# Evaluation of Post-Fault Restoration Strategies in Multi-Domain Networks

Feng Xu[1], Tamal Das[2], Min Peng[3], Nasir Ghani[1]
[1]*University of New Mexico,* [2]*IIT Bombay,* [3]*Wuhan University*

***Abstract:*** *Although multi-domain survivability is a major concern, few studies have considered post-fault restoration schemes. This paper proposes two such strategies (based upon hierarchical routing and signaling crankback) to handle single link failures in multi-domain IP/MPLS networks (also extendible to optical DWDM networks). The performance of the proposed solutions is then compared via simulation.*

## I. INTRODUCTION

Network survivability is a major issue and pre-planned *protection* and post-fault *restoration* schemes have been developed for IP *multi-protocol label switching* (MPLS) and optical *generalized MPLS* (GMPLS) networks [1]. In the former, backup paths are pre-computed and reserved for rapid recovery, i.e., fast switching, whereas in the latter, active re-routing procedures are used to recompute routes after faults. However, most of these strategies have only been applied to single-domain instead of *multi-domain* settings where only selected nodes have partial "global" views due to scalability and confidentiality concerns [2]. Although some studies have also looked at multi-domain recovery, most have focused on pre-provisioned protection. For example, [3] and [4] propose sequential/parallel strategies for working/backup routes with improved path diversity, but they are subject to inter-domain "trap" topology or significant routing overheads. Moreover, most multi-domain protection schemes can only perform single-fault recovery. Given the above, it is imperative to study multi-domain post-fault restoration. Even though this approach may yield longer recovery times, it is well-suited to most data services. More importantly, restoration is more resource efficient than protection and can offer a viable "last-gap" alternative in case of multiple failures as part of a tiered survivability framework. Along these lines, two multi-domain restoration schemes are studied here based upon hierarchical routing and signaling crankback strategies. These solutions are presented in Section II and then compared in Section III.

## II. MULTI-DOMAIN RESTORATION SOLUTIONS

Both hierarchical routing and signaling crankback strategies have been studied for working-mode multi-domain networks recently; see [5] and [6]. These setups assume the availability of full intra-domain link-state at interior nodes, e.g., via OSPF-TE protocol. Conversely, inter-domain visibility is more limited. Namely, hierarchical routing runs a second level of link-state routing between border nodes and can provide limited view of internal domain resources. Meanwhile, crankback signaling only assumes the availability of very limited path-vector state, as provided by BGP protocols. In both schemes, however, setup signaling is done using the *RSVP-traffic engineering* (RSVP-TE) protocol.

Before introducing the schemes, the requisite notation is defined. A multi-domain network is comprised of $D$ domains and is modeled as a set of domain sub-graphs, $\boldsymbol{G}^i(\boldsymbol{V}^i,\boldsymbol{L}^i)$, where $\boldsymbol{V}^i = \{v^i_1, v^i_2, \ldots\}$ are the domain nodes and $\boldsymbol{L}^i=\{l^{ii}_{jk}\}$ are the intra-domain links in domain $i$ ($1 \le i \le D$). Each domain also hosts a *path computation element* (PCE) with full access to the interior topology graph [5],[6]. The *inter-domain* link between border node $v^i_k$ in domain $i$ and $v^j_m$ in domain $j$ is further denoted as $l^{ij}_{km}$. Each node also maintains a list of traversing connections, $\underline{\boldsymbol{A}}^i_j$ for node $v^i_j$, where each entry in $\underline{\boldsymbol{A}}^i_j$ is a route vector. Finally, the RSVP-TE messaging fields include a path route vector, $\underline{\boldsymbol{R}}$, and an exclude link vector, $\underline{\boldsymbol{X}}$, to track failed links. Additionally, dual intra/inter-domain counters, $h_1$ and $h_2$, are also defined for signaling crankback.

### A. Restoration with Hierarchical Routing

Hierarchical routing implements *topology abstraction* to summarize domain-level state. Namely, designated *routing area leader* (RAL) nodes, possibly co-located with the PCE, run specialized graph transformation algorithms to reduce physical domain-level topology/resource graphs, i.e., $\boldsymbol{G}^i(\boldsymbol{V}^i,\boldsymbol{L}^i)$. This abstracted state is then flooded across border nodes in all domains to build "global" network views and allow them to perform inter-domain path computation. As per [5], two specific topology abstraction schemes are used here:

**Simple-Node Abstraction:** This scheme reduces each domain into a virtual node emanating physical inter-domain links. This scheme has low routing overheads but provides no domain-internal visibility. However, border nodes do provide full updates for their physical inter-domain links, i.e., $l^{ij}_{km}$, $i \neq j$.

**Full-Mesh Abstraction:** This scheme computes/advertises "abstract links" between all domain border nodes in addition to physical inter-domain link updates. The goal is to provide a summarization of domain traversal costs via abstract links $l^{ii}_{jk}$; see [5] for algorithmic details. As expected, this approach entails notably higher computational/routing complexity.

Using the above information, PCE entities in the hierarchical routing setup compute "skeleton path" sequences for inter-domain requests. Namely, the *loose routes* (LR) are generated by running modified shortest path algorithms over the "global" topology/resource graphs. Now leveraging from [5], two different LR path computation approaches are studied here, i.e., *minimum hop* and *minimum relative distance*. Specifically, each physical/abstract link in the inter-domain network graph is assigned a "cost", $\omega^{ij}_{kmn}$, as follows:

**Minimum hop:** $\omega^{ij}_{kmn} = 1$

**Min. relative distance:** $\omega^{ij}_{kmn} = 1 / (u \cdot c^{ij}_{kmn} / C^{ij}_{kmn})$

where $C$ and $c$ are the maximum and unreserved link capacities, respectively, and $u$ is a constant. The former metric tries to achieve (inter-domain) resource minimization whereas the latter aims for load-balancing. Finally, if the above LR computation is successful, full *explicit route* (ER) expansion is done along the LR sequence using RSVP-TE signaling.

Now consider the application of hierarchical routing for post-fault restoration, i.e., after an intra- or inter-domain link failure. It is assumed the link end-points can quickly discover the failure via rapid lower layer mechanisms (out of scope herein). These endpoint nodes then loop through their active



978-1-4244-9854-3/10/$26.00 ©2010 IEEE 

connection lists, $\underline{\boldsymbol{A}}^i_j$, to search for any traversing connections using the failed link. Here, all such connections are removed from $\underline{\boldsymbol{A}}^i_j$ and appropriate restoration procedures initiated for each. Specifically, the detecting nodes first send out resource takedown messages along the failed upstream and/or downstream path segments. Next, each node on the upstream side of the failed link notifies the source of the connection via a notification message with the failed link noted in exclude route vector, $\underline{\boldsymbol{X}}$. This message also sets an appropriate restoration flag to indicate that this is a failed connection.

Upon receiving the above failure notifications from downstream nodes, the source nodes query their respective PCE's to trigger (post-fault) hierarchical inter-domain path re-computation and subsequent re-routing. Taking into account the propagation delays of global link state (via hold-off timers), the source domain PCE extracts the failed inter-domain link from $\underline{\boldsymbol{X}}$ and prunes it from its inter-domain graph before computing a revised LR. The subsequent setup procedures are same as described for working-mode operation. Hierarchical routing restoration is shown in the top of Fig. 1, where a failure on link $l_{22}^{56}$ between domains 5 and 6 causes the endpoint nodes to issue takedown sequences to free resources along the failed paths. The source node then initiates ER signaling expansion along the least-cost LR and the restored path is signaled via domains 2, 7 and 4.

### B. Restoration with Enhanced Signaling Crankback

Recently, some (working-mode) signaling crankback schemes have also been studied for multi-domain networks, leveraging related RSVP-TE extension (RFC4920). In general, these strategies assume minimal inter-domain state knowledge and instead rely upon distributed "per-domain" computation. Namely, domain PCE's iteratively select the "next-hop" domain and then initiate intra-domain ER towards the specific border node/egress link. Path computation concludes when the fully-expanded route sequence reaches the destination node. However, as resource deficiencies can arise during setup, various crankback procedures have been defined. Along these lines, this effort leverages the joint intra/inter-domain crankback scheme of [6]. This solution limits intra/inter-domain crankback attempts to $H_1$ and $H_2$, respectively, and implements two key steps, i.e., crankback *notification* and *re-computation*. Specifically, the first node detecting a resource-deficient link sends a *PATH_ERR* message to its domain ingress node. Here the intra-/inter-domain crankback counters values are set to $h_1$=$H_1$ and $h_2$=$H_2$ and the problematic link noted in the exclude link vector, $\underline{\boldsymbol{X}}$. Upon receiving this *PATH_ERR* notification, an ingress border node performs *re-computation*. Namely, it decrements $h_1$ and if it is not zero, selects another egress node for ER expansion, i.e., intra-domain crankback. Alternatively, if $h_1$ expires, inter-domain crankback is done by sending a *PATH_ERR* message to the ingress node in the upstream domain (and $h_2$ decremented). The connection request is considered ultimately failed if both counters expire. To avoid random searches, a key innovation in [6] is the intelligent next-hop domain selection. Namely, each PCE now maintains a pre-computed multi-entry *distance vector table* that lists up to $K$ next-hop domains/egress links to each destination domain. This table can be generated using limited inter-domain path state, e.g., from BGP tables; see [6].

Now consider the extension of the above multi-domain crankback strategy for post-fault restoration. Again, it is assumed that failure end-point nodes quickly detect faults and can issue notification and takedown sequences (Section II.A). Upon receiving these notifications, source nodes can pursue two different restoration approaches:

**End-to-end (E2E):** Here the source node simply queries its PCE for another next-hop domain and sends a new *PATH* setup sequence towards the destination, i.e., clear route vector $\underline{\boldsymbol{R}}$, reset crankback counters $h_1$ and $h_2$, and note failed links to outgoing exclude route list, $\underline{\boldsymbol{X}}$. This request is then processed as per the regular working-mode setup procedures.

**Intermediate (IM):** Here the source node tries to preserve as much of the original path as possible, i.e., up to the domain preceding the failure. Hence the re-issued *PATH* message will only prune the route of the failed connection up to this domain and then initiate crankback processing from there, i.e. using same sequence as above-detailed working mode setup.

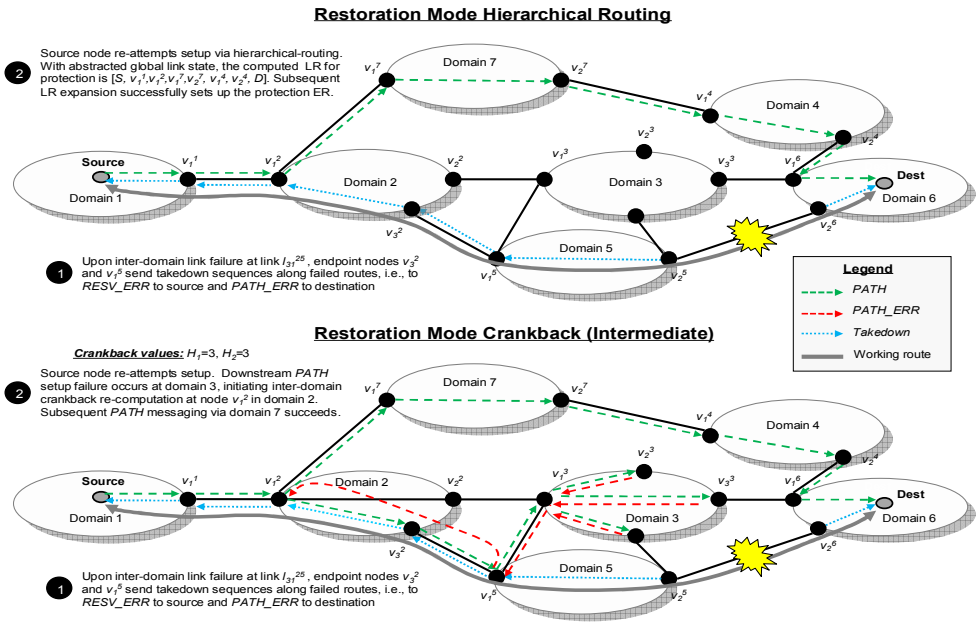


Figure 1: Joint intra/inter-domain crankback scheme for working and (intermediate) restoration mode

The case of IM post-fault crankback restoration is also shown in the lower part of Fig. 1, where the section of original path from source node to $v_1^5$ is preserved at first. After $H_1$=3 failed intra-domain crankback attempts in domain 3, inter-domain crankback is performed and finally the restored path succeeds via domain 2, 7 and 4 to the destination node.

## III. PERFORMANCE EVALUATION

The two multi-domain post-fault restoration schemes are tested using specially-developed *OPNET Modeler*™ models for a modified NSFNET topology with nodes replaced by domains. This topology has 16 domains averaging 7-10 nodes/domain and all intra/inter-domain links are set to 10 Gbps. Furthermore, all requests are generated between random nodes in random domains, and each run is averaged over 200,000 connections with mean holding times of 600 s (exponential) and variable inter-arrival times (as per desired load). The actual request sizes are further varied uniformly between 200 Mbps–1 Gbps in increments of 200 Mbps. Meanwhile, single link failures are limited to inter-domain links with exponential mean inter-arrival times of 12,000 s. Finally, $K$=5 next-hop domain entries are computed in the distance vector table for crankback restoration.

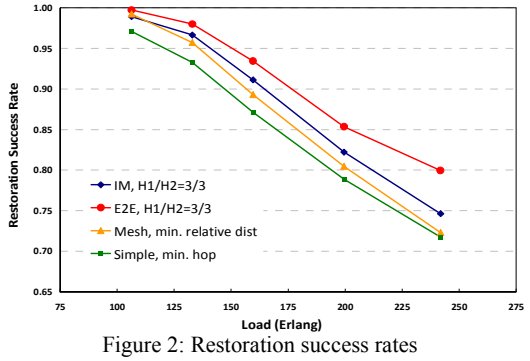


Figure 2: Restoration success rates

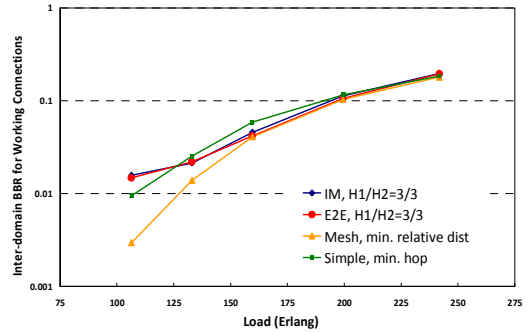


Figure 3: BBR for working-mode connection only

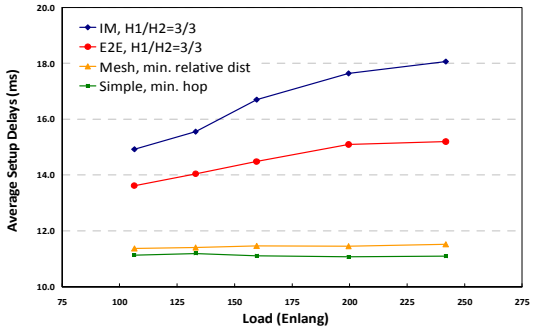


Figure 4: Average setup delay for protection paths

Post-fault restoration is first evaluated by measuring recovery success rates for hierarchical routing (simple-node, full-mesh abstraction) and signaling crankback (IM, E2E recovery) in Fig. 2. For signaling crankback, the intra/inter-domain crankback counters are set to $H_1$=$H_2$=3 as these values are shown to provide good inter-domain *bandwidth blocking rate* (BBR) reduction in working mode operation; see [6]. The results indicate that crankback schemes give better recovery than both forms of hierarchical routing, particularly at higher loads. This is likely due to the fact that crankback strategies allow multiple retries and hence can better handle setup failures during post-fault race conditions. In addition, E2E crankback always outperforms IM crankback as it tends to distribute attempts across the whole network (and not concentrate them to domains immediately prior to the failed links). Overall, these results show very good multi-domain recovery, i.e., over 70% at even high loads.

Generally, post-fault recovery may affect the performance of regular working-mode operation as resource usages may increase after failure events. To gauge this effect, Fig. 3 plots the BBR for working-only connections for all schemes. Overall, these results show that signaling crankback strategies yield blocking regimes in between simple node (upper bound) and full-mesh (lower bound) topology abstraction. As a result, it can be stated that improved signaling restoration comes at the expense of slightly-increased blocking of regular working connection requests (versus full mesh hierarchical routing).

Finally, recovery delays are also measured in Fig. 4. As expected, the two hierarchical routing schemes have notably lower recovery times, albeit these gains come at the expense of sizable increases in inter-domain routing overheads (not shown). In addition, the E2E crankback scheme gives notably lower delay than the IM scheme (about 17-25%), as it generally yields fewer retries and lower resource contention.

## IV. CONCLUSIONS

This paper proposes two enhanced solutions for post-fault restoration in multi-domain networks, i.e., hierarchical routing and signaling crankback. These schemes extend upon respective working-mode strategies by adapting them for link failure recovery. Overall, these results show that signaling crankback strategies yield the best performance in terms of recovery rates, with a slight increase in restoration delay.

## REFERENCES


[1] P. Cholda, *et al*, "A Survey of Resilience Differentiation Frameworks in Communication Networks", *IEEE Communications Surveys & Tutorials*, 4th Quarter 2007.

[2] N. Ghani, *et al*, "Control Plane Design in Multidomain/Multilayer Optical Networks," *IEEE Communications Mag.*, Vol. 46, No. 6, June 2008, pp. 78-87.

[3] T. Takeda, *et al*, "Analysis of Inter-Domain Label Switched Path (LSP) Recovery," *IETF Internet draft-ietf-ccamp-inter-domain-recovery-analysis-02.txt*, September 2007.

[4] A. Sprintson, *et al*, "Reliable Routing with QoS Guarantees for Multi-Domain IP/MPLS Networks," *IEEE INFOCOM 2007*, Alaska, May 2007.

[5] Q. Liu, *et al*, "Distributed Grooming in Multi-Domain IP/MPLS-DWDM Networks", *IEEE GLOBECOM 2009*, Hawaii, Nov. 2009.

[6] F. Xu, *et al*, "Enhanced Crankback Signaling for Multi-domain Traffic Engineering", *IEEE ICC 2010*, South Africa, May 2010.