Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 23, 2007

Pierre Francois
Olivier Bonaventure
Universite catholique de Louvain
Mike Shand
Stewart Bryant
Stefano Previdi
Cisco Systems

October 20, 2006

# Loop-free convergence using oFIB
# draft-francois-ordered-fib-02

## Status of this Memo



This Internet-Draft will expire on April 23, 2007.

## Copyright Notice



## Abstract

This draft describes a mechanism for use in conjunction with link state routing protocols which prevents the transient loops which would otherwise occur during topology changes. It does this by correctly sequencing the FIB updates on the routers.

This mechanism can be used in the case of non-urgent link or node shutdowns and restarts or link metric changes. It can also be used in conjunction with a FRR mechanism which converts a sudden link or node failure into a non-urgent topology change. This is possible where a complete repair path is provided for all affected destinations.

After a non-urgent topology change, each router computes a rank that defines the time at which it can safely update its FIB. A method for accelerating this loop-free convergence process by the use of completion messages is also described.

## 1. Introduction

With link-state protocols [1][2], each time the network topology changes, some routers need to modify their Forwarding Information Base (FIB) to take into account the new topology. Each topology change causes a convergence phase. During this phase, routers may transiently have inconsistent FIBs, which may lead to packet loops and losses, even if the reachability of the destinations is not compromised after the topology change. Packet losses and transient loops can also occur in the case of a link down event implied by a maintenance operation, even if this operation is predictable and not urgent. When the link state change is a metric update and when a new link is brought up in the network, there is no direct loss of connectivity, but transient packet loops and loss can still occur.

For example, in Figure 1, if the link between X and Y is shut down by an operator, packets destined to X can loop between R and Y when Y has updated its FIB while R has not yet updated its FIB, and packets destined to Y can loop between X and S if X updates its FIB before S. According to the current behaviour of ISIS and OSPF, this scenario will happen most of the time because X and Y are the first routers to be aware of the failure, so that they will update their FIBs first.

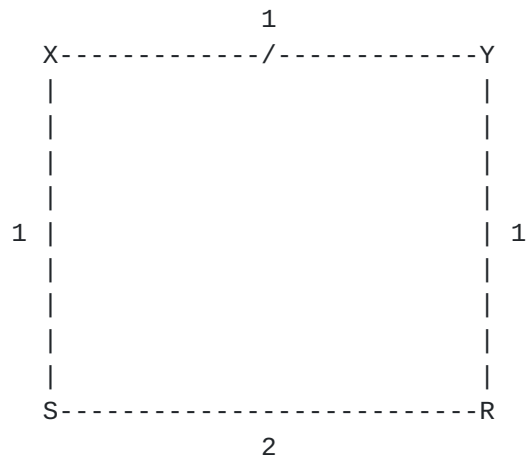


Figure 1: A simple topology

It should be noted that the loops can occur remotely from the failure, not just adjacent to it.

The goal of this draft is to define a mechanism which sequences the router FIB updates to maintain consistency throughout the network. By correctly setting the FIB change order no looping or packet loss can occur. As described in [4] this mechanism may be applied to the case of managed link-state changes, i.e. link metric change, manual link down/up, manual router down/up, and managed state changes of a set of links attached to one router. It may also be applied to the case where one or more network elements are protected by a fast re-route mechanism [3] [7]. The mechanisms that are used in the failure case are exactly the same as those used for managed changes. For simplicity this draft makes no further distinction between managed and unplanned changes.

## 2. The required FIB update order

This section provides an overview of the required ordering of the FIB updates. A more detailed analysis of the rerouting dynamics and correctness proofs of the mechanism can be found in [6].

### 2.1. Single Link Events

For simplicity the correct ordering for single link changes are described first. The draft then builds on this to demonstrate that the same principles can be applied to more complex scenarios such as line card or node changes.

### 2.1.1. Link Down / Metric Increase

First consider the non-urgent failure of a link or the increase of a link metric. In this case, a router R MUST NOT update its FIB until all other routers that send traffic via R and the affected link have first updated their FIBs.

The following argument shows that this rule ensures the correct order of FIB change when the link X->Y is shut down or its metric is increased.

An "outdated" FIB entry for a destination is defined as being a FIB entry that still reflects the shortest path(s) in use before the topology change. Once a packet reaches a router R that has an outdated FIB entry for the packet destination, then, provided the oFIB ordering is respected, the packet will continue to X only traversing routers that also have an outdated FIB entry for the destination. The packet thus reaches X without looping and will be forwarded to Y via X->Y (or in the case of FRR, the X->Y repair path) and hence reach its destination.

Since it can be assumed that the original topology was loop-free, Y will never use the link Y->X to reach the destination and hence the path(s) between Y and the destination are guaranteed to be unaffected by the topology change. It therefore follows that the packet arriving at Y will reach its destination without looping.

Since it can also be assumed that the new topology is loop-free, by definition a packet cannot loop while being forwarded exclusively by routers with an updated FIB entry.

In other words, when the oFIB ordering is respected, if a packet reaches an outdated router, it can never subsequently reach an updated router, and cannot loop because from this point on it will only be forwarded on the consistent path that was used before the event. If it does not reach an outdated router, it will only be forwarded on the loop free path that will be used after the convergence.

According to the proposed ordering, X will be the last router to update its FIB. Once it has updated its FIB, the link X->Y can actually be shut down (or the repair removed).

If the link X-Y is bidirectional a similar process must be run to order the FIB update for destinations using the link in the direction Y->X. As has already been shown, no packet ever traverses the X-Y link in both directions, and hence the operation of the two ordering processes is orthogonal.

### 2.1.2. Link Up / Metric Decrease

In the case of link up events or metric decreases, a router R MUST update its FIB BEFORE all other routers that WILL use R to reach the affected link.

The following argument shows that this rule ensures the correct order of FIB change when the link X->Y is brought into service or its metric is decreased.

Firstly, when a packet reaches a router R that has already updated its FIB, all the routers on the path from R to X will also have updated their FIB, so that the packet will reach X and be forwarded along X->Y, ultimately reaching its destination.

Secondly, a packet cannot loop between routers that have not yet updated their FIB. This proves that no packet can loop.

## 2.2. Multi-link events

The following sections describe the required ordering for single events which may be manifest as multiple link events. For example, the failure of a router may be notified to the rest of the network as the individual failure of all its attached links. The means of identifying the event type from the collection of received link events is described in Section 3.

### 2.2.1. Router Down events

In the case of the non-urgent shut-down of a router, a router R MUST NOT update its FIB until all other routers that send traffic via R and the affected router have first updated their FIBs.

Using a proof similar to that for link failure, it can be shown that no loops will occur if this ordering is respected [6].

### 2.2.2. Router Up events

In the case of a router being brought into service, a router R MUST update its FIB BEFORE all other routers that WILL use R to reach the affected router.

A proof similar to that for link up, shows that no loops will occur if this ordering is respected [6].

### 2.2.3. Linecard Failure/Restoration Events

The failure of a line card involves the failure of a set of links all of which have a single node in common, i.e. the parent router. The ordering to be applied is the same as if it were the failure of the parent router.

In a similar way, the restoration of an entire linecard to service as a single event can be treated as if the parent router were returning to service.

## 3. Deducing the topology change

As has been described, a single event such as the failure or restoration of a single link, single router or a linecard may be notified to the rest of the network as a set of individual link change events. It is necessary to deduce from this collection of link state notifications the type of event that has occurred in the network and hence the required ordering.

There are some events (for example a subsequent failure with conflicting repair requirements occurring before the ordered FIB process has completed) that cannot be correctly processed by this mechanism. In these cases it is necessary to ensure that convergence falls back to the conventional mode of operation (see Section 6).

In all cases it is necessary to wait some hold-down period after receiving the first notification to ensure that all routers have received the complete set of link state notifications associated with the single event.

At any time, if a link change notification is received which would have no effect on the receiving router's FIB, then it may be ignored.

When a link change event is received which impacts the receiving router's FIB, the routers at the near and far end of the link are noted.

If no other event is received during the hold-down time, the event is treated as a link event. Note that the reverse connectivity check means that only the first failure event, or second up event have an effect on the FIB.

If all events received within the hold-down period have a single router (R) in common, then it is assumed that the change reflects an event (line-card or router change) concerning the common router (R).

If an event is received within the hold down period which does NOT reference the common router (R) then in this version of the specification normal convergence is invoked immediately (see Section 6).

In the case of a link change event, the router at the far end of the link is deemed to be the common router (R).

All ordering computations are based on treating the common router R as the root for both link and node events.

## 4. Calculation of the ordering

This section describes how the required ordering is calculated.

### 4.1. Link or Router Down or Metric Increase

To respect the proposed ordering, routers compute a rank that will be used to determine the time at which they are permitted to perform their FIB update. In the case of a failure event rooted at router Y or an increase of the metric of link X->Y, router R computes the reverse Shortest Path Tree in the topology before the failure (rSPT_OLD) rooted at Y. This rSPT gives the shortest paths to reach Y before the failure. The branch of the reverse SPT that is below R corresponds to the set of shortest paths to R that are used by the routers that reach Y via R.

The rank of router R is defined as the depth (in number of hops) of this branch. In the case of ECMP, the maximum depth of the ECMP path set is used.

Router R is required to update its FIB at time

T0 + H + rank * MAX_FIB

where T0 is the arrival time of the link-state packet containing the topology change, H is the hold-down time and MAX_FIB is a network-wide constant that reflects the maximum time required to update a FIB irrespective of the change required. The value of MAX_FIB is network specific and its determination is out of the scope of this document. This value must be agreed by all the routers in the network. This agreement can be performed by using a capability TLV as defined in [8].

All the routers that use R to reach Y will compute a lower rank than R, and hence the correct order will be respected. It should be noted that only the routers that used Y before the event need to compute

their rank.

### 4.2. Link or Router Up or Metric Decrease

In the case of a link or router up event rooted at Y or a link metric decrease affecting link Y->W, a router R must have a rank that is higher than the rank of the routers that it will use to reach Y, according to the rule described in Section 2. The rank of R is thus the number of hops between R and Y in its renewed Shortest Path Tree. When R has multiple equal cost paths to Y, the rank is the length in hops of the longest ECMP path to Y.

Router R is required to update its FIB at time

T0 + H + rank * MAX_FIB

It should be noted that only the routers that use Y after the event have to compute a rank, i.e. only the routers that have Y in their SPT after the link-state change.

## 5. Acceleration of Ordered Convergence

The mechanism described above is conservative, and hence may be relatively slow. The purpose of this section is to describe a method of accelerating the controlled convergence in such a way that ordered loop-free convergence is still guaranteed.

In many cases a router will complete its required FIB changes in a time much shorter than MAX_FIB and in many other cases, a router will not have to perform any FIB changes at all.

This section describes the use of completion messages to speed up the convergence by providing a means for a router to inform those routers waiting for it, that it has completed any required FIB changes. When a router has been advised of completion by all the routers for which it is waiting, it can safely update its own FIB without further delay. In most cases this can result in a sub-second re-convergence time comparable with that of normal convergence.

Routers maintain a waiting list of the neighbours from which a completion message must be received. Upon reception of a completion message from a neighbour, a router removes this neighbour from its waiting list. Once its waiting list becomes empty, the router is allowed to update its FIB immediately even if its ranking timer has not yet expired. Once this is done, the router sends a completion message to the neighbours that are waiting for it to complete. Those routers are listed in a list called the Notification List.

Completion messages contain an identification of the event to which they refer.

Note that, since this is only an optimization, any loss of completion messages will result in the routers waiting their defined ranking time and hence the loop-free properties will be preserved.

## 5.1. Construction of the waiting list and notification list

### 5.1.1. Down events

Consider a link or node down event rooted at router Y or the cost increase of the link X->Y. A router R will compute rSPT_OLD(Y) to determine its rank. When doing this, R also computes the set of neighbors that R uses to reach the failing node or link, and the set of neighbors that are using R to reach the failing node or link. The Notification list of R is equal to the former set and the Waiting list of R is equal to the latter.

Note that R could include all its neighbors except those in the Waiting list in the Notification list, this has no impact on the correctness of the protocol, but would be unnecessarily inefficient.

### 5.1.2. Up Events

Consider a link or node up event rooted at router Y or the cost decrease of the link Y->X. A router R will compute its new SPT (SPT_new(R)). The Waiting list is the set of nexthop routers that R uses to reach Y in SPT_new(R).

In a simple implementation the notification list of R is all the neighbours of R excluding those in the Waiting list. This may be further optimized by computing rSPT_new(Y) to determine those routers that are waiting for R to complete.

## 5.2. Format of Completion Messages

The format of completion messages and means of their delivery is routing protocol dependent and is outside the scope of this document.

The following information is required:-

. Identity of the sender.

. Identity of the set of routing notifications being considered in the associated FIB change.

## 6. Fall back to Conventional Convergence

In circumstances where a router detects that it is dealing with incomplete or inconsistent link state information, or when a further topology event is received before completion of the current ordered FIB update process it may be expedient to abandon the controlled convergence process and revert to conventional convergence by immediately expiring all the associated ranking timers. This mechanism is similar to the one described in section 3.1 of [5].

Abandoning the controlled convergence process may be instigated by any router within the network.

## 7. Acknowledgments

We would like to thank Clarence Filsfils and Jean-Philippe Vasseur for their useful suggestions and comments.

## 8. References

[1] "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless mode Network Service (ISO 8473). ISO/IEC 10589:2002, Second Edition".

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

[3] Shand, M. and S. Bryant, "IP Fast Reroute Framework", draft-ietf-rtgwg-ipfrr-framework-05.txt (work in progress), Oct 2006.

[4] Bryant, S. and M. Shand, "Applicability of Loop-free Convergence", draft-bryant-shand-lf-applicability-02.txt (work in progress), Oct 2006.

[5] Zinin, A., "Analysis and Minimization of Microloops in Link-state Routing Protocols", draft-ietf-rtgwg-microloop-analysis-01.txt (work in progress), Oct 2005.

[6] Francois, P. and O. Bonaventure, "Avoiding transient loops during IGP convergence in IP Networks", in Proceedings of INFOCOM'05, www.info.ucl.ac.be/people/OBO/papers/pfr-infocom05.pdf, March 2005.

[7] Pan, P. and al, "Fast Reroute Extensions to RSVP-TE for LSP

Tunnels", RFC 4090.

[8] Atlas, A., Bryant, S., and M. Shand, "Synchronization of Loop Free Timer Values", draft-atlas-bryant-shand-lf-timers-02.txt (work in progress), Oct 2006.

[9] Francois, P., Filsfils, C., Evans, J., and O. Bonaventure, "Achieving sub-second IGP convergence in large IP Networks", in ACM SIGCOMM Computer Communication Review, http://portal.acm.org/citation.cfm?id=1070873.1070877, July 2005.

## Appendix A. General SRLG Case

This appendix describes the operation of oFIB when multiple link events which DO NOT have a node in common occur at approximately the same time. The covered events are the failure of a set of links and the restoration of a set of links. Note that for the case of a sudden SRLG failure, it is assumed that this is fully protected by a Fast Reroute mechanism, thus converting it into an non-urgent event.

In order to be applicable, this solution requires that routers have the same, consistent, view of the set of events. This can be achieved by means of the hold down mechanism described in Section 3 and [5].

### A.1. SRLG Down Events

#### A.1.1. Determining the ordering

Consider the case where there are two failing components F and G. In the general case, the ranking for any given router R will be different for destinations reached through F and those reached through G. R must therefore partition its FIB changes into a number of destination sets. In the worst-case, the number of destination sets will equal the number of failing links.

Router R computes the ranks associated with each of the failing links. It does this by applying the same algorithm as for single link down events. The rank at which a router R must update its FIB for a destination D is equal to the minimum rank among the ranks of the links that R uses to reach the destination D.

#### A.1.2. Completion messages

As described above, a router R computes the Waiting and Notification lists associated with each of the failing links when it determines

the ranking.

When R has received a completion message from all the members of the waiting list associated with a link, it is allowed to update its FIB for all the destinations that it was previously reaching via that link.

A router will send a completion message to the members of the Notification list for a given link once it has updated its FIB for all the prefixes that it reached via the link.

### A.2. SRLG Up Events

#### A.2.1. Determining the ordering

Consider the case where a set of links is brought up in the network. R computes the rank associated with each link, by the means of its renewed SPT. The rank at which R must update its FIB for a destination D is the maximum rank among the ranks of the links that it will use to reach D.

#### A.2.2. Completion messages

As described above, a router R will compute the Waiting List and Notification List associated with each of the links that come up in the network.

When R has received completion messages for the links that it will use to reach a destination D, it can safely update its FIB for D.

When R has updated its FIB for all the destinations that it reaches via a link, it will send a completion message for this link towards the neighbors that are not in its Waiting List for this link.

## Authors' Addresses

Pierre Francois
Universite catholique de Louvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348
BE

Email: francois@info.ucl.ac.be

Olivier Bonaventure
Universite catholique de Louvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348
BE

Email: bonaventure@info.ucl.ac.be

Mike Shand
Cisco Systems
Green Park, 250, Longwater Avenue,
Reading RG2 6GB
UK

Email: mshand@cisco.com

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

Email: sbryant@cisco.com

Stefano Previdi
Cisco Systems
Via Del Serafico 200
00142 Roma
Italy

Email: sprevidi@cisco.com

## Full Copyright Statement



## Intellectual Property



## Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).