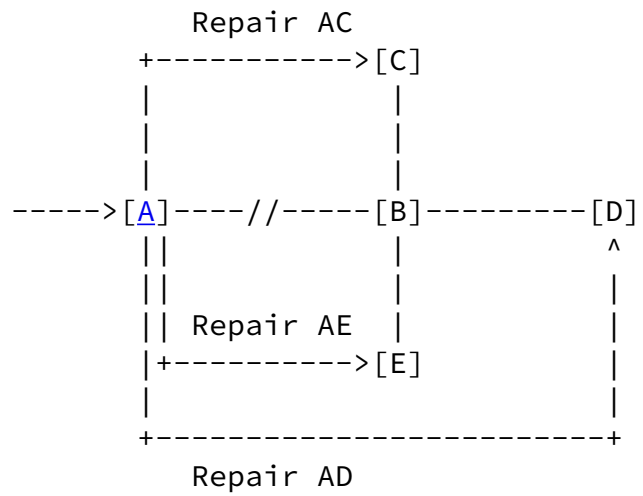


Figure 3: Repair paths in anticipation of a router failure

The set of repair paths in Figure 3 will function correctly in the case of link and router failure. However, in some network topologies they may not provide a means for traffic to reach router B itself. This is important in cases where B is a single point of failure and B is still functional (i.e. the failure was actually a failure of the A-B link). Hence, in addition to computing repair paths for the neighbors of its neighbor on a protected link, a router also

calculates a repair path for the neighbor itself. This is illustrated in Figure 4.

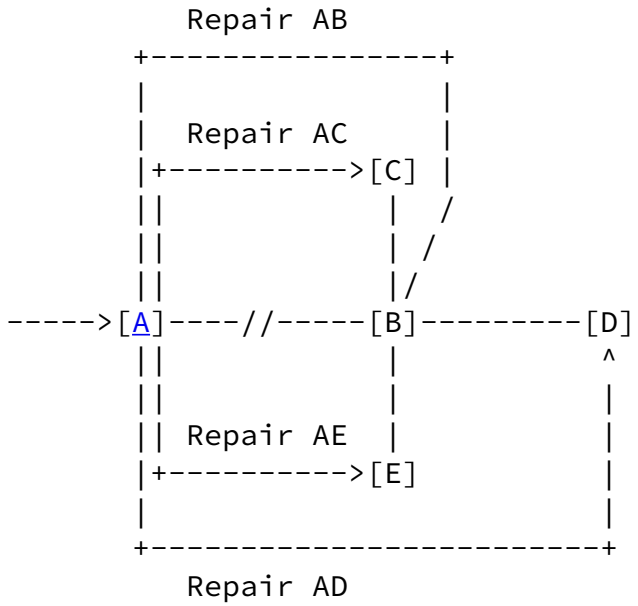


Figure 4 The full set of A-B repair paths.

In the event of a failure, the only traffic that is assigned to the link repair path (the AB repair) is that traffic which has no other path to its destination except via B. As we have already seen, there is a danger that traffic assigned to this link repair path may loop if B has failed, therefore, when the repair paths are invoked, a loop

detection mechanism is used which promptly detects the loop and, upon detection, withdraws the link (A-B) repair path from service.

3.2. Tunnel Requirements

The specific tunneling mechanism used to provide a repair path is outside the scope of this document. However the following sections describe the requirements for the tunneling mechanism.

3.2.1. Setup.

When a failure is detected, it is necessary to immediately redirect traffic to the repair paths. Consequently, the tunnels used must be provisioned beforehand in anticipation of the failure. IP fast re-

route will determine which tunnels it requires. It must therefore be possible to establish tunnels automatically, without management action, and without the need to manually establish context at the tunnel endpoint.

3.2.2. Multipoint

To reduce the number of tunnel endpoints in the network the tunnels should be be multi-point tunnels capable of receiving repair traffic from any IPFRR router in the network.

3.2.3. Directed forwarding.

Directed forwarding must be supported such that the router at the tunnel endpoint (P) can be directed by the router at the tunnel source (A) to forward the packet directly to a specific neighbor. Specification of the directed forwarding mechanism is outside the scope of this document.

3.2.4. Security

A lightweight security mechanism should be supported to prevent the abuse of the repair tunnels by an attacker. This is discussed in more detail in Section 12.

4. Construction of Repair Paths

4.1. Identifying Repair Path Targets

To establish protection for a link or node it is necessary to determine which neighbors of the neighboring node should be targets of repair paths. Normally all neighbors will be used as repair path

targets. However, in some topologies, not all neighbors will be needed as targets because, prior to the failure, no traffic was being forwarded through them by the repairing router. This can determined by examining the normal spanning tree computed by the repairing router.

In addition, the neighboring router B will also be the target of a

repair path for any destinations for which B is a single point of failure.

4.2. Determining Tunneled Repair Paths

The objective of each tunneled repair path is to deliver traffic to a target router when a link is observed to have failed. However, it is seldom possible to use the target router itself as the tunnel endpoint because other routers on the repair path, that have not learned of the failure, will forward traffic addressed to it using their least cost path which may be via the failed link. This is illustrated in Figure 5 in which all link costs are one in both directions. Router A's intended repair path for traffic to D when link A-B fails is the path W-X-Y-Z-D. However, if router A makes D be the tunnel endpoint and forwards the packet to router W, router W will immediately return it to A because its least cost path to D is A-B-D (cost 3 versus cost 4) and has no knowledge of the failure of link A-B.

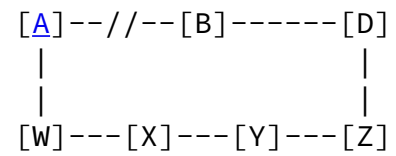


Figure 5. Repair path to target router D.

Thus the tunnel endpoint needs to be somewhere on the repair path such that packets addressed to the tunnel end point will not loop back towards router A. In addition, the release point needs to be somewhere such that when packets are released from the tunnel they will flow towards the target router (or their actual destination) without being attracted back to the failed link. By inspection, in Figure 5, suitable tunnel endpoints are routers X, Y, and Z.

Note that it is not essential that traffic assigned to a repair path actually traverse the target router for which the repair path was created. If, for example, in Figure 5, a packet's destination were normally reached via the path A-B-D-Z-?-?-?, once released at any of the possible tunnel endpoints, it would arrive at its destination by the best available route without traversing D.

In general, the properties that are required of tunnel endpoints are:

- the end point must be reachable from the tunnel source without traversing the failed link; and

- once released, tunneled packets will proceed towards their destination without being attracted back over the failed link or node.

Provided both of these conditions are met, packets forwarded on the repair path will not loop.

In some topologies it will not be possible to find a tunnel endpoint that exhibits both the required properties. For example, in Figure 5, if the cost of link X-Y were increased from one to four in both directions, there is no longer a viable endpoint within the fragment of the topology shown.

To solve this problem we introduce the concept of directed forwarding from the tunnel endpoint. Directed forwarding allows the originator of a tunneled packet to instruct that, when it is de-capsulated at the end of the tunnel, it be forwarded via a specific adjacency, and not be subjected to the normal forwarding decision process. This effectively allows the tunnel to be extended by one hop. So, for example, in Figure 5 with the cost of link X-Y set to four, it would be possible to select X as the tunnel endpoint with the directive that X always forward the packets it decapsulates via the adjacency to Y. Thus, router X is reached from A using normal forwarding, and directed forwarding is then used to force packets to router Y, from where D can be reached using normal forwarding.

Provided link costs are symmetrical, it can be proved that it is always possible to compute a tunneled repair path (possibly using directed forwarding) around a link failure.

The tunnel endpoint (P) and the release point (Q) may be coincident, or may be separated by at most one hop.

4.2.1. Computing Repair Paths

For a router A, determining tunneled repair paths around a neighboring router B, the set of potential tunnel end points includes all the routers that can be reached from A using normal forwarding without traversing the failed link A-B. This is termed the "P-space" of A with respect to the failure of B. Any router that is on an equal cost path split via the failed link is excluded from this set.

The resulting set defines all the possible tunnel end points that could be used in repair paths originating at router A for the failure of link A-B. This set can be obtained by computing a spanning tree rooted at A and excising the subtree reached via the A-B link.

The set of possible release points can be determined by computing the set of routers that can reach the repair path target without traversing the failed link. This is termed the "Q-space" of the target with respect to the failure. The Q-space can be obtained by computing a reverse spanning tree rooted at the repair path target, with the subtree which traverses the failed link (or node) excised.

The reverse spanning tree uses the cost towards the root rather than from it and yields the best paths towards the root from other nodes in the network.

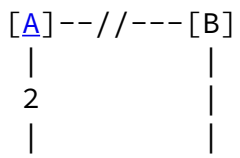
The intersection of the target's Q-space with A's P-space includes all the possible release points for any repair path not employing directed forwarding. Where there is no intersection, but there exist a pair of routers, P in A's P-space and Q in the target's Q-space, router P can be used as the tunnel endpoint with directed forwarding to the release point Q.

4.2.2. Extended P-space

The description in section 4.2.1 calculated router A's P-space rooted at A itself. However, since router A will only use a repair path when it has detected the failure of the link A-B, the initial hop of the repair path need not be subject to A's normal forwarding decision process. Thus we introduce the concept of extended P-space. Router A's extended P-space is the union of the P-spaces of each of A's neighbors. The use of extended P-space may allow router A to repair to targets that were otherwise unreachable.

4.2.3. Downstream Paths

Under certain circumstances, the target's Q-space will include a router that is a neighbor of A. This is traditionally referred to as a downstream path and has the property that a packet sent on it will not loop back because, having traversed this hop, it is then closer to its destination. A trivial example of this is shown in Figure 6.



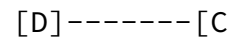


Figure 6. A topology that will permit a single-hop release point

When a downstream path exists, no tunneling is required.

4.2.4. Selecting Repair Paths

The mechanism described in section 4.2 will identify all the possible release points that can be used to reach each particular target. (The circumstances when no release points exist are described in section 4.4.) In a well-connected network there are likely to be multiple possible release points for each target, and all will work

correctly. For simplicity, one release point per target is chosen. All will deliver the packets correctly so, arguably, it does not matter which is chosen. However, one release point may be preferred over the others on the basis of path cost or some other criteria. It is an implementation matter as to how the release point is selected.

4.3. Assigning Traffic to Repair Paths

Once the repair path for each target has been selected, it is necessary to determine which of the destinations normally reached via the protected link should be assigned to which of the repair paths when the link fails.

This is achieved by recording which neighbor of B would be used to reach each destination reachable over A-B when running the original SPF. Traffic assignment is then simply a matter of assigning the traffic which B would have forwarded via each neighbor to the repair path which has that neighbor as its target.

Although the repair paths are calculated based on traffic addressed to specific targets, it can be proved that the traffic assignment algorithm guarantees that the repair path can be used for any traffic assigned to it.

Where B would normally split the traffic to a particular destination via two or more of its neighbors, it is an implementation decision whether the repaired traffic should be split across the corresponding set of repair paths.

The repair path to B itself is normally used just for traffic destined for B and any prefixes advertised by B. However, under some circumstances, it may be impossible to compute a repair path to one or more of B's neighbors, for example, because B is a single point of failure. In this case traffic for the destinations served by the otherwise irreparable targets is assigned to the repair path with B as its target, in the optimistic assumption that router B is still functioning. If router B is indeed still functioning, this will ensure delivery of the traffic. If, however, router B has failed, the traffic on this repair path will loop as previously shown in section 3.1. The way this is detected, and the course of action when it is detected, are described in section 5.3.

4.4. When no Repair Path is Possible

Under some circumstances, it will not be possible to identify a repair path to one or more of the targets. This can occur for the following reasons:

- The neighboring router that is presumed to have failed constitutes a single point of failure in the network.

- Severely asymmetric link costs may cause an otherwise viable physical repair path to be unusable.
- Interference may occur between the repair paths of individual targets.

In practice, these cases are unlikely to be encountered frequently. Networks that will benefit from the mechanisms described here will usually exhibit considerable redundancy and are normally operated with largely symmetric link costs. Note that a router's inability to compute a full set of repair paths for one of its links does not necessarily affect its ability to do so for its other links.

Example topologies illustrating each of the three cases above are described in the following subsections.

4.4.1. Unreachable Target

If the failure of a neighboring router makes one or more of its neighbors genuinely unreachable, clearly it will not be possible to establish a repair path to such targets. Such single points of failure are not expected to be encountered frequently in properly designed networks, and will probably occur only when the network has previously suffered other failures that have reduced its connectivity.

4.4.2. Asymmetric Link Costs

When link costs have been set asymmetrically, it is possible that a repair path cannot be constructed even using directed forwarding.

Although it is trivial to construct a network fragment with this property, this should not be regarded as a major problem. Firstly, asymmetric link costs are seldom used deliberately. And, secondly, even when an asymmetric link cost prevents one potential repair path being used, there will normally be other ones available.

4.4.3. Interference Between Potential Node Repair Paths

Under some circumstances the existence of one neighbor may interfere with a potential repair path to another. Consider the topology shown in Figure 7 in which all links have a symmetrical cost of one, with the exception of that between H and G, which has a cost of 3. In this example, the fact that router F is a neighbor of B prevents the discovery of a repair path from router A to router C despite the existence of an apparently suitable path.

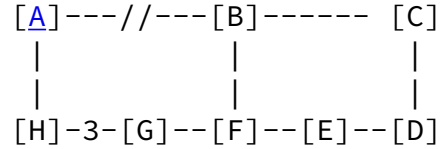


Figure 7. Interference between repair paths

A repair path from router A to F can use F itself as the release

point by employing directed forwarding from G. However, it is not possible to identify a suitable release point for a repair path to router C within the topology shown since there is nowhere that router A can reach that will subsequently forward traffic to router C except via the forbidden link B-C (F's least cost path to C is F-B-C). This is because the extended P-space of router A is separated by more than one hop from the Q-space of router C.

Since the topology shown in Figure 7 will typically form part of a much larger topology, a different, and possibly more circuitous repair path from A to C, that does not go via F, may be discovered. This is illustrated in Figure 8. In this enhanced topology, a repair path to C using Y as the release point can be used.

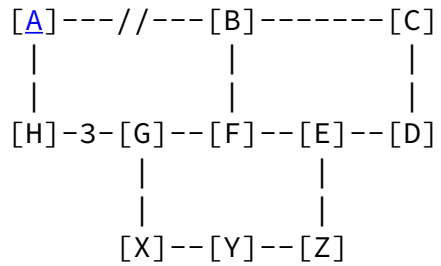


Figure 8. Resolving interference in a larger network

Note that, in Figure 8, if the traffic for C were assigned to the repair path for F, it would correctly reach C because F would assign it to its repair path to C. That is, packets from A to C would travel via two successive tunnels. Consequently, this is referred to as a "secondary repair path". However, it is not always the case that interference can be handled in this fashion and it is possible to create looping repair paths.

One possibility of looping repair paths is illustrated in Figure 9. All links have a symmetrical cost of one with the exception of HG, which is cost 3 in either direction, and ED and DC which are cost 5 in the indicated direction and cost 1 in the other.

[A]---//---[B]--------[C]

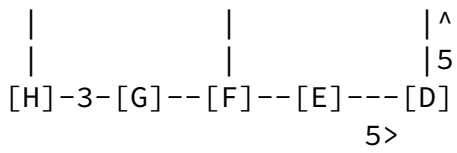


Figure 9 Looping secondary repair paths

In this topology, A can establish a repair path to F, but cannot establish a repair path to C because of interference. Router A might assign traffic intended for C onto its repair path to F expecting it to undergo a secondary repair towards C. However, because of the asymmetrical link costs, F is unable to establish a repair path to C. It is only able to establish a repair path to A. If F, like A, elected to forward repaired traffic to C using its (only) repair path to A, similarly expecting a secondary repair to get it to its destination, traffic for C would loop between A and F. Thus when interference occurs, the possibility of a secondary repair path cannot be relied upon to ensure that traffic reaches its destination.

In order to determine the viability of secondary repair paths, it is necessary for each router to take into account the repair paths which the other neighbors of router B can achieve. These can be computed locally by running the repair path computation algorithms rooted at each of those neighbors. It is only necessary to compute the repair paths from the routers to which router A can establish repair paths, with targets of those routers to which repair paths have not yet been established.

It is then possible to determine whether all routers can now be reached by invoking secondary (or if necessary tertiary, etc.) repair paths, and if so, to which primary repair path traffic for each target should be assigned.

There is another, more subtle, possibility of loops arising when secondary repair paths are used. This is illustrated in Figure 10, where all links are cost 1 with the exception of JI which has a cost 5 in that direction and cost 1 in the direction IJ.

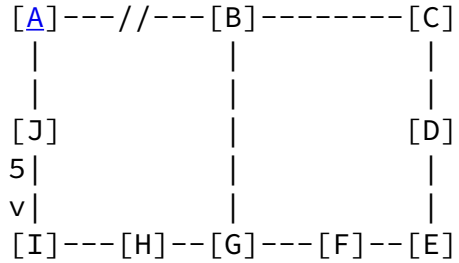


Figure 10 Example of an apparently non-looping secondary repair path which results in a loop.

Router A has a primary repair path to G (with a release point of I), and G has a primary repair path to C (with a release point of E). It

would appear that these form a non-looping secondary repair path from A to C. As usual, the primary repair path from A to G has been computed on the basis of destinations normally reachable through BG. However, when making use of the secondary repair path, the traffic inserted in the repair path from A to G will be destined not for one of the routers normally reachable via BG, but for C. Hence this repair path is not necessary valid for such traffic, and in this example it will have a 50% probability of being forwarded back along the path IJABC, and hence looping.

This problem can in general be avoided by choosing a release point for the initial primary repair with the property that traffic for the secondary target (C) is guaranteed to traverse the primary target (G). This can be achieved by computing the reverse SPF rooted at the secondary target (C) and examining the sub-tree which traverses the primary target. It can be proved that in the absence of asymmetric link costs, such a release point will always exist. Where asymmetric link costs prevent this, the traffic can be encapsulated to the intermediate router (G), which may require the use of double encapsulation. On reaching router G, the traffic for C is decapsulated and then forwarded in G's primary repair path to C (via router E, in the example).

4.5. Multi-homed Prefixes

Up to this point, it has been assumed that any particular prefix is "attached" to exactly one router in the network, and consequently only the routers in the network need be considered when constructing repair paths, etc. However, in many cases the same prefix will be attached to two or more routers. Common cases are: -

- The subnet present on a link is advertised from both ends of the link.
- Prefixes are propagated from one routing domain to another by multiple routers.
- Prefixes are advertised from multiple routers to provide resilience in the event of the failure of one of the routers.

In general, this causes no particular problems, and the shortest route to each prefix (and hence which of the routers to which it is attached should be used to reach it) is resolved by the normal SPF

process. However, in the particular case where one of the instances of a prefix is attached to router B, or to a router for which router B is a single point of failure, the situation is more complicated.

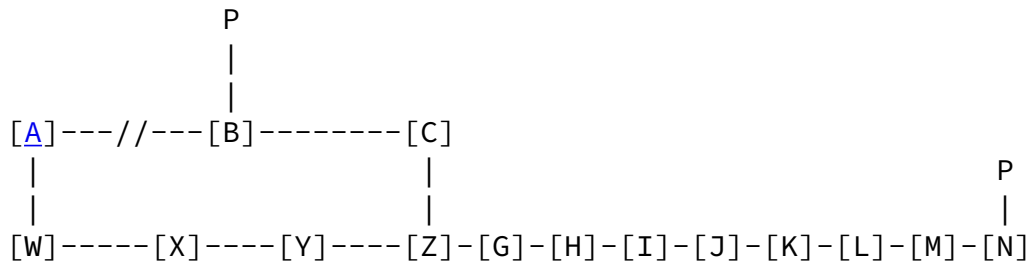


Figure 11 A multi-homed prefix p

Consider a prefix p, which is attached to router B and some other router N as illustrated in Figure 11. Before the failure of the link A-B, p is reachable from A via A-B. After the failure it cannot be assumed that B is still reachable. If traffic to p is assigned to a link repair path to B (as it would be if p were attached only to B), and router B has failed, then it would loop and subsequently be dropped. Traffic for p cannot simply be assigned to whatever repair path would be used for traffic to N, because other routers, which are not yet aware of any failure, may direct the traffic back towards B, since the instance of p attached to B is closer.

A solution is to treat p itself as a neighbor of B, and compute a repair path with p as a target. However, although correct, this solution may be infeasible where there are a very large number of such prefixes, which would result in an unacceptably large computational overhead.

Some simplification is possible where there exist a large number of multi-homed prefixes which all share the same connectivity and metrics. These may be treated as a single router and a single repair path computed for the entire set of prefixes.

An alternative solution is to tunnel the traffic for a multi-homed

prefix to the router N where it is also attached (see Figure 11). If this involves a repair path that was already tunneled, then this requires double encapsulation.

4.6. Equal Cost Path Splits

Equal cost path splits may be used as a repair mechanism, but link and node repairs need to be considered separately.

4.6.1. Equal Cost Path Splits as Link Repair Paths

When a link is used as a member of one or more path-split sets, by definition, the destinations served could be equally well served by any other member of the path-split set. Therefore, when the link fails, any destinations that use the link as a path-split may be immediately assigned to another member of the set. Clearly, if traffic to some destinations can be repaired using a path split, it

should not also be subject to repair by tunneling. Such destinations should be identified before performing traffic assignment to tunneled repair paths.

4.6.2. Equal Cost Path Splits and Node Failure

An equal cost path split may traverse the failed node (router B). In this case, the path split may not be an appropriate repair path. There are two cases: -

- the path split is a parallel link, having router B as a direct neighbor, and
- the path split does not have router B as a direct neighbor, but the route traverses router B at some point further downstream.

These are illustrated in Figure 12 and Figure 13 respectively.

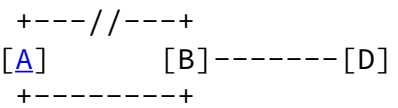


Figure 12 A parallel link path split

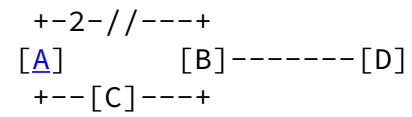


Figure 13 A path split via an intermediate node

In both cases it must be assumed that router B has failed and some other repair path, diverse with respect to router B, must be used.

4.7. LANs and pseudonodes

In link state protocols a LAN is represented by a construct known as a pseudonode in IS-IS and a network LSA in OSPF.

In order to deal correctly with this representation of LANs, the algorithms described in this draft require certain modifications. There are four cases which require consideration. These are described in the following subsections.

4.7.1. The Link between Routers A and B is a LAN

In this case, the link which is being protected is a LAN, and the router B which has potentially failed is reachable over the LAN. This is illustrated in Figure 14.

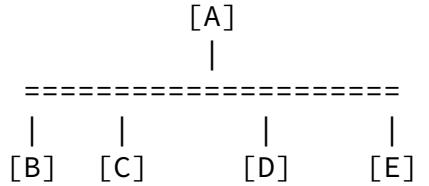


Figure 14 The link between routers A and B is a LAN

There are two possible failure modes in this case.

4.7.1.1. Case 1

Router B or its interface to the LAN may have failed independently of the rest of the LAN. In this case the remaining routers on the LAN (routers C, D and E) will remain reachable from router A. These routers do not appear as direct neighbors of router B in the link state database and are not treated as neighbors of router B for the purposes of this specification because no traffic from router A would be directed through router B to any of these routers. However, each of these neighboring routers will have router B as a neighbor and they will initiate their own repair paths in the event of the failure of router B or its LAN interface.

Repair paths are computed with the non-LAN neighbors of B as targets, and also B itself (the "link-failure" repair path). Note that since the remaining neighbors of A on the LAN are assumed to be still reachable when the link to B has failed, these repair paths may traverse the LAN.

A separate set of repair paths is required in anticipation of the potential failure of each router on the LAN.

4.7.1.2. Case 2

Router A's interface to the LAN may have failed (or the entire LAN may have failed). In either event, simultaneous failures will be observed from router A to all the remaining routers on the LAN (routers B, C, D and E). In this case, the pseudonode itself can be treated as the "adjacent" router (i.e. the router normally referred to as "router B"), and repairs constructed using the normal mechanisms with all the neighbors of the pseudonode (routers B, C, D and E) as repair path targets. If one or more of the routers had failed in addition to the LAN connectivity, treating it as a repair

path target would not be viable, but this would be a case of multiple simultaneous failures which is out of scope of this specification.

The entire sub-tree over A's LAN interface is the failed component and is excised from the spanning tree when computing A's extended P-space. For the Q-spaces of the targets, the sub-tree over the LAN interface of the target is excised.

4.7.1.3. Simplified LAN repair

A simpler alternative strategy is to always consider the LAN and all routers attached to it as failing as a single unit. In this case, a single set of repair paths is computed with targets being the entire set of non-LAN neighbors of all the routers on the LAN, together with "link-repair" paths with all the routers on the LAN as targets. Any failure of one or more LAN adjacencies results in these repair paths being invoked for all neighbors on the LAN. These repair paths must not traverse the LAN, and so must be computed by excising the entire sub-tree reachable over A's LAN interface from A's spanning tree (i.e. the entire LAN is the failed component). The Q-spaces are computed as normal, with the LAN neighbors or their interface to the LAN being excised as appropriate. This is simpler than the approach proposed above, but will fail to make use of possible repair paths (or even path splits) over the LAN. In particular, if the only viable repair paths involve the LAN, it will prevent any repair being possible.

4.7.2. A LAN exists at the release point

When computing the viable release points, it may be that one or more of the leaf nodes are actually pseudonodes. In this case, the release point is deemed to be any of the parent nodes on the LAN by which the pseudonode had been reached, and when computing the extended set of release points (reachable by directed forwarding), all the remaining routers on the LAN may be included.

4.7.3. A LAN between B and its neighbors

If there is a LAN between router B and one or more of B's neighbors (other than router A), then rather than treating each of those neighbors as a separate target to which a repair path must be computed, the pseudonode itself can be treated as a single target for which a repair path can be computed. If there are other neighbors of B which are directly attached to B, including those which may also be attached to the LAN, they must still be treated as an individual repair path target.

Normally a repair path with the pseudonode as its target will have a release point before the pseudonode. However it is possible that the release point would be computed as the pseudonode itself. This will occur if the reverse spanning tree rooted at the pseudonode includes

no routers other than itself. In this case a single repair with the pseudonode as target is not possible, and it is necessary to compute individual repair paths whose target are each of the neighbors of B on the LAN.

4.7.4. The LAN is a Transit Subnet.

This is the most common case, where a LAN is traversed by a repair path, but is not in any of the special positions described above. In this case no special treatment is required, and the normal SPF mechanisms are applicable.

5. Failure Detection and Repair Path Activation

The details of repair path activation are inherently implementation-dependent and must be addressed by individual design specifications. This section describes the implementation independent aspects of the failover to the repair path.

5.1. Failure Detection

The failure detection mechanism must provide timely detection of the failure and activation of the repair paths. The failure detection mechanisms may be media specific (for example loss of light), or may be generic (for example BFD). Multiple detection mechanisms may be used in order to improve detection latency. Note that in the case of a LAN it may be necessary to monitor connectivity to all of the adjacent routers on the LAN.

5.2. Repair Path Activation

The mechanism used by the router to activate the repair path following failure will be implementation specific.

An implementation that is capable of withdrawing the repair may delay the start of network convergence in order to minimize network disruption in the event that the failure was a transient.

5.3. Node Failure Detection Mechanism

When router A detects a failure of the A-B link, it will invoke the link repair path from itself to router B. This A-B link repair is always invoked because even if all other traffic can be re-routed, B

is always a single point of failure to itself. If router B has failed, the A-B link repair can result in a forwarding loop. A node failure detection mechanism is therefore needed. A suitable mechanism might be to run BFD [BFD] between A and B, over the A-B link repair path.

When the node failure detection mechanism has determined that router B has failed it withdraws the A-B link repair path. The node failure detection and revocation of the A-B link repair needs to be expedited, in order to minimize the duration of collateral damage to the network cause by packets looping around the A-B link repair path.

If B is a single point of failure to some destinations, then withdrawing the A-B link repair has no impact on network connectivity, because those destinations will have been rendered unreachable by the failure of router B.

If B is not a single point of failure, but traffic to some destinations is being repaired via the A-B link because of the inability to provide suitable repair paths, then there are destinations that are rendered temporarily unreachable by IPFRR. The IPFRR loop free convergence mechanism delays normal convergence of the network. Consideration therefore has to be given to the relative importance of the traffic being protected and the traffic being black-holed. Depending on the outcome of that consideration, the IPFRR loop-free strategy may need to be abandoned.

6. Loop Free Transition

Once the repair paths have been activated, data will again be forwarded correctly. At this stage only the routers directly adjacent to the failure will be aware of the failure because no routing information concerning the failure has yet been propagated to other routers. The network now has to be transitioned to normal operation using the available components.

During network transition inconsistent state may lead to the formation of micro-loops. During this period, packets may be prevented from reaching the repair path, may expire due to transiting an excessive number of hops, may be subject to excessive delay, and the resultant congestion may disrupt the passage of other packets through the network. The use of a loop free transition technique allows the network to re-converge without packet loss or disruption.

Four loop free transition strategies are described:

- Incremental cost advertisement
- Single Tunnel
- Distributed Tunnels
- Ordered SPF

6.1. Incremental Cost Advertisement

When a link fails, the cost of the link is normally changed from its assigned metric to "infinity". However it can be proved that: if the

link cost is increased in suitable increments, and the network is allowed to stabilize before the next cost increment is advertised, then no micro-loops will form.

This approach has the advantage that it requires no change to the routing protocol, and will work with non-IPFRR capable routers. However the loop-free transition is slow, particularly if large metrics are used, and during this time the network is vulnerable to a second failure.

6.2. Single Tunnel Per Router

When a failure is detected, the routers adjacent to the failure issue a "covert" announcement of the failure, which is propagated through the network by all routers, but which is understood only by IPFRR capable routers. These routers each build a tunnel to the closest IPFRR router adjacent to the failure. They then determine which of their traffic would transit the failure and place that traffic in the tunnel. When all of these tunnels are in place, the failure is then announced as normal. Because the tunnel will be unaffected by the transition, and because the IPFRR router at the tunnel endpoint will continue the repair, no traffic will be disrupted by the failure. When the network has converged, the IPFRR routers can withdraw the tunnels. The order of tunnel insertion and withdrawal is not important, provided the tunnels are all in place before the normal announcement.

This technique has the disadvantage that it requires traffic to be

tunneled during the transition.

A further disadvantage of this method is that it requires co-operation from all the routers within the routing domain to fully protect the network against micro-loops. However it can be shown that micro-loops will be confined to contiguous groups of non-IPFRR capable routers, and will only affect traffic arriving at the network through one of those routers.

6.3. Distributed Tunnels

This is similar to the single tunnel per router approach except that all IPFRR capable routers calculate a set of repair paths using the same algorithms as for traffic that will be affected by the failure.

This reduces the load on the tunnel endpoints, but the length of time taken to calculate the repairs increases the convergence time.

This method suffers from the same disadvantages as the single tunnel method.

6.4. Ordered SPFs

Micro loops occur when a router closer to the failed component revises its routes to take account of the failure before a router which is further away. By analyzing the reverse spanning tree over which traffic is directed to the failed component, it is possible to determine a strict ordering which ensures that routers closer to the root always process the failure after any routers further away, and hence micro loops are prevented.

When the failure has been announced, each router waits a multiple of some time delay value. The multiple is determined by the router's position in the reverse spanning tree, and the delay value is chosen to guarantee that a router can complete its processing within this time. The convergence time may be reduced by employing a signaling mechanism to notify the parent when all the children have completed their processing, and hence when it was safe for the parent to instantiate its new routes.

The property of this approach is therefore that it imposes a delay which is bounded by the network diameter although in most cases it will be much less.

It requires all routers in the domain to operate according to these procedures, and the presence of non co-operating routers can give rise to loops for any traffic which traverses them (not just traffic which is originated through them).

7. Restoring Failed Components to Service

When a neighbor or failed link is restored to service, it will be detected according to the normal operation of the routing protocols by the formation of an adjacency. Normally this would result in the information about the link being included in newly generated routing information. However, just as in the case with increasing costs, the sudden decrease in cost from "infinity" to the configured value of the link cost may give rise to loops. Each of the loop-free transition mechanism described above has a corresponding mechanism that can be used to add a link to the network without the formation of micro-loops.

8. Implications for Network Management

It will be clear from the above that topology changes introduced by management action, such as enabling or disabling a link or router, or changing the cost metric of a link may result in disruption of traffic due to the formation of micro-loops. It will equally be clear that the loop-free convergence strategies described above can equally be applied to the prevention of such micro-loops.

9. IPFRR Capability

In the previous sections it has been assumed that all routers in the network are capable of acting as IPFRR routers, performing such tasks as tunnel termination and directed forwarding. In practice this is unlikely to be the case, partially because of the heterogeneous nature of a practical network, and partially because of the need to progressively deploy such capability. IPFRR therefore needs to support some form of capability announcement, and the algorithms need

to take these capabilities into account when calculating their path repair strategies. For example, the ability of routers to function as tunnel end points and perform directed forwarding will influence the choice of repair path. However, routers which are simply traversed by repair paths (tunneled or not) do not need to be IPFRR capable in order to guarantee correct operation of the repair paths.

10. Enhancements to routing protocols

It will be seen from the above that a number of enhancements to the appropriate routing protocols are needed to support IPFRR. The following possible enhancements have been identified:

- The ability to advertise IPFRR capability
- The ability to advertise tunnel endpoint capability
- The ability to advertise directed forwarding identifiers
- The ability to announce the start of a loop-free transition, and to abort a loop-free transition.
- The ability to signal transition completion status to neighbors.
- The ability to advertise that a link is protected.

Capability advertisement should make use of existing capability mechanisms in the routing protocols. The exact set of enhancements will depend on specific IPFRR design choices.

11. IANA considerations

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

12. Security Considerations

Changes to the IGPs to support IPFRR do not introduce any additional security risks.

The security implications of the increased convergence time due to the loop avoidance strategy depend on the approach to multiple failures. If the presence of multiple failures results in the network aborting the loop free strategy, then the convergence time will be similar to that of a conventional network. On the other hand, an attacker in a position to disrupt part of a network might use this to disrupt the repair of a critical path.

The tunnel endpoints need to be secured to prevent their use as a facility by an attacker. Performance considerations indicate that tunnels cannot be secured by IPsec [IPSEC]. A system of packet address policing, both at the tunnel endpoints and at the edges of the network would prevent an attacker's packet arriving at a tunnel endpoint and would seem to be the best strategy.

When a fast re-route is in progress, there may be an unacceptable increase in traffic load over the repair path. Network operators need to examine the computed repair paths and ensure that they have sufficient capacity.

Acknowledgments

The authors acknowledge the significant technical contributions made to this work by their colleagues: John Harper and Kevin Miles.

IPR Disclosure Acknowledgement

By submitting this Internet-Draft, we certify that any applicable patent or other IPR claims of which we are aware have been disclosed, and any of which we become aware will be disclosed, in accordance with RFC 3668.

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).

IPSEC Kent, S., Atkinson, R., "Security Architecture for the Internet Protocol", RFC 2401

Authors' Addresses

Stewart Bryant
Cisco Systems,
250, Longwater Avenue,
Green Park,
Reading, RG2 6GB,
United Kingdom. Email: stbryant@cisco.com

Clarence Filsfils
Cisco Systems,
De Kleetlaan 6a,
1831 Diegem,
Belgium Email: cfilsfil@cisco.com

Stefano Previdi
Cisco Systems,
Via Del Serafico 200
00142 Roma,
Italy Email: sprevidi@cisco.com

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

Full Copyright statement

INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.