Network Working Group
Internet Draft

Expiration Date: Januray 2005

Naiming Shen
Redback Networks
Ping Pan
Ciena Corp

July 2004

Nexthop Fast ReRoute for IP and MPLS

<draft-shen-nhop-fastreroute-01.txt>

Status of this Memo

Abstract

This document describes a mechanism called NFRR (Nexthop Fast ReRoute) to perform fast re-route for any type of traffic in the event of a link/node failure or a nexthop unreachable.

The protected traffic can be IP, MPLS, unicast or multicast. The re-routed traffic can either be destined to the nexthop router or to the next-nexthop router. RSVP explicitly routed LSPs are used as a tool to perform the local patch for minimizing the packet loss.

1. Introduction

This document describes a simple mechanism to quickly re-direct IP

and/or MPLS traffic away from a local link or a nexthop failure. The mechanism presented is mainly to facilitate the needs of real-time IP applications over native IP unicast/multicast networks or LDP based MPLS networks. The goal is to limit the IP packet loss duration in the network to 10s of milliseconds in the event of link/node failures. RSVP [1] signaled LSP is used with explicitly routed path as the re-direct tunnel, while the protected traffic can be either MPLS traffic engineered LSPs, LDP based LSPs, IP unicast, IP multicast traffic or the mix of them. This mechanism can be applied to both point-to-point links and multi-access links in the cases of the link protection and node protection.

An optional RSVP Bypass Nexthop object is defined to allow a modified RPF checks for re-directed IP multicast data traffic. It can also be used to detect misconfigured re-direct LSPs.

The node failure fast protection of native IP traffic is also described in this document. Link State IGP can be used to make the IP prefixes association with Next-NextHop nodes and the re-direct LSPs configured towards them. For node protection of LDP and multicast IP traffic, extension for both protocols are needed and is beyond the scope of this document.

The re-direct LSP is no different from any other RSVP explicitly routed LSPs, except that it does not carry data traffic under normal condition. It is pre-built to reach the nexthop or the next-nexthop router in the case of the protected link/node fails or the nexthop over that protected link is unreachable. Any type of data traffic intended to use this nexthop may be switched onto this re-direct LSP. Since the LSP is built to protect local links or adjacent nodes, the explicit path can be easily calculated either statically or dynamically.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [5].

LSP - An MPLS Label Switched Path. In this document, an LSP will always refer to an explicitly routed LSP.

LDP LSP - LSP signaled by LDP protocol.

RSVP LSP - LSP signaled by RSVP protocol.

NFRR - Nexthop Fast ReRoute Scheme

NFRR LSP - Nexthop Fast ReRoute LSP, which is same as bypass
LSP or re-direct LSP in this document.

PLR - Point of Local Repair. The head-end LSR of a bypass LSP.

MP - Merge Point. The tail-end LSR of a bypass LSP.

Nexthop Node - The router is directly connected to the PLR node.

Next-Nexthop Node - The router is the Nexthop node of PLR's Nexthop node.

3. Motivations

IP applications such as VoIP and PWE3 are highly desirable to have the packet loss with less than 10s of milliseconds during network elements failure. Currently there are three approaches in practice or proposed to speed up the recovery from such failures as discussed below.

First, MPLS LSP Fast ReRoute mechanism [2] is used to quickly re-route the RSVP LSP traffic onto a detour or bypass LSP when a local link failure is detected. Since the detour or bypass LSPs are pre-built before the local link failure, this re-route operation can be accomplished within 10s of milliseconds. If the IP backbone deploys network-wide MPLS TE, this MPLS Fast ReRoute approach is the best solution. The Fast ReRoute is just another application using the existing MPLS infrastructure. This mechanism does not offer protection of IP, LDP or multicast data traffic.

Second, IGP fast convergence is another mechanism in reducing the packet loss time in network element failure. This mechanism also includes the improvement of LDP convergence. Comparing with the first solution, the recovery time is usually an order of magnitude higher, which is in the 100s of milliseconds range. For certain real-time applications, that duration is still acceptable and this is a big improvement over "normal" IGP convergence time of seconds or even 10s of seconds. Since this mechanism is purely based on the control plane convergence in the network, there is no guarantee for the convergence to be limited under 100s of milliseconds.

Third, pre-calculated alternative nexthops are downloaded into forwarding engines. As in the first mechanism, when a local link failure is detected, those alternative nexthops are used to continue forwarding the data traffic. If an alternative nexthop does exist, then the re-route time can be accomplished within 10s of milliseconds. There are a couple of shortcoming of this approach. There may not exist such an alternative nexthop for the IP destinations along with the links it intends to protect. When such alternative nexthops do exist, if there are many IGP interfaces and adjacencies on the node, this requires to run

many instances of SPF in order to find a loop-free alternative. This scheme can not be used to protect MPLS TE LSPs since they are not constructed from the native IP routing. Last, if the local link failure is shortly after some network events and the IGP on the node is busy calculating those SPFs, then the alternative nexthop picture is incomplete at that time and the re-route action may not be reliable.

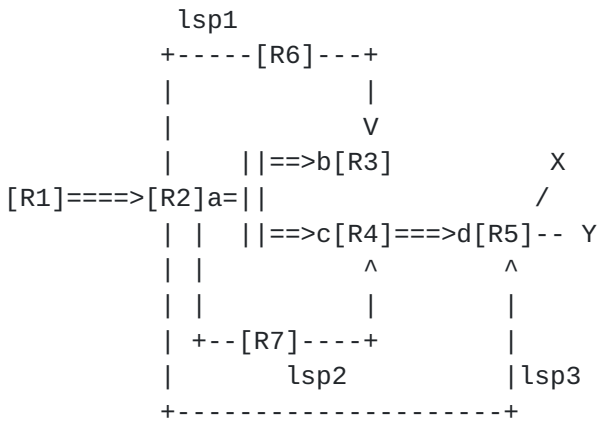
There is obviously a need for providers to protect not only the RSVP based LSP traffic, but also any data traffic in general; there is a need for providers not only to protect traffic in the case of link failure, but also with node failure for any type of traffic; there is a need for providers not only to protect native IP unicast traffic, but also the IP multicast data traffic. The Nexthop Fast ReRoute mechanism does not make any assumption of the protected traffic is MPLS tunneled or not. If a network-wide MPLS traffic engineering is not the goal of the network design, the network can either stay with native IP or LDP LSPs but still get the reliable Fast ReRoute benefits for link/node protection. In this scheme, RSVP signaled LSP will be used as the re-direct tunnel for protected links/nodes. Those LSPs are explicitly routed to get around the intended failure points. In this scheme, the LSP is used in the network as a tool to fast re-direct data traffic. If a network has an external Frame Relay circuit in replacement for the RSVP based LSP, this Nexthop Fast ReRoute will also work as specified in this document.

4. Nexthop Fast ReRoute(NFRR) Operation Scheme

Over a point-to-point there will be only one nexthop in general. There are multiple nexthops over a LAN interface. The NFRR is used to re-route traffic around the IP nexthop over the protected interface. Assume there is an alternative path to reach that nexthop node other than the protected link, we can always pre-build an explicitly routed LSP to the node which owns this nexthop address. If the local link is down, or the nexthop is unreachable, the PLR router can quickly re-direct the data traffic intended for this nexthop onto the NFRR LSP for that nexthop node. Thus any traffic can be fast re-routed in a loop-free fashion using NFRR. Over a LAN, the nexthop unreachable status can possibly be quickly detected using some link level aliveness protocols, and this is beyond the scope of this document. When using NFRR protecting MPLS traffic, global label space scheme is assume on the MP node. The NFRR for link

protection is assumed in this section. Node protection is described in section 5.

The below network diagram will be used in this document as an example topology for discussion purpose:



R2 is the PLR router.
R2, R3 and R4 share a LAN connection with "a", "b" and "c" belong to the same prefix.
R5 is a Next-Nexthop node from the PLR through R4.

lsp1 is sourced from R2 and explicitly route through R6 to reach R3. lsp1 is configured to protect "b" over interface "a".

lsp2 is sourced from R2 to reach R4. lsp2 is configured to protect "c" over interface "a".

lsp3 is sourced from R2 to reach R5. lsp3 is configured to protect "c" or node R4 over interface "a".

Figure 1: An example of NFRR

4.1 NFRR to protect LDP LSPs

When the protected traffic is LDP LSPs, LDP process can pre-build the association of the NFRR tunnel and the LDP LSPs which use this particular nexthop. When the link is down or the nexthop is unreachable, the forwarding engine can quickly switch the traffic onto that NFRR tunnel by pushing an outbound label and send it out. This is no different from re-directing MPLS TE LSPs as described section 4.5 only without any RSVP signaling involved in the LDP case.

4.2 NFRR to protect IP unicast traffic

The PLR node can pre-build the association of the NFRR tunnel and the IP prefixes which use that same nexthop in the route lookup. When the nexthop failure is detected, the forwarding engine will be able to re-route the IP traffic to those affected destinations onto the NFRR tunnel. The only difference from the LDP case is that, it only has one label on the label stack of the packet when being switch out to the NFRR LSP.

4.3 NFRR to protect IP multicast traffic

Similar to the IP unicast case, the PLR node can pre-build the association of the NFRR tunnel and the (S, G) entries on the protected interface. In the point-to-point case, there needs to be only one NFRR tunnel to be referenced in the (S, G) entries of the protected interface. In the LAN case, multiple NFRR tunnel references can exist in the (S, G) entries. When the protected interface is down, or one of the multicast forwarding downstream neighbor is unreachable, all or part of the NFRR tunnels can be applied to re-route multicast traffic to the downstream nodes. In the example in Figure 1, assume R2 has both R3 and R4 as downstream for some (S, G)s, when the link "a" is down, the references in those (S, G)s should point to both lsp1 and lsp2 as its "new" downstream interfaces. R2 needs to forward the multicast traffic through both LSPs.

This document defines a new RSVP Bypass Nexthop object which can be optionally inserted into the PATH message by the head-end of the NFRR LSP and the RESV message by the tail-end of the NFRR LSP. The bypassed link nexthop IP address of the NFRR tunnel can be conveyed to the tail-end node using this new RSVP object. Multicast RPF check algorithm can be modified to accept the multicast traffic for the (S, G)s on the alternative inbound interface even though the RPF check may currently point to the protected link which has that nexthop IP address.

4.4 NFRR to protect IPv6 traffic

The same NFRR LSP can protect native IPv6 traffic going to the same neighbor node over the protected interface. In this case, an IPv6 nexthop address can be configured along with the NFRR LSP. The same operation for unicast and multicast of IPv4 traffic mentioned above applies here.

4.5 NFRR to protect MPLS TE LSPs

When some of the protected traffic to the nexthop belongs to MPLS TE LSPs, the mechanism is the same as described in [2] as facility based link-protection bypass tunnel scheme. All the RSVP signaling extension described in that document

applies here. NFRR only explicitly extends this into the protection of a nexthop to deal with multi-access case instead of protection of local link only, but The technique used is identical.

4.6 Protocol packets over the protected link

There are two types of protocol packets with regard to this scheme. One requires an IP route lookup such as BGP, OSPF VL or RSVP packets;

The other is sent directly over a local interface to neighbors such as ISIS, OSPF, PIM or LDP adjacency packets. When the protected link is down or the protected nexthop is unreachable, the affected routable protocol packets MUST be re-routed over the NFRR tunnel while the directly transported protocol packets SHOULD be dropped in order to time out the protocol adjacency (even if those protocol packets are re-directed over the NFRR LSPs, they will be dropped by the neighbors due to inbound interface does not match the protocol packets). The link/node down event will be eventually propagated across the network and the entire network can converge into a new topology.

5. Node Protection

The NFRR scheme described in section 4 is for link and nexthop failure protection. This section describe the technique to use NFRR for node protection.

5.1 Node Protection for IP Unicast Traffic

As described in section 4, we make the association from the route nexthop to the NFRR LSP. When the link or nexthop fails, the forwarding engine switches the traffic using this route nexthop onto the NFRR LSP. In the node protection case, as long as we have the knowledge of which routes using this nexthop also going to the Next-Nexthop node, we are able to make the same association to re-direct the traffic onto the NFRR LSP which has the Next-Nexthop node as the tail-end.

For IP unicast traffic, this knowledge of routes association with nexthop and Next-Nexthop nodes can be easily obtained from link state IGPs. IGP Shortcut [4] is a technique to dynamically direct IP traffic through TE LSPs. In the NFRR node protection case, the PLR can use those shortcuts not for normal IP traffic, it will only be used when the nexthop element fails. In other words, this shortcut to the Next-Nexthop node is suddenly enabled by forwarding engine when it detects the link, nexthop or nexthop node failure. Those IGP shortcut LSPs can also be called conditional IGP shortcuts with the conditions being the nexthop link or node failure.

In the example of Figure 1, lsp3 is a NFRR LSP source from PLR R2

with destination of R5 to protect node R4 as well as link "a" and nexthop "c". IGP uses modified shortcut technique to associated prefixes X and Y with nexthop "c1" over interface "a". The IGP also installs a modified shortcut lsp3 to be associated with nexthop "c1". The nexthop "c1" is just like "c", but it contains the NFRR LSP lsp3 reference information. Basically if R4 has N nexthop nodes, R2 will have "c1" through "cn" nexthops, each references a NFRR LSP to it's Next-Nexthop node. The IP traffic to X and Y normally is forwarded to R4, in the event of "a", "c" or R4 node failure, the traffic is re-directed to lsp3 into R5.

The algorithm for IGP Shortcut described in section 4 of [4] has three possible ways to determine the first-hop information. The first way can be modified for NFRR conditional shortcut as follows:

- Examine the list of tail-end routers directly reachable via a NFRR-tunnel. If there is a NFRR-tunnel to this node, we will copy the first-hop information from the parent node(s) to the this node. We also attach the NFRR-tunnel information to the first-hop information on this node.

This first-hop information of the node can be used to construct the nexthop and its association with the NFRR LSP destined to that node. The rest of the NFRR operation is the same as in link protection case as described in section 4.2.

5.2 Node Protection for Other Traffic

NFRR Node Protection for MPLS TE LSP is the same as described in [2] as facility based node-protection bypass tunnel scheme. NFRR only explicitly extends this capability into the multi-access case. The technique used is the same.

NFRR Node Protection for LDP LSP and IP multicast traffic will be covered by two separate documents using the NFRR technique and is out side the scope of this document.

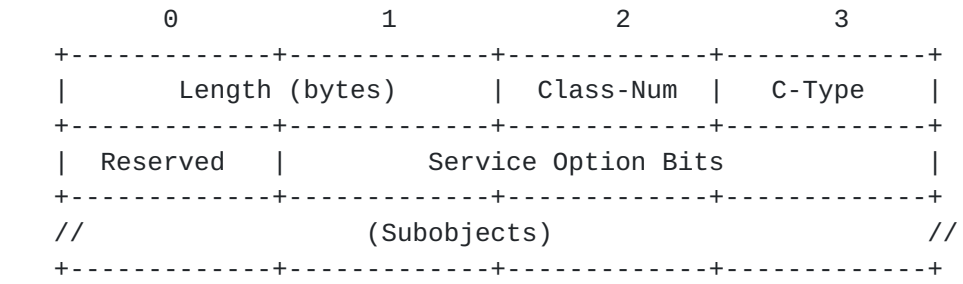
6. RSVP Bypass Nexthop Object

A NFRR LSP is just like any explicitly routed LSP defined in [1] and it is used by the PLR to fast re-route traffic to the same neighbor over alternative interfaces. As mentioned in section 4.3, re-routed multicast traffic will be dropped if the neighbor does not aware certain multicast traffic can come in an alternative interface. This Bypass Nexthop object is used to dynamically pass this information from the head-end NFRR node to the tail-end node so that the RPF check on that protected interface can be modified to accept with an alternative interface. The tail-end node can send back the same object to indicate whether the requested operation is supported.

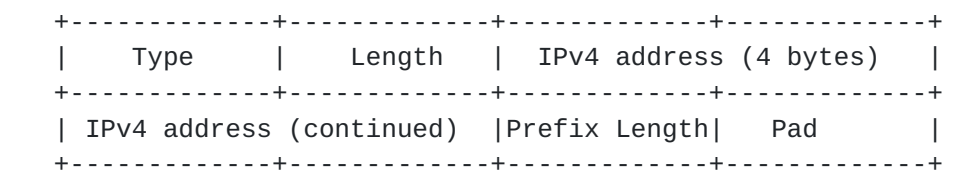
If NFRR LSP is used to protect the link failure, it is useful to know if the NFRR tail-end owns this bypassed nexthop address; When in node protection case, it is also useful to know the NFRR tail-end does not own this bypassed nexthop address; For IP multicast re-route application, the bypassed nexthop address needs to be local to the tail-end node in both link and node protections. This object can be inserted by the tail-end node in RESV message to confirm the acknowledged address is local or not to prevent NFRR LSP mis-configuration.

The Bypass Nexthop object has the following format:

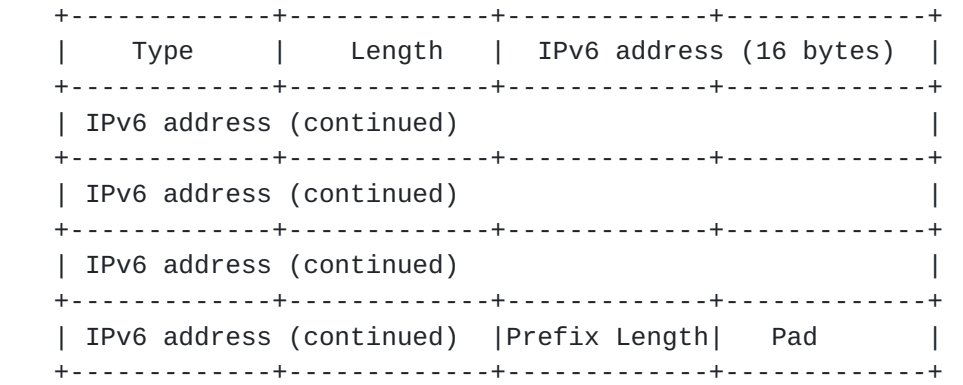
Class = TBD (use form 11bbbbbb for compatibility)
C-Type = 1 or 2



Subobject 1: Nexthop IPv4 address



Subobject 2: Nexthop IPv6 address



This object with C-type of 1 is used in PATH message and C-type of 2 is used in RESV message. They are called Request and Acknowledgement Objects. The Request Bypass Nexthop object MUST only be inserted into PATH message by the head-end of the NFRR LSP node , and MUST not be changed by downstream LSRs.

The Ack Bypass Nexthop object MUST only be inserted into RESV message by the tail-end of the NFRR LSP node, and MUST not be changed by the upstream LSRs.

Only two bits are currently defined for Service Option Bits field in this document as following (position from the right most to left most):

Bits	Description
1	Request/Ack this bypass nexthop address to be local to the NFRR LSP tail-end node. The head-end LSR sets this bit in PATH message requesting the information; the tail-end LSR responds with RESV message by setting or clearing this bit depends on the bypass nexthop address is local to the LSR or not.
2	Request/Ack this bypass nexthop address to be used to support modified multicast RPF checks as defined in section 4.3. If the tail-end LSR does not support this extension, then the multicast data traffic SHOULD NOT be switched over to the NFRR LSP in the event of link/node failure.

7. Forwarding Entry Update and NFRR LSP Reversion

The link down or nexthop unreachable event will eventually reach the protocols such as OSPF or LDP. Regardless of IGP fast convergence is used or not, the new forwarding entry downloading should be hold for a little longer than the expected network convergence time. This is to guarantee all the nodes in the routing area have converged onto the new topology to avoid the possibility of forwarding loops. It is safe to send traffic over the NFRR LSP even after the network is converged. Since before the link failure, the PLR was using the nexthop node to reach some IP destination; it will be highly unlikely that the nexthop node sends the traffic back to the PLR after the adjacent link/node fails in network steady state.

Since the NFRR scheme is nexthop entry based, when those entries are updated after the NFRR re-route takes place, the re-route action for those entries will be reverted to normal operation. In the example in Figure 1, R2 used to reach prefix X and Y through nexthop "c" of R4. When "a" or "c" is down, the traffic towards X and Y will be tunneled in lsp2 to use the same R4 for further forwarding. After the holdown time expires and R2 decides to use R7 as the new IP nexthop for X and Y. When R2 downloads the X and Y into the forwarding engine, this update will revert the re-route operation for traffic to X and Y. If the link "a" comes up before the holdown time expires, R2 will use "c" as the nexthop

for X and Y again. The reversion time and operation is the same in both cases.

Fast convergence of IGP will improve the network performance even with the NFRR presents. It can help to quickly reach the more optimal forwarding state in the routing domain with topology changes.

8. Operational Considerations

8.1 ECMP Cases

The NFRR scheme is independent of ECMP case and the loadsharing algorithm should be the same. THe NFRR LSP is used to protect one particular nexthop, only the portion of traffic used to use this nexthop will be re-routed in the failure event.

8.2 Bandwidth Reservation

Even in the case the network does not use MPLS TE for normal traffic, bandwidth reservation for NFRR LSPs can still be applied. The RSVP interface bandwidth will reflect the amount of link bandwidth reserved for re-routed traffic purpose.

8.3 Type of Traffic To Be Re-routed.

Since NFRR can be applied to any traffic in link protection case, it is an implementation or configuration issue to decide which type of traffic will be applied, others will be dropped. Even within the same type of traffic, filters can be designed to select only the traffic using certain destination, service or labels will be re-routed if the bandwidth is an issue.

8.4 MPLS and IP In The Network

The NFRR scheme uses RSVP explicitly routed LSP to protect data traffic while data traffic itself may not use MPLS TE LSPs in the network. In this case, MPLS TE is not the goal of the network design, the NFRR LSPs are used as a tool to accomplish the fast re-route goal. In order for the re-routed traffic to be reliable and loop-free for any network topology, the traffic has to be either source routed or tunneled independent of IP routing. RSVP signaled LSP is widely supported technology and can be easily fit into NFRR application. If the network only needs to protect a few local links or nodes, the RSVP LSPs can be restricted to a limited scope in the network using NFRR. It is also useful for the network which has MPLS TE in the core and IP or LDP LSPs at the edge.

Even in the case the provider has outband circuits to protect the link or node, NFRR can also be used without RSVP signaled LSP involved. The multicast RPF extension for RSVP functionality can be statically applied on the MP node in this case.

8.5 Link Protection and Node Protection

Node protection in fast re-route is not without issues. Not all the LSRs have the node-protection option. For example, the PHP nodes can only perform link-protection for the last hop of LSPs.

The node-protection scheme also takes more network resource since the MP is further away from the nexthop and it requires more signaling work to identify the flow going through the next-nexthop node in IP, LDP and multicast cases.

With Non-stop forwarding, protocol graceful restart and software modular design make good inroad into provider's networks, a complete node failure will gradually become rare events and a node down often can be scheduled.

9. IANA Considerations

The NFRR proposal requires that IANA allocate a C-class number for Bypass Nexthop object.

10. Security Considerations

This document does not introduce new security issues. The security considerations pertaining to the original RSVP protocol [3] remain relevant.

11. Acknowledgments

The authors would like to thank George Apostolopoulos, Enke Chen, Albert Tian, Liming Wei and Jun Zhang for their contributions and comments to this document.

12. References

[1] D. Awduche, et al, "RSVP-TE: Extensions to RSVP for LSP tunnels", RFC3029, December 2001.

[2] Pan, P., Gan, D., Swallow, G., Vasseur, J.Ph., Copper, D., Atlas, A., Jork, M., "Fast Reroute Technique in RSVP-TE", Interne draft, draft-pan-rsvp-fastreroute-06.txt, work in

progress.

[3] R. Braden, Ed., et al, "Resource ReSerVation protocol (RSVP) -- version 1 functional specification," RFC2205, September 1997.

[4] Shen, N., Smit, H.,"Calculating IGP Routes Over Traffic Engineering Tunnels", draft-ietf-rtgwg-igp-shortcut-01.txt, Work In Progress.

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

13. Author Information

Naiming Shen
Redback Networks, Inc.
300 Holger Way
San Jose, CA 95134
Email: naiming@redback.com

Ping Pan
CIENA Corp.
5965 Silver Creek Valley Road
San Jose, CA 95138
Email: ppan@ciena.com

IPR Notice

Director.

Full Copyright Notice

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.