A Fast and Scalable Inter-Domain MPLS Protection Mechanism

Changcheng Huang and Donald Messier

Abstract: **With the fast growth of Internet and a new widespread interest in optical networks, the unparalleled potential of Multi-Protocol Label Switching (MPLS) is leading to further research and development efforts. One of those areas of research is Path Protection Mechanism. It is widely accepted that layer three protection and recovery mechanisms are too slow for today's reliability requirements. Failure recovery latencies ranging from several seconds to minutes, for layer three routing protocols, have been widely reported. For this reason, a recovery mechanism at the MPLS layer capable of recovering from failed paths in 10's of milliseconds has been sought. In light of this, several MPLS based protection mechanisms have been proposed, such as end-to-end path protection and local repair mechanism. Those mechanisms are designed for intra-domain recoveries and little or no attention has been given to the case of non-homogenous independent inter-domains. This paper presents a novel solution for the setup and maintenance of independent protection mechanisms within individual domains and merged at the domain boundaries. This innovative solution offers significant advantages including fast recovery across multiple non-homogeneous domains and high scalability. Detailed setup and operation procedures are described. Finally, simulation results using OPNET are presented showing recovery times of a few milliseconds.**

Index Terms: **Protection, MPLS, inter-domain, optical networks.**

I. INTRODUCTION

Real-time communication services have become increasingly critical to businesses and governments. Unlike traditional datagram services where quality of service (QoS) is measured in terms of availability and throughput, real time services require additional QoS criteria such as delay, delay variations, packet drop rate, error rate, and increasingly, fault tolerance and fast recovery. Those QoS requirements combined with the convergence of networking technologies towards an IP based infrastructure have placed significant new demands on the traditional best effort IP network. As a result, several new technologies have emerged to improve the QoS of the IP domain and enable real time applications to converge to this domain. Multi-Protocol Label Switching (MPLS) [1] is one such technology enabling improved QoS control, granularity and traffic engineering in an IP or other network layer domain. Two signaling protocols are defined to bind MPLS labels to Forward Equivalency Class (FEC) and distribute those labels to MPLS peers. First, the Label Distribution Protocol (LDP) designed specifically for this task is described in RFC 3036 [2]. Second, the Resource Reservation Protocol (RSVP) defined in RFC 2205 [3] and extended to support label binding and distribution with RSVP-TE [4]. MPLS, when combined with a differentiated services model, offers faster switching operations and traffic engineering with pre-determined and guaranteed QoS using pre-established and reserved Label Switch Paths (LSP). This improved QoS offering opens the IP door to potential customers traditionally bound, due to their high QoS demands, to circuit switch and connection oriented services. Furthermore the virtual circuit approach of MPLS makes it readily applicable to optical networks that are based on circuit switch technology. Efforts are being made to generalize MPLS architecture for supporting optical networks. Unfortunately, the slow fault recovery mechanism of IP, which can take several seconds to minutes to recover from a failure, is still keeping the IP door closed for some service providers, applications, and users who can not tolerate such long outages. As a result, a mechanism to quickly recover from failures at the MPLS layer has been sought to complement existing higher layer recovery mechanism. The goal is to provide a recovery mechanism at the MPLS layer capable of restoring services around a failure point in tens of milliseconds (10's ms). This fast recovery time would be comparable to SONET recovery as specified in GR253 [5] and therefore make MPLS satisfy the reliability requirements of optical networks.

The latency in Internet path failure, failover, and repair has been well documented over the years. This is especially true in the inter-domain case due to excessively long convergence properties of Border Gateway Protocol (BGP) [6]. Research by C. Labovitz *et al.* [7] presents results, supported by a two year study, demonstrating the delay in Internet inter-domain path failovers averaging three minutes and some percentage of failover recoveries triggered routing table fluctuations lasting up to fifteen minutes. Furthermore the report states that "Internet path failover has significant deleterious impact on end-to-end performance-measured packet loss growth by a factor of 30 and latency by a factor of four during path restoration". Although networks are becoming more and more resilient, there are still frequent network failures that are becoming a cause for concern [8]–[11]. Future optical networks will carry a tremendous amount of traffic. A failure in an optical network will have a disastrous effect. The FCC has reported that network failures in the United States, with an impact on more than 30,000 customers, occur in the order of one every two days, and the mean time to repair them is in the order of five to ten hours [12]. The problem is worse with optical communications technologies because a single failure may affect millions of users. Strategic planning at Gartner Group suggests in [13] that through 2004, large U.S.

Manuscript received October 25, 2002; approved for publication by Thomas Hou, Division III Editor, April 19, 2003.

Changcheng Huang is with Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6 Canada, email: huang@sce.carleton.ca.

Donald Messier's affiliation is omitted upon the request of the author.

enterprises will have lost more than $500 million in potential revenue due to network failures that affect critical business functions. This Internet path failover latency is one of the driving factors behind advances in MPLS protection mechanism.

Protection and restoration issues have been widely studied under various contexts such as SONET rings, ATM, and optical networks [14]–[16]. Several recovery mechanisms have been proposed over the last few years. End-to-end schemes provide protection on disjoint paths from source to destination and may rely on fault signaling to effect recovery switching at the source [17]. Local repair mechanisms for their part affect protection switching at the upstream node from the point of failure, the point of local repair (PLR) and do not require fault signaling [18], [19]. Local repair has the advantage of fast recovery, but in general is not efficient in capacity. Path protection, on the other hand, can optimize spare capacity allocation on an end-to-end basis. Therefore it is typically more efficient.

As briefly discussed in the last section, MPLS being a relatively new technology, the research in advanced protection mechanism for MPLS is still in its infancy. This is especially true for inter-domain protection mechanism. The research conducted, and is still ongoing, has identified several possible solutions to the MPLS intra-domain recovery problem [20]. Each of those solutions presents its own strengths and weaknesses. As a first cut, MPLS restoration schemes can be separated into on-demand and pre-established mechanisms. On-demand mechanism relies on the establishments of new paths after the failure event while pre-established mechanism computes and maintains restoration paths for the duration of the communication session. Due to the fast recovery times sought, this work focuses exclusively on pre-established protection switching. Of the pre-established recovery mechanisms, one of the first commercial product of this being implemented is Cisco Systems' Fast Re-route (FRR) algorithm in the Gigabit Switch Router family. FRR provides very fast link failure protection and is based on the establishment of pre-established bypass tunnels for all Label Switch Routers. The FRR algorithm can switch traffic on a failed link to a recovery path within 20 ms but is limited to the global label assignment case. Several other methods have been proposed based on individual backup LSPs established on disjoint paths from source to destination. An immediate benefit of end-to-end mechanism is scalability. Reference [21] shows that given a network of N nodes, local repair schemes require $N * L * (L - 1)$ backup paths to protect a network if each node has L bi-directional links. For end-to-end schemes, a network with M edge nodes, the total number of backup paths is proportional to $M * (M - 1)$. If M is kept small, a significant scalability advantage is realized. The following paragraphs provide an overview of the most promising intra-domain protection schemes.

The proposal at [22] is an improvement over the one hop CISCO FRR and describes mechanisms to locally recover from link and node failures. Several extensions to RSVP-TE are introduced to enable appropriate signaling for the establishment, maintenance, and switchover operations of bypass tunnels and detour paths. The Fast Reroute method will be referred to as Local Fast Reroute (LFR) in this paper. In the Local Fast Reroute, one-to-one backup LSPs can be established to locally bypass a point of failure.

A key part of this proposal is to setup backup LSPs by making use of label stack. Instead of creating a separate LSP for every backed-up LSP, a single LSP is created which serves to backup a set of LSPs. Such an LSP backing up a set of primary LSPs is called a bypass tunnel.

The key advantage of LFR is the very fast recovery time while its disadvantages are scalability issues due to the potential large number of bi-directional links and complexity in maintaining all the necessary label associations for the various protected paths.

The first end-to-end path protection scheme is presented at [21] and uses signaling from the point of failure to inform the upstream LSRs that a path has failed. Here a Reverse Notification Tree (RNT) is established and maintained to distribute the fault and recovery notifications to all ingress nodes which may be hidden due to label merging operations along the path. The RNT is based on the establishment of a Path Switch LSR (PSL) and a Path Merge LSR (PML). The PSL is the origin of the recovery path while the PML is its destination. In the case of multipoint-to-point tree, the PSLs become the leaves of the multicast trees while the PMLs are the roots. The main advantages of RNT protection are scalability and efficient use of resources while its disadvantage is long recovery time due to the propagation of failure notification messages.

Another end-to-end path protection mechanism presented at [23] is called End-to-end Fast Reroute (EFR). It can achieve nearly the same protection speed as LFR, but is extremely inefficient in terms of bandwidth resource. It requires about two times the bandwidth of the protected path for protection path. For more about this approach, readers are referred to [23].

Current research and recent proposals deal with the intra-domain case or assume homogeneity and full cooperation between domains. Recognizing the growing need to provide a solution for the more general case, this paper proposes a new and innovative solution to solve the inter-domain protection problem for LSPs spanning multiple inhomogeneous and independent domains. The proposed solution is based on the use of concatenated primary and backup LSPs, protection signaling and a domain boundary protection scheme using local repair bypass tunnels. We will call our scheme Inter-Domain Boundary Local Bypass Tunnel (IBLBT) in this paper to distinguish with other solutions.

II. PROBLEM STATEMENT AND PROPOSED SOLUTION

The MPLS protection mechanisms presented in Section I include LFR, EFR, and RNT. All were designed for intra-domain failure recovery and will generally not function when the primary LSP is not bounded to a single administrative domain. The scalability problem with LFR will be stretched further if multiple domains are involved because each domain may have hundreds of nodes and links that require bypass tunnels for protection. While both EFR and RNT suffer longer delays due to the long LSPs that span several domains, EFR becomes more inefficient compared to RNT because of its extra bandwidth requirements.

A unique issue for inter-domain protection is that separate

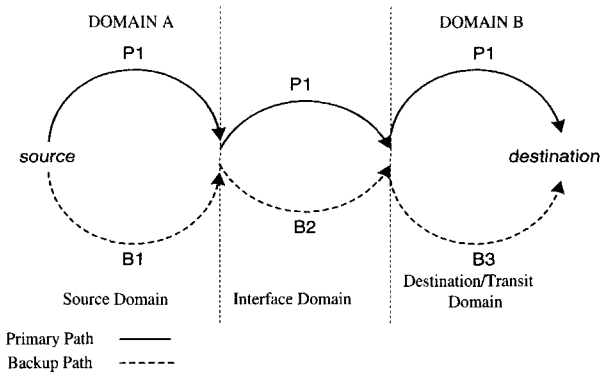

Fig. 1. Concatenated primary and backup LSPs.

domains may not cooperate with each other. Each domain is administered through a different authority. Some authorities, such as carriers, are not willing to share information with each other. Certain critical information may have significant impact on the operation of public carriers if they are disclosed. For example, failure information is typically considered negative on the image of a public carrier and competitors may exploit this information to their advantages. Most carriers will consider this information confidential and will not likely share this information with their customers and other carriers. When an internal failure happens, a carrier will try to contain this information within its own domain and try to recover from the failure by itself. Both end-to-end RNT and end-to-end EFR require some kind of failure signaling to all the upstream domains. Containing this failure signaling to the originating domain will make end-to-end RNT and EFR almost impossible.

A complete solution to the inter-domain protection problem can be found if we turn the apparent difficulties in end-to-end RNT into advantages. Such is the case for the independence of domains. Accepting the fact that domains will be independent and inhomogeneous leads to the idea of establishing an independent path protection mechanism within each domain while at the same time being able to guarantee protection throughout the path from end to end. What is required is a solution at the domain boundaries to ensure protection continuity. For the solution to work, each domain must provide its own RNT protection scheme which it initiates, establishes, maintains, and hands over to the next protection domain at the domain boundary. A domain protection scheme must therefore be executed completely within that domain with no involvement from other domains. The first step towards this solution is to allow the primary and backup LSPs to be concatenated at the domain boundaries. Usage of concatenation, in this context, means that specific actions must be taken at this point in the LSP to ensure continuity of service and protection across domain boundaries. Fig. 1 illustrates the fundamental principles behind this solution. The primary path P1 is protected through three separate backup paths namely B1, B2, and B3. B1 is initiated in the source domain, B2 at the domain boundary, and B3 in the destination domain. Each of those backup paths is independent of each other and does not require fault notification beyond its own domain.

This innovative solution permits the isolation of the protection mechanism to a single domain or domain boundary. Furthermore, domains can now use independent protection mechanisms and signaling schemes and do not need to propagate their internal failure notifications to adjacent domains. This solution combines the advantage of fast local repair at the domain boundaries and the scalability advantage of end-to-end protection within domains.

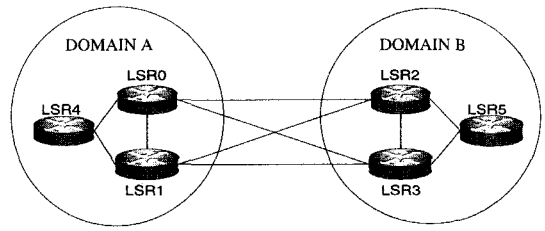

Fig. 2. Dual exit LSRs fully meshed.

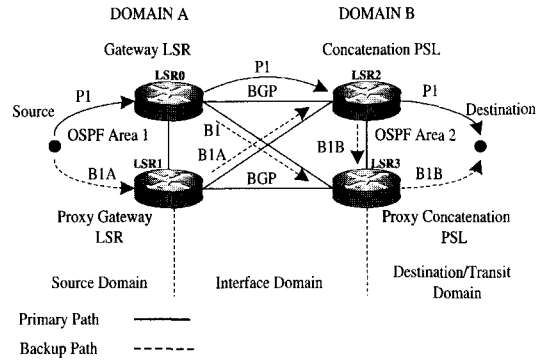

Fig. 3. Inter-domain protection.

In summary, the proposed solution to solve the inter-domain MPLS recovery problem is based on the establishment of independent protection mechanisms within domains using concatenated primary and backup LSPs, minimal protection signaling between domains, and local repair at the domain boundaries. Viewed from end-to-end in Fig. 1, the primary LSP is protected by three or more distinct and independent protection regions merged at their respective boundaries. Those protection regions are the Source Protection Domain, the Domain Interface Protection and the Destination/Transit Protection Domain. In addition to those three protection regions, transit protection regions are also possible when a protected LSP transits one or more independent domains before reaching its destination. In such a case, there would be several domain interface protections in place.

Our solution introduces and makes use of Gateway LSRs and Concatenation Path Switch LSRs (CPSLs) as well as Proxy Concatenation PSLs (PCPSL) and Proxy Gateway LSRs (PGL). Those new protection elements are used to pre-establish inter-domain local bypass tunnels and guarantee protection against node and link failures when sufficient protection elements are present.

In the following discussions, we assume that there are at least two separate links connecting two pairs of border routers between any neighboring domains. This will allow us to provide protection for two neighboring domains without counting on the support of a third domain under the context of single point failure. One example that satisfies this requirement is shown in

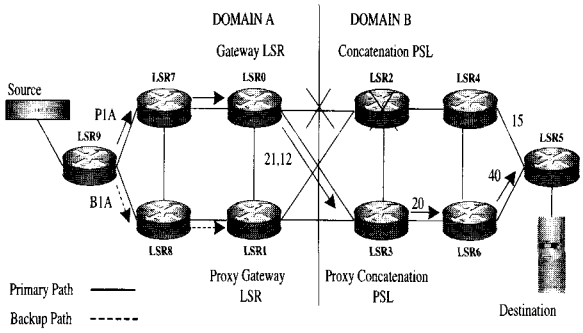

Fig. 4. Inter-domain link or concatenation PSL failure recovery.

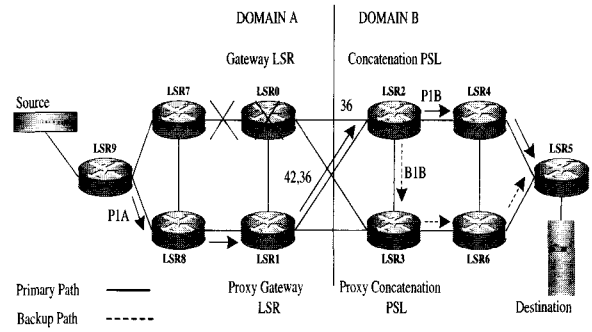

Fig. 5. Source domain failure recovery.

Fig. 2. Our focus is therefore on removing the interdependency among domains that are not directly linked and further limiting the dependency between neighboring domains as discussed in the next section. We will use the scenario of Dual Exit LSRs Fully Meshed (Fig. 2) as our example case. The principles of our solution can be readily applied to all other scenarios that satisfy the condition stated at the beginning of this paragraph.

Fig. 3 illustrates the topology where a primary protected LSP P1A is protected in Domain A via backup path B1A, protected at the boundary via local backup path B1, and protected in Domain B through backup path B1B. LSR 0 is the selected Gateway LSR for path P1 while LSR 1 is its corresponding PGL. LSR 2 is the CPSL for the same primary path while LSR 3 is the PCPSL. The PGL and PCPSL are responsible to maintain end-to-end path integrity in the event of a Gateway or CPSL failure. The selection of the PCPSL and its significance in the recovery process is critical for the operation of this scheme. This point is evident when looking at Fig. 3. In Fig. 3, we note that B1 and B1B are routed through the PCPSL LSR 3. Although the identification of the PCPSL is simple in a Dual Exit LSR topology, its role is nevertheless important. It is the merging of the local inter-domain backup path B1 and the destination domain backup path B1B at the PCPSL LSR 3 that permits full and continuous protection across domains. Without this action, recovery traffic on B1 would be dropped at LSR 3 since it could not make the necessary label association. The merging action of the PCPSL ensures label binding between B1 and B1B, enabling the recovery traffic from the Gateway LSR to be switched to the destination.

III. DETAILED PATH RECOVERY PROCEDURES

This section describes the detailed procedures to be executed by each participating node in a protection domain.

A. Gateway LSR Procedures

Refer to Fig. 4. When a Gateway LSR (LSR 0) detects a failure in its external link, either a link or a LSR 2 node failure, the following protection algorithm is executed.

1) The Gateway LSR determines what LSPs are affected by the failure and which of those are protected LSPs;
2) Consults the Incoming Label Map (ILM) and the Next Hop Label Forwarding Entry (NHLFE) and extracts the associated local inter-domain bypass tunnel label and the label bound by the PCPSL (LSR 3) for that protected LSP (labels 21 and 12 respectively from Fig. 4);
3) Removes current labels from the protected LSP and inserts new labels. The label stack is now (21,12);
4) Forwards packets onto the bypass tunnel;

As a result of those actions, the primary path P1 is switched around the link failure LSR 0-LSR 2 or LSR 2 node failure, to the PCPSL (LSR 3) which merges the backup traffic with the pre-established backup LSP to the egress LSR. The recovery path is LSRs 9-7-0-3-6-5.

B. Proxy Gateway Procedures

A failure occurring along the primary path in the source domain as illustrated at Fig. 5, either a link or Gateway LSR node failure, will result in the primary path being switched to the backup path. The switch to the backup path will occur at the PSL. In Fig. 5, the backup path leads to the CPSL through the PGL (LSR 1) and global label space is used. The PGL pushes its local repair bypass tunnel label (42) onto the stack and forwards the packet to the CPSL. Label 36 was inserted by the PSL to properly associate the traffic at the PML. The recovery path is LSRs 9-8-1-2-4-5.

IV. SIMULATION RESULTS

To verify the potential for the proposed IBLBT solution, three separate models were built. The first one is a simple BGP-4 network consisting of three independent domains. This model was used to gather baseline data on recovery times using traditional layer three inter-domain routing protocols. This data will be used to compare recovery times with the proposed MPLS protection scheme. The second model implements MPLS recovery using an end-to-end path protection mechanism. The model was built using dynamic LSPs. For the end-to-end recovery to work, it is necessary for all nodes in the model to share a common signaling and recovery mechanism. This is necessary in the extended intra-domain end-to-end scheme since domains have to fully cooperate in the recovery process. As discussed in previous chapters, this naive extended intra-domain solution would likely not be found in real networks. Nevertheless, the model is useful to serve as a comparison point with IBLBT solution proposed in this paper. In contrast to end-to-end recovery, IBLBT isolates recovery to the domain boundary or to an individual domain. The third and final model built is the model implement-

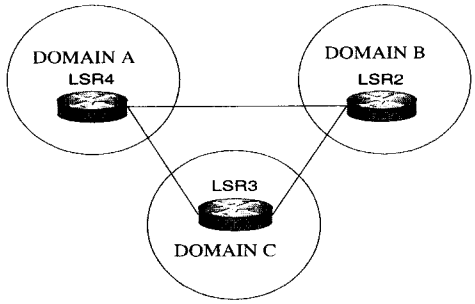

Fig. 6. Border gateway baseline model.

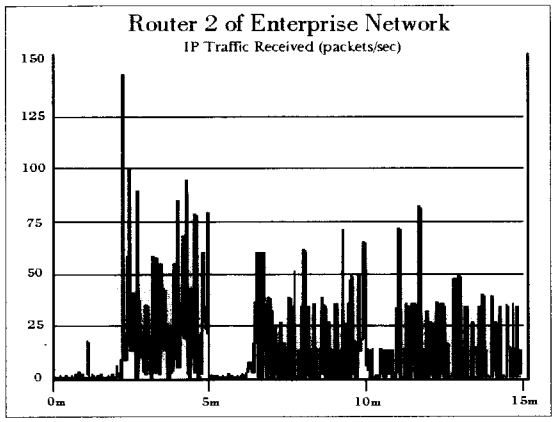

Fig. 7. IP packets received at destination node (router 2) and sourced from router 4.

ing the proposed solution with its inter-domain boundary local repair tunnels. All models are run with various failure scenarios to collect data on recovery time for further analysis and comparison.

A. Border Gateway Baseline Model Setup

Although the model is very simple, it does provide a glimpse into very long failure recovery latencies experienced in traditional layer three networks when multiple domains are involved. Fig. 6 represents the model built using Opnet 8.0. The links are all DS-3 links running BGP-4 and carrying Email, File Transfers, Database access, and Web browsing. The traffic was set at approximately 1.5 Mbps and the links are all 100 km apart. The simulation was run for 15 minutes with a failure point in link 4-2 set for 300 seconds. The first 2 minutes are reserved for network setup and convergence. No user traffic is generated.

For intra-domain recovery, the layer three protocols may be able to recover from failures comparatively rapidly. In the case of OSPF (Open Shortest Path First [24]) for example, its ability to maintain alternate routes in its Link State Database enables OSPF to quickly re-compute a new route upon failure and introduce this new route in its routing table at the point of local repair. Recovery can therefore take place within a few milliseconds while the complete domain convergence to the new route may take 10s of seconds. For the inter-domain situation studied in the paper, the use of inter-domain routing protocols such as BGP4 makes the recovery latencies much worse [7].

In the simulation, upon a link failure between router 4 and router 2, BGP required on average 70 seconds to converge to a new route to router 2 through router 3. Fig. 7 presents a graph of IP packets received at router 2 (sourced from router 4 only) and clearly shows a 73 seconds gap between the link failure and recovery through route LSR4-LSR3. The link failure was programmed to occur at 300 seconds. Due to the 73 seconds gap before recovery, TCP retransmissions reach the maximum and the connection is reset with an abort issued to the source applications.

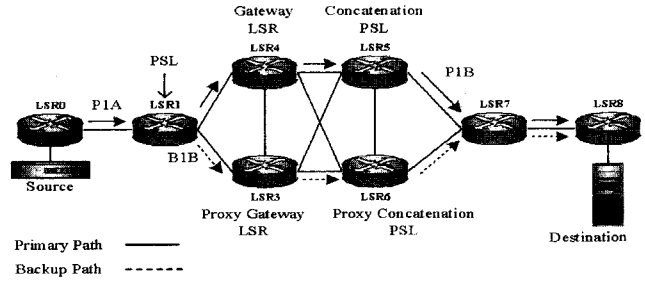

Fig. 8. MPLS end-to-end protection model.

B. MPLS End-to-End Protection Model

This MPLS model was built to measure recovery time for an end-to-end protection case and is represented in Fig. 8. As stated earlier, it is recognized that such an inter-domain end-to-end protection mechanism is naive for the reasons discussed in Section II. However, to obtain comparative data from such a scheme, LSPs were configured using dynamic LSPs and all nodes share the same signaling and protection mechanism. Traffic was generated using two separate Gigabit Ethernet LANs, each with twenty-five users running high-resolution video conferencing applications over UDP. Additional applications were configured such as heavy database access and email, file transfers, and print sessions using TCP. Traffic entered the MPLS network at the ingress node LSR 0. The Egress and Ingress LSRs were modeled as CISCO 7609 while the transit LSRs were CISCO 7000 routers. The Egress and Ingress LSRs were selected based on the number of Gigabit Ethernet ports available for the source LANs. IP forwarding processor speeds were increased to 50,000 packets/sec on all nodes to permit higher traffic volumes for the simulation. High traffic volume was necessary to ensure high link utilization for measurement purposes. Traffic was switched between LSRs based on the Forward Equivalency Class (FEC) associated with the incoming traffic and the established Paths. The selected Primary Path is shown in Fig. 8 and follows path LSRs 0-1-4-5-7-8 while the pre-established end-to-end backup LSP follows LSRs 1-3-6-7-8 (LSR 1 is the PSL). All model links are OC-12 with 0.8 ms delay for inter-domain links and 4 ms delay for intra-domain links. This approximates 1200 km intra-domain links and 240 km inter-domain links. The average load on the network was kept at approximately 125 Mbps.

Several failure scenarios were studied as follows:

1) Source domain failure (Link 1–4 failure);
2) Domain interface Failure (Link 4–5 and node 5 failure);
3) Destination domain failure (Link 5–7 failure).

A failure and recovery process was configured in Opnet to effect at 170 seconds from the simulation start time. All simulations were run for a total of 180 seconds. The 170 seconds time before failure was selected to ensure sufficient time for all routing processes to complete their initial convergence, for traffic generation processes to reach steady state prior to the network failure, and for the MPLS processes to establish LSPs after the initial layer three routing protocol convergence. The simulation end time is selected to allow sufficient time for recovery and steady states to return while being kept at a minimum to reduce the simulation run time. The large amount of application traffic generated during the simulation caused the simulation run time to be in excess of one hour.

This model makes use of CR-LDP keep-alive messages to detect node failures while link failures are detected through lower layer Loss of Signal (LOS). The keep-alive message interval was configured for 10 ms while the hold off timer was set at 30 ms. Those short intervals were selected arbitrarily but taking into account the OC-12 line rate with a view to reduce packet loss during failover. Upon detecting a failure the node upstream from the point of failure sends an LDP notification message to the source node informing it of the failure and the affected LSPs. Triggered by this notification message, the source node switches to the recovery path. This LDP notification message is encapsulated in an IP packet and forwarded to the ingress node for action. Several network probes were configured to collect data on recovery times, routing tables, link state databases, and traffic in and out of all LSPs as well as forwarding buffer utilization.

For this work, recovery time was measured at the merge point of the primary and backup paths (i.e., PML). This recovery time is from the receiver's perspective and represents the difference in time between the reception of the last packets on the primary path and reception of the first packets on the recovery path. The recovery time includes failure detection time, time for the transmission of failure notification messages, protection-switching time, and transmission delay from the recovery point to the merge point. To obtain the necessary LSP traffic data for the measurement of recovery time, LSP output traffic for primary and backup LSPs at the merge point was sampled every 100 seconds. This sampling time was selected to provide sufficient granularity into the recovery process while maintaining simulation output files to a manageable size.

Link 5–7 failure recovery results: In this failure scenario, LSR 5 detects the failure through the LOS, transmits the LDP notification message to the PSL (LSR 1) which switches traffic to the backup path. Based on five replications and a 95% confidence interval, the average recovery time for this scenario was 12.09 ms $\leq R \leq$ 14.39 ms with an average of 13.24 ms.

Link 1–4 failure recovery results: In this failure scenario, LSR 1 detects the failure with the LOS. At this time LSR 1 switches the traffic to the backup path that merges at the egress LSR 8. In the same fashion as in the previous case, the output of the primary and backup paths are monitored and the recovery time is measured to average 7.18 ms with a 95% confidence interval 6.45 ms $\leq R \leq$ 7.91 ms. This short recovery time reflects the local failure detection through LOS.

Link 4–5 failure recovery results: In this scenario, LSR 4 detects the failure through LOS and forwards a notification message to the PSL (LSR 1) that switches traffic to the recovery path LSP 1–8. The recovery time R is measured as 12.0 ms on average with 11.56 ms $\leq R \leq$ 12.44 ms. The increase in recovery time is due to the 4 ms link delay experienced by the failure notification message propagated from LSR 4 to LSR 1.

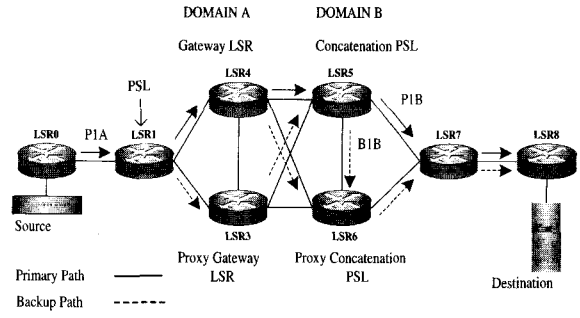

Fig. 9. MPLS end-to-end protection model.

Node 5 failure recovery results: In this failure scenario, LSR 4 detects the failure using the keep-alive message after the 30 ms hold off period. LSR 4 sends an LDP notification message to LSR 1 that is the PSL responsible to switch to the backup LSP. Recovery is 46.18 ms on average with 45.88 ms $\leq R \leq$ 46.88 ms. The long recovery time is largely characterized by the fault detection time.

C. Inter-Domain Boundary Bypass Tunnel Model

In the IBLBT model, the primary and backup paths were established following the proposed inter-domain protection algorithms. As described in previous sections and depicted at Fig. 9, concatenated LSPs were setup within each domain with backup paths using bypass tunnels established manually as described in Section II. The simulations were run for 130 seconds with failures programmed for 125 seconds. Shorter simulation time is possible with this model because static LSPs are used and no setup time is required during the simulation. Other than the recovery mechanism, the model was setup identically to the previous end-to-end MPLS model.

Link 5–7 failure recovery results: In this scenario, border LSR 5 detects the failure through LOS. LSR 5 is also the PSL for this domain and switches the affected traffic to Domain B backup path through LSR 6. The merge point of the primary and backup LSP is LSR 7 and the output of both primary and backup paths are measured at that node. The average recovery time was measured to be 8.56 ms with 8.26 ms $\leq R \leq$ 8.86 ms. This recovery time is characterized mostly by the cumulative 8 ms link delay consisting of 4 ms on link 5–6 and 4 ms on link 6–7. As opposed to the end-to-end model, the failure is isolated to domain B.

Link 1-4 failure recovery results: In this failure scenario, LSR 1 detects the failure through LOS and switches traffic immediately to the backup path LSP 1–5. Once again LSP output is monitored at the merge point (LSR 5) and recovery time is measured at 4.7 ms with 4.5 ms $\leq R \leq$ 4.9 ms which is largely composed of the 4 ms link 1–3 delay. The failure is isolated to

Table 1. Comparisons of end-to-end recovery and IBLBT.

	Link 1-4	Link 4-5	Link 5-7	Node 5
IBLBT	4.7 ms	1.19 ms	8.56 ms	32.02 ms
End-to-end recovery	7.18 ms	12.0 ms	13.24 ms	46.38 ms

the source domain.

Link 4–5 failure recovery results: In this failure scenario, LSR 4 detects the failure through LOS and switches traffic to the local repair inter-domain bypass tunnel from LSR 4 to LSR 5 that switches the traffic to merge point at LSR 7. Output traffic is monitored at the merge LSR. For a sample LSP output file, the last traffic received on the primary path LSP 5–7 is at 125.0041 seconds while the first traffic received on the recovery path is at 125.0052 seconds. The recovery time is therefore 1.1 ms largely characterized by the inter-domain link 4–6 delay of 0.8 ms. On average the recovery time R was measured to be 1.19 ms and 1.01 ms $\leq R \leq$ 1.37 ms.

Node 5 failure recovery results: In this scenario, node 4 detects the failure only after the 30 ms hold off timer expires and the overall recovery is measured at 32.02 ms on average with 31.6 ms $\leq R \leq$ 32.44 ms.

D. Results Summary

When compared to BGP recoveries of over 70 seconds demonstrated in this paper and recoveries of several minutes widely reported, there is little argument on the recovery benefits of MPLS. Comparing end-to-end recovery with the IBLBT case as shown in Table 1 is not so evident however. The recovery speed benefits of IBLBT over the end-to-end case would have been much more evident had the simulation model included several more independent domains. Of course the further away the failure is from the point of repair, the longer the recovery time. Given the simplicity of the models in this work, the significant advantages of IBLBT could not be exploited fully against the end-to-end case.

V. CONCLUSIONS

The growing demand for QoS has led to significant innovations and improvements on the traditional best effort IP networks. Technologies such as MPLS provide important advantages over the classical hop-by-hop routing decision processes. The ability of MPLS to apply equally well to various layer 1 technologies, including Wave Division Multiplexing (WDM), makes this technology a strong contender for current leading edge and future networks. Furthermore, due to its label switching architecture, MPLS can provide very fast recovery mechanism complementing existing lower layer protection schemes. The development of new techniques to provide path protection at the MPLS layer will certainly continue. The proposed IBLBT protection mechanism presented in this paper is an innovative and unique scheme to provide protection across multiple independent domains. It relies on only a very basic amount of information provided by neighboring domains and makes no assumption on protection mechanisms of other domains and level of cooperation. Simulation results show recovery times of a few milliseconds which displays the potential for this proposed solution for MPLS inter-domain protection.

In general, our solution permits the isolation of the protection mechanism to a single domain or domain boundary. Furthermore, domains can now use independent protection mechanisms and signaling schemes and do not need to propagate their internal failure notifications to adjacent domains. This solution combines the advantages of fast local repair at the domain boundaries and the scalability advantages of path protection within domains. A recent proposal to IETF has addressed the issue of extending RSVP to support inter-domain protection and restoration schemes such as the one proposed by this paper [25].

REFERENCES

[1] E. Rosen, A.Viswanathan, and R. Callon, "Multiprotocol label switching architecture," *IETF RFC 3031*, Jan. 2001.

[2] L. Andersson *et al.*, "LDP specification," *IETF RFC 3036*, Jan. 2001.

[3] R. Braden *et al.*, "Resource reservation protocol (RSVP) - version 1 functional specification," *IETF RFC 2205*, Sept. 1997.

[4] D. O. Awduche *et al.*, "RSVP-TE: Extensions to RSVP for LSP tunnels," *IETF RFC 3209*, Dec. 2001.

[5] Bellcore, "SONET common generic criteria," *GR 253 Core*, Dec. 1995.

[6] Y. Rekhter, T. J. Watson, and T. Li, "A border gateway protocol 4 (BGP-4)," *IETF RFC 1771*, Mar. 1995.

[7] C. Labovits *et al.*, "Delayed Internet routing convergence," *IEEE/ACM Trans. Networking*, vol. 9, no. 3, pp. 293–306, June 2001.

[8] T. G. Griffin and G. Wilfong, "An analysis of BGP convergence properties," in *ACM SIGCOMM'99*, Cambridge, Sept. 1999.

[9] S. J. Jeong *et al.*, "Policy management for BGP routing convergence using inter-AS relationship," *J. Commun. Networks*, vol. 3, no.4, Dec. 2001.

[10] R. Mahajan, D. Wetherall, and T. Anderson, "Understanding BGP misconfiguration," in *ACM SIGCOMM 2002*, Pittsburgh, Aug. 2000.

[11] K. Varadhan, R. Govindan, and D. Estrin, "Persistent route oscillations in inter-domain routing," *Comput. Netw.*, vol. 32, no. 1, 1999.

[12] P. Demeester, T. Wu, and N. Yoshikai, "Survivable communications networks," *IEEE Commun. Mag.*, vol. 37, no. 8, Aug. 1999.

[13] B. Hafner and J. Pultz, "Network failures: Be afraid, be very afraid," *Gartner Group, Note SPA-09-1285*, 15 Sept. 1999.

[14] O. J. Wasem, "An algorithm for designing rings for survivable fiber networks," *IEEE Trans. Rel.*, vol. 40, 1991.

[15] T. Frisanco, "Optimal spare capacity design for various switching methods in ATM networks," in *IEEE ICC'97*, Montreal, June 1997.

[16] B. T. Doshi *et al.*, "Optical network design and restoration," *Bell Labs Tech. J.*, Jan.–Mar. 1999.

[17] P.-H. Ho and H. T. Mouftah, "Reconfiguration of spare capacity for MPLS-based recovery," to appear in *IEEE/ACM Trans. Networking*.

[18] M. Kodialam and T. V. Laksman, "Dynamic routing of locally restorable bandwidth guaranteed tunnels using aggregate link information," in *IEEE INFOCOM 2001*, Anchorage, April, 2001.

[19] W. D. Grover and D. Stamatelakis, "Cycle-oriented distributed preconfiguration: Ring-like speed with mesh-like capacity for self-planning network restoration," in *IEEE ICC'98*, Dresden, June 1998.

[20] V. Sharma and F. Hellstrand, "Framework for multi-protocol label switching (MPLS) based recovery," *IETF RFC 3469*, Feb. 2003.

[21] C. Huang *et al.*, "Building reliable MPLS networks using a path protection mechanism," *IEEE Commun. Mag.*, Mar. 2002.

[22] P. Pan *et al.*, "Fast reroute extensions to RSVP-TE for LSP tunnels", *IETF Draft*, work in progress, <draft-ietf-mpls-rsvp-lsp-fastreroute-00.txt>, Jan. 2002.

[23] D. Haskin and R. Krishnan, "A method for setting an alternative label switched paths to handle fast reroute", *IETF Draft*, work in progress <draft-haskin-mpls-fast-reroute-05.txt>, Nov. 2000.

[24] J. Moy, "OSPF version 2," *IETF RFC 2328*, Apr. 1998.

[25] C. Pelsser and O. Bonaventure, "RSVP-TE extensions for interdomain LSPs," *IETF Draft*, work in progress <draft-pelsser-rsvp-te-interdomain-lsp-00.txt> , Oct. 2002.

Changcheng Huang was born in Beijing, China. He received the B. Eng. in 1985 and the M. Eng. in 1988 both in Electronic Engineering from Tsinghua University, Beijing, China. He received a Ph.D. degree in Electrical Engineering from Carleton University, Ottawa, Canada in 1997. He worked for Nortel Networks, Ottawa, Canada from 1996 to 1998 where he was a systems engineering specialist. From 1998 to 2000 he was a systems engineer and network architect in the Optical Networking Group of Tellabs, Illinois, USA. Since July 2000, he has been an assistant professor in the Department of Systems and Computer Engineering at Carleton University, Ottawa, Canada.

Donald Messier Photograph and biography are omitted upon the request of the author.