

The Bi-Mesh Survivable Network – Fast Recovery from Node Failure in MPLS enabled Networks

Dr. Bruce Northcote
Teletraffic Research Centre
University of Adelaide
Adelaide, Australia
bruce.northcote@adelaide.edu.au

Abstract—Today’s converged networks are increasingly based on MPLS technology. To provide reliable carrier-grade service, MPLS-based recovery procedures must be able to quickly address any link or node failure. Current standardized MPLS fast recovery mechanisms cannot cope with egress node failure. We present a dual plane core network architecture, a path naming convention and a simple recovery mechanism that together can remove that fundamental limitation.

Keywords-MPLS; fast re-route; node failure; survivability.

I. INTRODUCTION

Multi-Protocol Label Switching (MPLS) [1] [2] is one of the preferred models for providing carriage of data across high-speed, wide area networks. In MPLS enabled networks, incoming datagrams are assigned a label by an edge device, typically a Label Switch Router (LSR) in an Internet Protocol (IP) network. The datagrams are then forwarded along Label Switched Paths (LSPs) to the network egress router. MPLS provides significant potential advantages over simple IP based routing in wide area networks through improved network management, traffic engineering and failure recovery as compared with IP routing, and intrinsic support for Virtual Private Networks (VPNs) at the cost of some increased operational overhead.

To achieve redundancy in carrier-grade wide area networks, core network elements are often deployed as mated pairs. An extension of the mated pairs deployment is the “dual plane” architecture. In this architecture, traffic in regional centres is aggregated at points of presence (PoPs) comprising pairs of high-speed “core” routers, before being transported to other regional centres. Preferably, directly connected PoPs will be interconnected via pairs of physically diverse links to provide for maximum reliability. In the dual plane architecture, the core routers within a PoP, labelled A and B for example, will be connected to each other, and to the correspondingly named core routers in adjacent PoPs.

When a component (link or LSR) fails in an MPLS enabled network, LSPs traversing that component also fail. In those situations the LSP headend (the LSR at the beginning of the LSP) determines a new route for the LSP through the network. While this approach can help to ensure best use of network resources, time to detect and report the failure to the LSP headend can result in significant delays in establishing a new

path, with consequent disruption to traffic being carried on the failed LSP.

Consequently the IETF has defined the MPLS Traffic Engineering Fast Re-Route (FRR) function [3] [4]. Fast Re-Route diverts traffic around a failed component along a pre-assigned backup LSP that intersects the original LSP at some point downstream. This reduces the time to recovery down to that of the failure detection times, typically of the order of 50ms or less, thus potentially significantly decreasing path down-time and thus improving network reliability.

Reference [3] describes a framework for MPLS-based recovery schemes, and identifies the necessary components of those schemes with respect to fault detection, fault notification and fault recovery. In that context, FRR can be considered to be a protection switching mechanism with local scope that can operate in revertive mode. The LSR immediately upstream of the fault initiates the recovery and becomes the Point of Repair (PoR), switching protected traffic on to the FRR LSP for the duration of the failure, then automatically allowing the traffic to revert to the original LSP if/when the failure has cleared.

A fundamental limitation of FRR as standardized is that it cannot protect from LSR failure whenever the PoR is the penultimate hop on the LSP [4]. This is because it is not possible to construct a FRR LSP that bypasses the failed LSR and re-joins the LSP downstream. As a result, an upstream core LSR would have to make routing decisions for all protected traffic packets after reconvergence of the routing protocol has determined an alternate path. This can take in the order of several seconds (for interior gateway protocols) or minutes (for exterior gateway protocols such as Border Gateway Protocol (BGP)).

In this paper we apply a novel LSP labelling scheme to a dual-plane MPLS architecture to derive the Bi-Mesh Survivable Network (BSN), which overcomes this limitation of the conventional use of Fast Re-Route. Thus all MPLS-enabled link and node failure re-configuration times can be reduced to FRR detection times, namely of the order of 50ms, thereby helping to improve network reliability.

This paper is structured as follows:

Section 2: Describes dual plane architectures in Wide Area Networking.

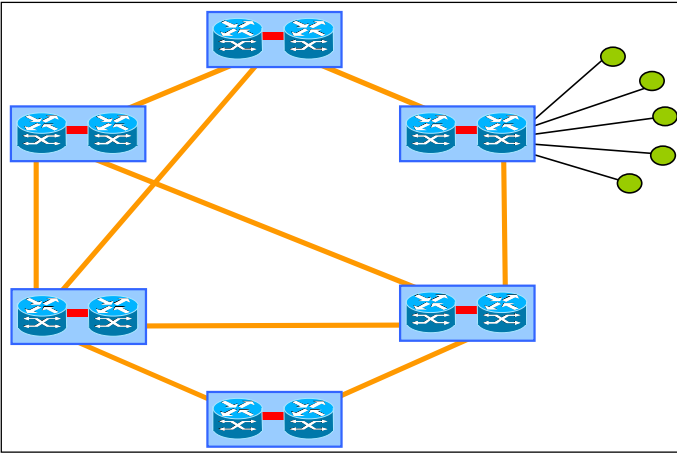


Figure 1. Example Carrier Wide Area Network showing physical connectivity between Points of Presence and intra-regional routers.

Section 3: Describes Fast Re-Route (FRR) in MPLS enabled networks and a shortcoming in its response to failure of egress nodes on MPLS paths.

Section 4: Describes the mechanism for MPLS-enabled fast recovery in the Bi-mesh Survivable Network that addresses the shortcoming discussed in Section 3.

II. MPLS DUAL PLANE ARCHITECTURE

Our basic network model assumes that MPLS is deployed ubiquitously throughout the core network, with physical link connectivity between Points of Presence (PoP). A theoretical six node example is shown in Figure 1. The dual plane architecture assumes that connectivity between PoPs comprises a pair of physically diverse links, and that each PoP is comprised of a pair of directly connected core-network LSRs. These assumptions ensure path diversity between every pair of PoPs.

We can emulate direct link connectivity throughout the network by establishing MPLS LSPs between any pair of PoP LSRs that are not directly connected by a physical link. Doing so allows simplification of the routing tables, because each core destination becomes reachable via a single logical hop from any core origin. That single hop will be either a physical link, or an LSP. We further assume that traffic that needs to be routed within a region does *not* pass to the core network routers.

Therefore, we are able to abstract the geographic and link-specific physical implementation details of the core network architecture to obtain an arbitrary logical connectivity topology (reflected in the routing tables of the core nodes) between PoPs.

We exploit this freedom by electing to create two virtual meshes of LSPs. This is shown in Figure 2.

We denote the mated LSRs within a region as “A” or “B” routers as appropriate. We can establish a fully connected inter-PoP mesh network of LSPs between the “A” core LSRs (the “solid” mesh), and another between the “B” core LSRs (the

“dashed” mesh), with the two meshes connected only via LSPs between the mated “A” and “B” LSRs within a region.

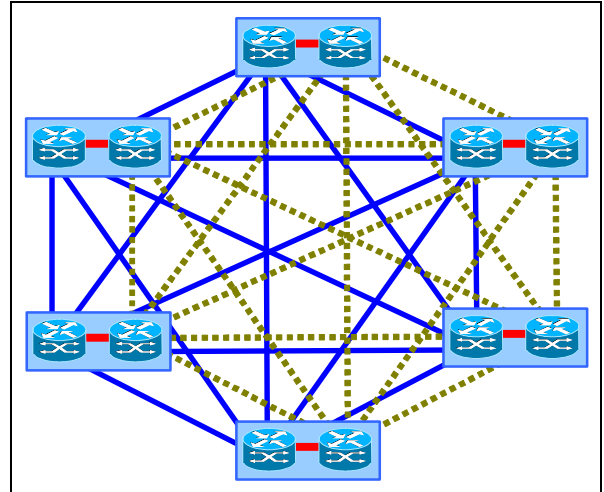


Figure 2. Redundant core network LSP topology

It is possible to construct the dual plane architecture such that the “A” LSRs only have inter-PoP links to other “A” LSRs, and similarly for the “B” LSRs. Note that due to our assumption of path diversity this then ensures that none of the paths in either virtual mesh need ever utilize the direct link between LSRs within the same PoP. Those intra-PoP links are then 100% free to be used to carry protected traffic during failure scenarios.

A major advantage of the MPLS mesh architecture is the simplicity of its routing tables, as each core-ingress PoP can reach each core-egress PoP in a single virtual hop.

III. FAST RE-ROUTE OPERATION AND A SHORTCOMING

A. Fast Re-Route Background

The following description of Fast Re-Route is derived primarily from [6].

Fast Reroute enables all traffic carried by LSPs that traverse a failed network component to be rerouted around the failure along pre-established backup FRR LSPs. The reroute decision is completely controlled locally by the LSR immediately upstream of the failure (the PoR). The headend of the LSP is also notified of the failure through the appropriate routing protocol (BGP or RSVP) and the headend may then attempt to establish a new LSPs that bypass the failure.

Local reroute is intended to prevent any further packet loss caused by the failure, subject to encountering congestion on the backup paths. This gives the headend of the LSP time to re-establish each failed LSP along a new, optimal route. If the headend still cannot find another path to take, it will continue using the appropriate backup LSP.

The example in Figure 3 (adapted from [6]) illustrates how Fast Reroute link protection is used to protect traffic carried on an LSP between devices R1 and R4, as it traverses the mid-point link between devices R2 and R3. The LSP from R1 to R4 is considered to be the primary LSP and is defined by labels

37, 14, and employs penultimate hop popping (PHP) [2]. To protect that R2-R3 link, a backup LSP is created that runs from

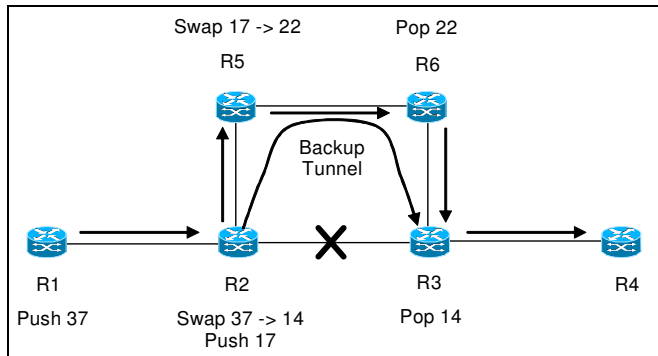


Figure 3. Backup LSP – Fast Re-route

R2 to R3 by way of R5 and R6. This backup FRR LSP is defined by labels 17, and 22 with PHP.

When R2 is notified that the link between it and R3 is no longer available, it simply forwards traffic destined for R3 through the backup LSP. That operation implements label stacking [2] by pushing label 17 onto packets destined to R3 after the normal swap operation (which replaces label 37 with label 14) has been performed. Pushing label 17 onto packets forwards them along the backup LSP, thereby switching traffic around the failed link. The decision to switch packets from the primary LSP to the backup LSP is made solely by R2 upon detection of link failure.

The Fast Re-Route mechanism can also deal with failures of routers on an LSP provided, as noted in [4], that the failed router is not the *egress router* of the LSP. With that exception, FRR LSPs can be established that detour around the next LSR in the path, rejoining the LSP further downstream.

To emphasise the point, Fast Re-Route *does not* provide a mechanism dealing with failure of the egress router of an LSP and such failures must generally be handled using IP layer route failure detection and recovery. The conventional use of FRR cannot handle scenarios in which the egress router of the LSP fails, as there is no downstream segment of the LSP that bypasses the failed LSR to which the backup path can re-join.

Consequently, the 50ms recovery times expected from using FRR are not available when an egress router fails, and the recovery times are likely to be of the order of 5 to 30 seconds as IGP reconvergence takes effect.

IV. A NOVEL SOLUTION TO EGRESS LSR FAILURE – THE BI-MESH SURVIVABLE NETWORK

The Bi-mesh Survivable Network (BSN) strategy is a remarkably simple one, achieved through smart naming of the LSPs within the standard dual mesh architecture and configuration of a straightforward path selection in the event of failure.

In the BSN approach, we label LSPs *solely* by the combination of ingress PoP name and egress PoP name. Therefore, the two meshes of the core network have *exactly* the same LSP names in use. However, since the two meshes are

independent of each other, this does *not* result in any confusion in routing tables. This naming strategy is sufficient to allow switching of the packet through the network as the core-ingress LSR for a given packet is a member of only one mesh and therefore has a well-defined route to the router in the egress region of the packet.

We now describe normal operation under this scheme, and then show how it can be exploited to overcome the fundamental limitation of conventional FRR during certain failure scenarios.

A. Normal operation

Consider the routing tables of the edge (intra-region) routers and core LSRs in a stable environment (no failures), and examine how a packet traverses the core. An inter-region packet arriving at an edge router must always be routed to the core. In our network this means the packet is routed to one of the core-ingress LSRs in the local PoP. Any packet arriving at a core-ingress LSR can be routed/switched in a single “hop” (over a direct link or over a single LSP) to the appropriate core-egress LSR.

Now assume that the label placed on the packet at core-ingress identifies the ingress PoP – egress PoP names only. Without link failures, all subsequent core LSRs along the LSP will be able to look at the label and immediately switch the packet to the correct egress interface and the packet will eventually arrive at the core-egress LSR. That LSR will recognise that the packet has reached the end of the LSP, e.g. it will recognise its own PoP name in the egress segment of the label, or see that PHP is in effect, and will route it to the appropriate egress edge LSR.

Locally routed (e.g. intra-region) packets are of no interest in this scenario – they either get routed entirely within the local access network, or between edge routers via single hops to one of the dual-homed core LSRs. They never need traverse a core link.

B. Failure situations

Now consider a scenario in which there is failure of a core link somewhere along the LSP, and the packet in question has just arrived at the core LSR immediately upstream of the failure.

The upstream LSR recognises that it cannot switch the packet to the failed link. If there are no pre-defined survivability rules in place (e.g. a FRR backup path) that LSR would need to make a routing decision. For example, the LSR may route the packet to its mate LSR (within the same PoP) – across the link between the mated “A” and “B” LSRs in Figure 2.

We wish to eliminate the need for core LSRs to make routing decisions so as to lessen router capacity requirements. Exploiting the use of FRR backup paths is the standard possibility. As mentioned above, the FRR backup paths are typically configured to go around a failed link or router, rejoining the original LSP further down the path. If, however, the failure is at the last LSR in the path (the egress router from the core network), there is no point at which a FRR path can

rejoin the original LSP. In such situations the MPLS network relies on IP layer detection of the failure and subsequent updating of IP layer routing tables via the IGP. This can take from 5 to 30 seconds.

The BSN scheme can be exploited to achieve FRR recovery times in this situation.

Under the BSN scheme, all core routers are configured so that upon detecting an immediately-downstream failure (either core-to-core or core-to-edge link failure, or as a result of failure of the router on the far end of the link) all protected packets are switched (not routed) across the intra-PoP link to the mate LSR. That is, the intra-PoP link is set to be the backup path for all egresses (LSP or link) to an LSR. An option would be to mark the packet with a generic “failure” label stacked on top of the current ingress-PoP-egress-PoP label.

Upon receiving one of these protected “failure” packets a core LSR need simply remove that outer label (if it was used), look at the next label in the stack that identifies the core ingress / core-egress PoPs, which is a locally valid label since we are using *exactly* the same naming of LSP within each mesh, and switch the packet to the appropriate LSP within its own mesh. This approach works even when the failure point is the core-egress router of the LSP, as the mate of the failed node would simply become the new egress point. The changeover would occur in the time taken to physically detect the failure, namely of the order of 50ms. Here the BSN scheme is making use of the “alternate egress repair” option described in [3] that is not implemented in standardized FRR [4].

Another option would be to ensure that the traffic reverts to its original LSP as soon as the failure has been bypassed. A link failure is bypassed after a single inter-PoP link is traversed, whereas an LSR failure would require two inter-PoP link traversals. The nature of the failure and/or the number of inter-PoP hops remaining can be encoded into additional labels that are stacked on to the packets as they traverse the alternate mesh. An LSR receiving a packet with such a generic label would pop that label, determine the appropriate egress based on the ingress-Pop-egress-PoP label, and (if necessary) push another generic label back on to the packet. Once the necessary inter-PoP links have been traversed in that mesh, the packets can be switched back to their original mesh, with all generic labels removed.

Finally, the BSN scheme allows for sophisticated differentiation in the survivability treatment provided to different classes of traffic.

We note also that we are explicitly labelling the LSPs, and moreover, re-using the names in both meshes. This feature is not part of the standard auto-configuration function of current

generation LSRs and hence the paths must be manually defined. This adds to the operational complexity in initially setting up the LSPs, but with the major benefit of improved speed of response to certain failure conditions. Part of this configuration would be to configure the routers such that the packet switching tables do not use the ingress interface as part of the switching decision. This is because we need to allow packets coming in over the intra-PoP link to be switched by the same rules as those coming in over the regularly defined meshes.

The capacity of the intra-PoP links in the proposed architecture would need to be set such that all traffic from a single failure can be carried. Using MPLS-TE it would be prudent to ensure that protected traffic not consume more than 50% of link bandwidth in non-failure operation, so that all protected (high priority) traffic can be delivered when a link fails.

V. CONCLUSION

Imposing specific structure on a core network of PoPs (each consisting of mated LSRs) enables a dual plane inter-region architecture to be established. Adding MPLS functionality allows for logical single hop routing between PoPs. Applying the BSN LSP naming scheme allows for enhanced survivability that extends FRR functionality to cope with egress-LSR failure. Consequently we have been able to establish localized MPLS-based recovery mechanisms for *any* link or LSR failure along an LSP – thereby minimizing recovery times and helping to improve network reliability.

ACKNOWLEDGMENT

This work was performed under contract to Telstra and the author gratefully acknowledges Telstra’s permission to publish.

REFERENCES

- [1] D. Awduche, J. Malcolm, J. Agogbua, M. O’Dell and J. McManus, “Requirements for traffic engineering Over MPLS”, IETF RFC 2702, September 1999.
- [2] E. Rosen, A. Viswanathan and R. Callon, “Multiprotocol Label Switching Architecture”, IETF RFC 3031, January 2001
- [3] V. Sharma and F.Hellstrand (Editors), “Framework for MPLS-based recovery”, IETF RFC 3469, February 2003.
- [4] P. Pan, G. Swallow and A. Atlas (Editors), “Fast reroute extensions to RSVP-TE for LSP tunnels”, IETF RFC 4090, May 2005.
- [5] T. Nadeau, M. Morrow, G. Swallow, D. Allan and S. Matsushima, “Operantions and mangement requirements for MPLS networks”, IETF RFC 4377, February 2006.
- [6] J. M. Soricelli, “Tutorial: MPLS fast reroute”, NANOG30, February 2004, <http://www.nanog.org/mtg-0402/pdf/soricelli.pdf>.