9-10-2010

Multi-domain crankback operation for IP/MPLS & DWDM networks

Fareena Saqib

Fareena Saqib
Candidate Name

Electrical and Computer Engineering
Department

This thesis is approved, and it is acceptable in quality and form for publication on microfilm:

Approved by the Thesis Committee:

Dr Nasir Ghani Chairperson

Dr Jim Plusequallic

Dr Payman Zarkesh-Ha

Accepted:

Dean, Graduate School

Date

MULTI-DOMAIN CRANKBACK OPERATION FOR IP/MPLS & DWDM NETWORKS

BY

FAREENA SAQIB

BACHELOR OF INFORMATION TECHNOLOGY

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

Electrical Engineering

The University of New Mexico
Albuquerque, New Mexico

July, 2010

DEDICATION

This thesis is dedicated to my family, for all the love, support, and the many sacrifices made.

ACKNOWLEDGMENTS

First of all I wish to offer my gratitude to Almighty God for the Blessings bestowed upon me.

I would like to express my sincerest gratitude and indebtedness to my advisor and supervisor Dr. Nasir Ghani for his untiring assistance, timely guidance, encouragement and creativity at every stage of this study. Working with him has been a true privilege, and I have benefited tremendously from his knowledge of science, both in depth and broadness, and his enthusiasm in supporting me to carry out and complete the research work. He is a role model that I can always look up to.

My deepest appreciation and thanks are extended to Dr. Wennie Shu, Dr. Jim Plusquellic, and Dr. Payman Zarkesh-Ha for their continuous support, encouragement and valuable suggestions. I am also grateful to all the Optical Networks group members in the Electrical and Computer Engineering Department at the University of New Mexico as well as my post-graduate colleagues for their assistance, motivation and valuable comments to improve the dissertation. I am especially grateful to Mostafa Esmaeili for his dedicated assistance, guidance and professional contributions.

It would have not been possible for me to complete the research project without the motivation that my family members provided. Their constant reassurance and tireless optimism provided me the impetus to find ways out of seeming dead-ends.

It is almost impossible to make note of all those, whose inspirations have been vital in the completion of this dissertation. I am grateful to all of them.

ENHANCED CRANKBACK SIGNALING FOR MULTI-DOMAIN IP/MPLS NETWORKS USING STANDARD RSVP-TE PROTOCOL

BY

FAREENA SAQIB

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

Electrical Engineering

The University of New Mexico
Albuquerque, New Mexico

July, 2010

ENHANCED CRANKBACK SIGNALING FOR MULTI-DOMAIN IP/MPLS NETWORKS USING STANDARD RSVP-TE PROTOCOL

by

Fareena Saqib

B.I.T., National University of Sciences & Technology (NUST), 2007
M.S., Computer Engineering, University of New Mexico, 2010

ABSTRACT

Network carriers and operators have built and deployed a very wide range of networking technologies to meet their customers' needs. These include ultra scalable fibre-optic backbone networks based upon *dense wavelength division multiplexing* (DWDM) solutions as well as advanced layer 2/3 IP *multiprotocol label switching* (MPLS) and Ethernet technologies as well. A range of networking control protocols has also been developed to implement service provisioning and management across these networks.

As these infrastructures have been deployed, a range of new challenges have started to emerge. In particular, a major issue is that of provisioning connection services between networks running across different *domain* boundaries, e.g., administrative geographic, commercial, etc. As a result, many carriers are keenly interested in the design of multi-domain provisioning solutions and algorithms. Nevertheless, to date most such efforts have only looked at pre-configured, i.e., static, inter-domain route computation or more complex solutions based upon hierarchical routing. As such there is significant

scope in developing more scalable and simplified multi-domain provisioning solutions. Moreover, it is here that *crankback signaling* offers much promise.

Crankback makes use of active messaging techniques to compute routes in an iterative manner and avoid problematic resource-deficient links. However very few multi-domain crankback schemes have been proposed, leaving much room for further investigation. Along these lines, this thesis proposes crankback signaling solution for multi-domain IP/MPLS and DWDM network operation. The scheme uses a joint intra/inter-domain signaling strategy and is fully-compatible with the standardized *resource reservation* (RSVP-TE) protocol. Furthermore, the proposed solution also implements and advanced next-hop domain selection strategy to drive the overall crankback process. Finally the whole framework assumes realistic settings in which individual domains have full internal visibility via link-state routing protocols, e.g., open *shortest path first traffic engineering* (OSPF-TE), but limited "next-hop" inter-domain visibility, e.g., as provided by inter-area or *inter-autonomous system* (AS) routing protocols.

The performance of the proposed crankback solution is studied using software-based discrete event simulation. First, a range of multi-domain topologies are built and tested. Next, detailed simulation runs are conducted for a range of scenarios. Overall, the findings show that the proposed crankback solution is very competitive with hierarchical routing, in many cases even outperforming full mesh abstraction. Moreover the scheme maintains acceptable signaling overheads (owing to it dual inter/intra domain crankback design) and also outperforms existing multi-domain crankback algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AS	Autonomous system
ATM	Asynchronous transfer mode
BGP	Border gateway protocol
CSPF	Constrained shortest path first
DES	Discrete event simulation
DWDM	Dense wavelength division multiplexing
EGP	Exterior gateway protocol
GMPLS	Generalized multi-protocol label switching
IAT	Inter-arrival time
IGP	Interior gateway protocol
IP	Internet Protocol
ITU-T	Telecommunication Standardization Sector of the ITU
LAN	Local area networks
LPCS	Lightpath circuit switching
MAH	Minimum average hop
MAC	Minimum average cost
MPLS	Multiprotocol label switching
NSFNET	National Science Foundation network
OC	Optical carrier
OSPF	Open shortest path first
PAC	Protection at connection

PAL	Protection at lightpath
PCC	Path computation client
PCE	Path computation element
PL	Private line
QoS	Quality of service
RWA	Routing and wavelength assignment
SP	Shared protection
TE	Traffic engineering
VCAT	Virtual concatenation

Chapter 1

INTRODUCTION

The last two decades have seen tremendous progress in networking technologies. Here, the traditional "best-effort" paradigms of Internet networking service have now been replaced by full-blown *quality of service* (QoS) provisions for multiple service types, e.g., data, voice, video, etc. For example at the IP (Layer 3) level, new *multi-protocol label switching* (MPLS) technologies [1] have been introduced to support direct circuit setup between router nodes. As a result, network carrier can now achieve advances *traffic engineering* (TE) capabilities over then service backbones, improving vastly upon earlier hop-based routing setups.

Meanwhile there have also been many advances at the lower fiber-optic level, i.e., Layer1. Most notably, new *dense wavelength division multiplexing* (DWDM) technologies [2] have been developed to carry multiple signals over a single fibre using separate wavelength frequencies. Current DWDM systems can easily support over 100 wavelengths per fiber, giving over 1 terabit/sec capacity. Moreover, advanced optical add-drop and switching technologies have also ushered in new lightpath current routing capabilities, i.e., allowing a wavelength channel to be routed across a network of optical switches with little/no backbone processing. Finally, the MPLS framework has also been

extended to support these newer optical technologies, i.e., termed as the *generalized MPLS* (GMPLS) framework [3].

1.1 Background

As the above techniques have been deployed, network provisioning issues have received much focus. Namely a wide range of *constraint-based* routing solutions have been proposed for IP/MPLS networks [4]. Similarly, many studies have also been done for lightpath circuit *routing and wavelength assignment* (RWA) [5] in optical DWDM networks. However most of these efforts have only focused on single "domain" settings in which a provisioning entity has complete "network-wide" topology/resource views, e.g., single link-state routing domain [6]. Clearly, as user demands grow there is now a strong desire to achieve TE provisioning across *multiple* domains, both at the IP/MPLS and optical DWDM layers. Owing to obvious scalability and confidentiality concerns here [7],[8] it is clear that this must be achieved in a distributed, decentralized manner.

To address multi domain provisioning challenges, a diverse set of provisions have emerged to help improve multi-domain TE support, both at the IP/MPLS and underlying optical GMPLS layers. On the standards side, many ubiquitous routing protocols already provide varying levels of inter-domain visibility, e.g., next-hop/path-vector dissemination in *exterior gateway protocol* (EGP) [7] and hierarchical link-state dissemination in two-level *open-shortest-path-first* (OSPF-TE) [9]. Furthermore, the new IETF *path*

computation element (PCE) [10] framework also defines a comprehensive framework for multi-domain path computation and TE.

Meanwhile on the research side, a host of multi-domain TE schemes have been studied, see survey in [7] and Chapter 2. A key focus here is to address the tradeoff between inter-domain visibility and control plane complexity (i.e., dissemination, computation). For example, some have developed hierarchical link-state routing solutions to increase inter-domain visibility. The major contributions here are graph-theoretic topology abstractions for compressing domain-level state in IP/MPLS and DWDM networks. However, even though hierarchical routing delivers good blocking performance, associated routing overheads are very high, i.e., low scalability across large networks. Hence these schemes will likely be problematic in real-world settings where carriers tend to prefer EGP distance/path-vector protocols, e.g., *border gateway protocol* (BGP) variants. Nevertheless, these latter protocols only provide next-hop domain and end-point reachability state and most operational versions do not support any QoS parameters, e.g., delay, bandwidth, etc. As a result, hierarchical routing solutions do not represent a complete framework for all multi-domain provisioning scenarios.

1.2 Motivation

In light of the above, there is growing need to develop scheme to provision guaranteed bandwidth connections across multiple IP/MPLS and/or optical DWDM domains. Ideally, these schemes should yield effective provisioning and high scalability

[7],[8]. Along these lines crankback signaling schemes [11] offer a very promising approach for developing new solutions for the multi-domain TE. Namely, it is envisioned that these resultant schemes will potentially yield very good performance gains (in terms of blocking) at the same time as reducing overheads. However even though some crankback schemes have been studied [12]-[15], most of these strategies pursue more basic "exhaustive" search methodologies and hence entail significant signaling overheads. Moreover, none of these solutions have been gauged against alternate hierarchical routing schemes. Along these lines the focus of this thesis is to study the design of advanced crank back strategies for multi domain networks.

1.3 Problem Statement

This thesis focuses on the design of multi-domain crankback operation (MCO) for IP and optical DWDM networks. These solutions are also gauged against competing "global" hierarchical routing schemes.

1.4 Scope

The thesis focuses on the design of the distributed TE algorithms for multi-domain networks. The proposed solution addresses realistic scenarios where individual domains have full internal visibility via link-state routing, e.g., OSPF-TE protocol [9], but generally limited "next-hop" inter-domain visibility, e.g., as provided by inter-area or inter-domain routing protocols such as hierarchical OSPF or BGP. All evaluation is done

using the *OPNET Modeler*^{TM} Tool and more detailed analytical studies are left for future study.

1.5 Research Approach

To achieve the project aim, the work has focused on three key tasks. First, a detailed survey conducted on the existing literature in the multi-domain networking field. The next task focuses on the design of novel solutions for intra- and inter-domain crankback. Finally the third task addresses the coding and evaluation of these schemes using the *OPNET Modeler*^{TM} tool. Various best networks and scenarios are built and performance verified and gauged versus hierarchical routing.

1.6 Thesis Outline

The thesis is organized as follows. First Chapter 2 presents a survey of the latest work on multi-domain TE provisioning, including standards- and research-based activities. Next, Chapter 3 details the proposed enhanced intra/inter-domain crankback signalling solution. Chapter 4 then evaluates the simulation design and introduces the key performance evaluation matrices in the study. Detailed performance analysis results are then presented in Chapter 5. The results compared versus those from counterpart hierarchical inter-domain routing schemes. Finally, conclusions and future research directions are highlighted in Chapter 6.

1.7 Thesis Timelines

A Gantt timeline chart is shown in Figure 1.1 to summarize the key tasks for this thesis.

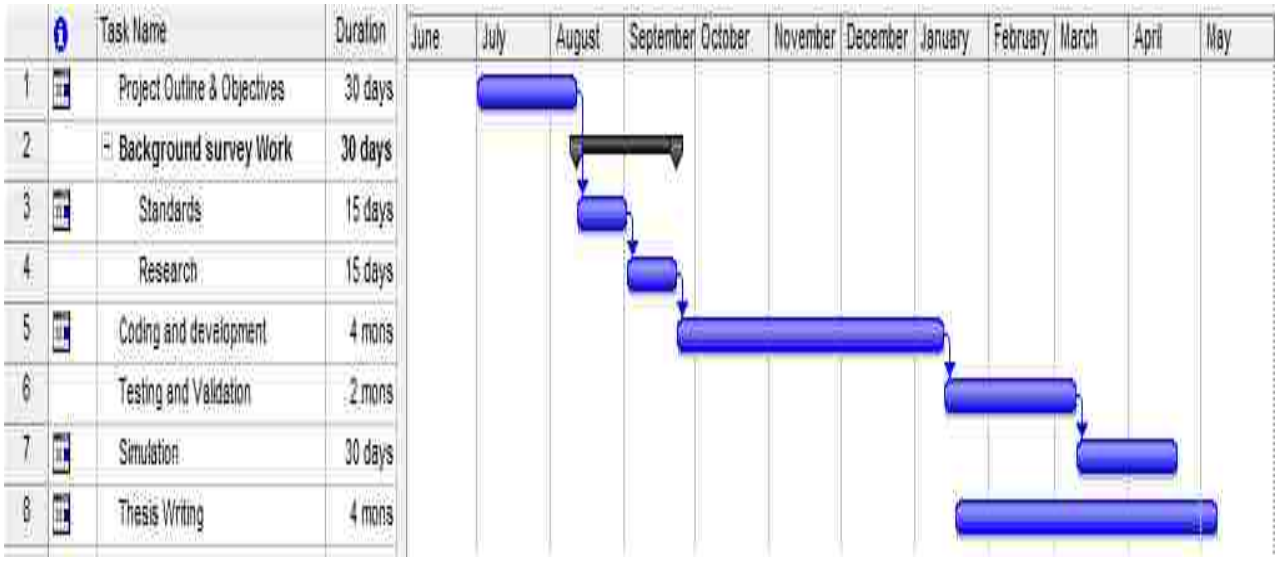

Figure1.1: Timeline of thesis work

Chapter 2

LITERATURE REVIEW

A review of the published literature is now presented with the goal of summarizing the latest standards and research work done in the broader area of multi-domain networking. Indeed these existing contributions encompass a wide range of efforts relating to IP/IMPLS and optical DWDM technologies and are now presented.

2.1 Multi-Domain Optical Networking Standards

A range of multi-domain networking standards have been developed by the *International Telecommunication Union* (ITU-T), the *Optical Internetworking Forum* (OIF), and the *Internet Engineering Task Force* (IETF) standardization bodies [3]. These are now detailed and a summary is also presented in Table 2.1

2.1.1: International Telecommunication Union (ITU-T)

The ITU-T *automatically switched transport network* (ASTN) framework [3],[7], formally termed as G.ason, presents one of the most well-defined set of frameworks for multi-domain operation. Specifically targeting optical transport networks, the ASTN

solution is based upon a hierarchical set up of *routing areas* (RA). Namely, a RA at the lowest hierarchical level represents a domain comprised of "physical" nodes and links, whereas the RA's at higher levels represents multiple "abstract" nodes and links. The state information from these abstractions can be distributed between domains to help improve "global" visibility levels.

International Telecommunication Union (ITU-T)	➢ ASTN framework: Hierarchical routing ➢ Call setup/ release/maintenance
Optical Internetworking Forum (OIF) Standards	➢ UNI: Client-network signal protocol ➢ NNI: Network-network protocol for inter-domain signaling and routing.
Internet Engineering Task Force (IETF) Standards	➢ OSPF-TE: Two level link-state routing protocol. ➢ BGP: Inter AS networking exchange protocol. ➢ RSVP-TE: Resource reservation protocol for setup signaling with inter-domain support. ➢ PCE: Path computation standard, functions with varying levels of inter-domain visibility.

Table 2.1: Summary of multi-domain standards

Overall, the ASTN framework bears some resemblance to the earlier routing standards developed for *asynchronous transfer mode* (ATM) technology [1]. Namely, ATM also defines a hierarchical design comprising of peer groups as part of its *private network-to-network interface* (PNNI) protocol [16]. However, the ASTN framework

further defines additional component groups to set up, maintain and release connections, and also provides standardized integration with IETF routing protocols. Interested readers are referred to [3] for more details on this framework.

2.1.2: Optical Internetworking Forum (OIF)

The OIF, as per its name, is more focused on *inter-networking* issues i.e., client-network and network-network. Along these lines it has tabled two protocols, namely the *user network interface* (UNI) [17] and *network-network interface* (NNI) [18]. First of all, the UNI defines a signaling protocol for clients to request and release "optical" connections from carrier domains running SONET/SDH or DWDM technology layers. The UNI protocol is based upon an *overlay* model design in which resource or topology information is not shared with clients. Overall, the initial UNI 1.0 standard was ratified almost 10 years ago and the newer UNI 2.0 provides further improvements for enhanced security and "hitless" bandwidth modification, see [17].

Conversely the OIF NNI protocol [7] implements inter-domain interconnection. Namely, NNI features are defined to support crucial address (reachability) information exchange as well as limited resource information exchange between domains. Furthermore, two NNI variants are defined, i.e., *internal-NNI* (I-NNI) and *external-NNI* (E-NNI) [18]. In particular, the E-NNI standard interfaces multiple vendor domains together and provides support for hierarchical routing by adapting the existing "two-layer" IETF OSPF-TE protocols. The overall goal here is to provide sufficient state

information between domain boundaries in order to automate connection setups across multi-technology regions.

2.1.3: Internet Engineering Task Force (IETF)

The IETF has developed perhaps the widest range of protocols standards for multi-domain/multi-AS networks. By many accounts, these architectures have become ubiquitous to most carriers and provide a rich set of capabilities. Foremost, at the routing level the IETF has standardized its comprehensive *exterior gateway protocol* (EGP) framework [7]. Here the most notable offering is the *border gateway protocol* (BGP) which provides inter-AS reachability exchange. However BGP is generally not sufficient for higher-end QoS provisioning needs, as it does not provide complete link state information exchange. Although various "QoS-enabled" BGP extensions have been developed, these are not well deployed as of today [6]. As a result, IETF has also extended its well-known OSPF protocol to provide new "QoS-capable" TE extensions, under its GMPLS framework [3]. Moreover, since OSPF-TE provides an additional level of routing, i.e., two levels of hierarchy [9], it can also be applied to multi-domains settings.

The IP/MPLS framework also supports a well-defined signaling protocol to setup connections, i.e., the *resource reservation-traffic engineering* (RSVP-TE) protocol [3]. In terms of multi-domain support, this standard offers some key features. Foremost is its ability to expand partial or *"loose" routes* (LR), either in a hierarchical or domain-domain manner. RSVP-TE also defines mechanisms to set up connections across domain

boundaries, i.e., contiguous, stitched, or nested [8]. Moreover, this protocol has recently been augmented to support crankback operation as well, a key facilitating provision for the research work in this thesis, see [11].

Finally, the IETF has also introduced a complimentary (inter-domain) path computation framework under its *path computation element* (PCE) standard [10]. The key goal here is to formally decouple TE path computation from signaling and routing operations. Namely, network domains are now provisioned with one or more logical PCE entities, i.e., in a standalone or co-located manner, that communicate with *path communication clients* (PCC) to resolve connection paths. Here, all PCC-PCE communication is also performed by a new *PCE protocol* (PCEP). These PCE entities have local domain resource databases and can function with varying levels of inter-domain visibility, e.g., low visibility in inter-carrier settings and high visibility in more trusted intra-carrier settings. Along these lines, two distributed path computation strategies are also outlined here, i.e., *per-domain* and *PCE-based* [10],[13]. There former computes paths in a *"domain-domain"* manner and is geared towards networks with lower visibility, i.e., no hierarchical routing support, hence the PCE's must iteratively compute path segments to the destination domain. Meanwhile the latter assumes enhanced inter-domain visibility and makes use of available inter-domain resource state information. Moreover the PCE framework also allows policy control at the domain boundaries.

2.2 Research Survey

Concurrent with the advances in multi-domain networking standards, a range of research studies have also emerged. These include studies on multi-domain packet-switched IP/MPLS networks, multi-domain optical DWDM networks, and survivability. The application of crankback signaling across domains has also received attention of late. The key contributors in these areas are now surveyed and also summarized in Table 2-2.

Multi-Domain IP/MPLS Routing Networks	➢ Range of studies on topology abstract actions for bandwidth and delay links ➢ Hierarchical and distance vector routing protocol analyses
Multi-Domain DWDM Routing Networks	➢ Studies on DWDM topology abstraction schemes. ➢ Extensions for survivability
Crankback Signaling	➢ Recent studies on "per-domain" and exhaustive crankback algorithms.

Table 2.2: Summary of multi-domain research studies

2.2.1 Multi-Domain IP/MPLS Routing

One of the key challenges in multi-domain networking is handling the reduced "visibility" between domains, i.e., links, nodes, resource levels. Along these lines, a host

of studies have been done in the area of *topology abstraction* or *topology aggregation* (TA) [7],[19]-[21]. Namely, these schemes use graph transformations to condense resource state via virtual graphs with fewer abstract vertices and edges. Typically, this is done by a designated domain-level entity, e.g., an ASTN RA controller [3] or domain PCE [10], which then propagates the abstract link state to other domains to build a global *aggregated graph*, i.e., via a hierarchical routing protocols such as OSPF-TE. This work has its origins in ATM technology, where various studies have proposed domain summarization for the PNNI protocol, see [16]. Overall, these earlier efforts have revealed good benefits from state reduction.

Extending the above work to IP packet-switched networks, [19] also proposed various topology abstraction TA solutions using star, mesh, tree, and spanner graphs. This effort also considered the interactions between the abstractions and various other factors such as routing overhead frequency reduction and different path computation algorithm. In addition, two other aggregation schemes (hybrid aggregation and weighted aggregation with protocol overhead similar to conventional star aggregation) were also devised. The results here showed lower bandwidth rejection rates for hybrid aggregation as compared to weighted aggregation, and were similar to full-mesh aggregation performance. The study also indicated strong improvements in routing scalability and route fluctuation reduction with the various schemes. Meanwhile, further work by [20] extended the above to incorporate delays into the abstraction formulation. Specifically, novel bandwidth/delay abstraction techniques were studied for directed graphs by leveraging information-theoretic and line-segmentation techniques. Overall, the results

showed good gains with aggregation yielding higher success rates and lower crankbank message loads.

Meanwhile, [21] also developed some *source-oriented* abstractions for efficient QoS-based routing in scalable networks. The goal here was to eliminate redundancy in the advertised state information by keeping in perspective the relevance of the information for path selection. Namely, three specific schemes were developed and evaluated including *unified quasi-stars*, *source-oriented simple node,* and *source-oriented star*. These solutions achieved different trade-offs between compaction and accuracy, and the work also proposed two new approaches for computing the weights of "logical links" with multiple QoS parameters i.e., to support multi-parameter QoS provisioning (see [21] for details). Extensive simulations for sparse and dense topologies here under static/dynamic scenarios were also performed and the results revealed that the source-oriented versions of the simple-node and star schemes showed better results than their conventional non-source-oriented counterparts. It was also noted that increasing routing update intervals resulted in more deleterious impacts on complex abstraction schemes (i.e., full-mesh) versus lossy schemes (i.e., simple node). This observation is due to the fact that an accurate state advertisement gradually loses its value as the routing update interval increases.

More recently, researcher has also applied topology abstraction to multi-domain survivability. Namely, [22] recently outlined an advanced solution to extract domain diversity state at the abstract graph level using *Surballe`s* algorithm. A distributed routing algorithm was also defined to leverage this specialized aggregated representation and to

compute two disjoint (primary, backup) QoS paths across domains, i.e., dedicated protection. However, the generated state was found to be quadratic in nature (posing high overhead complexity) and detailed performance evaluation studies were not presented, i.e., only mathematical proofs.

2.2.2 Multi-Domain DWDM Routing Networks

Extending upon the above studies for IP/MPLS settings, a host of DWDM-based topology abstraction schemes have also been proposed. The key goal here is to summarize DWDM node and link state information (i.e., wavelengths, converters) which is notably different from IP/MPLS link state (i.e., bandwidth, delay). Along these lines [23] presented a theoretical study of partial information models for domains with border node conversion. Here, lightpath selection was modeled as a *Bayesian* (probabilistic) decision and the findings showed that scalable information models achieve a good trade-off with loss (Bayes error rate). Although this study gave promising results the treatment was largely only theoretical and focused on bus topologies— rather unrealistic representations of DWDM mesh-domains. Moreover, *inter-domain* routing and RWA algorithms were not studied.

Next, [24] proposed a basic simple-node abstraction scheme for DWDM networks with a focus on "all-optical" networks. Although these schemes yielded good provisioning efficiency, this treatment did not address wavelength conversion–a critical necessity at domain boundaries which must perform regeneration and bit-level *service level agreement* (SLA) monitoring. Building upon this idea, [25] proposed a more

comprehensive multi-domain DWDM topology abstraction framework using simple-mode, full-mesh, and star abstractions. These algorithms also included further provisions for wavelength conversion, and related inter-domain RWA schemes were also detailed using skeleton-path computation/expansion. Overall, results showed the lowest blocking with the full-mesh scheme, albeit routing overheads were significantly higher, i.e., 3-4 times more than simple node.

Various other multi-domain DWDM studies have also been conducted as well. For example [26] proposed a domain-by-domain RWA scheme in which gateways maintain border alternate routes across all-optical and opto-electronic networks. Results showed good setup success rates, although path dissemination issues were not studied. Meanwhile, [27] studied a solution for RWA across a "multi-segment" DWDM networks. Here a graph-based heuristic method was used to transform the network into a multi-granularity graph and three path selection schemes were proposed, i.e., *end-to-end* (E2E), *concatenated shortest path* (CSP), and *hierarchical routing* (HIR). Namely the E2E scheme assumes a flat globalized graph, whereas the HIR scheme assumes a hierarchical graph with segments summarized as nodes, and finally the CSP scheme uses local information for segment-by-segment routing. Results for a specialized mesh-torus topology showed significant blocking reduction with the E2E and CSP schemes. The work did not, however, study associated intra or inter-domain routing overheads.

Finally, some work has also been done to extend topology abstraction for survivable DWDM networks. Namely [28] proposed multi-domain shared path protection schemes using aggregated full-mesh topology abstractions. These algorithms performed

sequential working/backup path computation and were tested to show very close performance to idealized "flat" routing. However, no details were presented on the actual virtual link computation algorithms and/or inter-domain routing overheads. Recently this work was also extended to consider back-up path re-optimization, yielding moderate blocking reductions, i.e., 5% range, see [29].

2.2.3 Crankback Signaling

The overall aim of crankback signaling is to use messaging (i.e., RSVP-based) to iteratively search for valid/feasible path routes in a *per-domain* manner [8],[10]. Namely, ingress border nodes receiving egress setup messages select appropriate egress border nodes and try to signal and expand local routes across their domains to these nodes. Here, if setup signaling fails across a domain, crankback messaging is sent to an appropriate upstream node in order to re-compute an alternate downstream path sequence.

Now various studies have investigated crankback in multi-domain MPLS/GMPLS networks. For example [14] presented a *compute while switching* (CSW) scheme in which the ingress border nodes used per-domain computation and crankback to setup an initial route. After setup, data transmission is started along this initial route, but further crankback signaling also was initiated to search for more "*optimal*" routes, i.e., shorter hop counts. To achieve this, new extensions to the RSVP-TE protocol are proposed to carry the necessary signaling state. Overall, the results of this study showed very high setup success with the CWS scheme, on a par with global state. Nevertheless, this was expected as the scheme essentially mimicked an exhaustive search strategy. Moreover,

the signaling overheads of this scheme were not analyzed in the study and are expected to be quite high.

Meanwhile, [12] also defined a basic *per-domain* (PD) crankback scheme which probed egress domain nodes for traversal routes, and upon failure, notified upstream border nodes. Specifically here the next-hop domains were selected as those with the closest border node to the ingress border node (performing path expansion). However, results show somewhat higher request blocking rates and setup delays, particularly when compared to alternate PCE-based strategies utilizing pre-determined inter-domain routes. Finally, [15] also studied crankback to minimize end-to-end path delays in multi-domain settings. Namely two next-hop domain selection strategies were presented here. The first approach selected the next-hop as the "nearest" egress border node in the domain, whereas the other approach relied upon detailed inter-domain *round-trip time* (RTT) measurements, i.e., pre-computed global state. In general the latter heuristic tended to shown to yield slightly higher carried load and less crankbacks, although it requires adoption of a specialized coordinates system [15]. Overall, the above crankback solutions represent some good initial contributions. However, new innovations are clearly possible for multi-domain settings.

2.3 Motivation

In light of the above reported research, it is obvious that multi-domain traffic engineering is a very challenging problem area. Moreover, the application of crankback

signaling here offers a very promising avenue along which to develop new and improved solutions. However, even though some initial crankback studies have been done for multi-domain settings, there is still significant latitude for designing new and improved solutions, e.g., with more advanced next hop domain selection, improved intra/inter-domain crankback strategies, etc. Along these lines, this thesis proposes to study these possibilities in realistic MPLS/GMPLS multi-domain network settings.

Chapter 3

ENHANCED CRANKBACK SOLUTION

An enhanced multi-domain crankback solution is now presented in this chapter. The solution utilizes key components of the evolved IETF MPLS/GMPLS framework detailed in Chapter 2, including protocols for routing (OSPF-TE, BGP), signaling (RSVP-TE), and path computation (PCE). Namely all domains are assumed to run OSPF-TE, providing nodes with full link-state knowledge. Meanwhile, selected border gateway nodes are also assumed to run inter-domain BGP, providing limited path-level views of the "global" inter-domain topology. Finally, each domain is assumed to have to have at least one PCE entity which has full access to domain-level OSPF-TE state as well as inter-domain BGP path vector state. These PCE entities then operate in a distributed "*per-domain*" manner to help compute end-to-end routes.

Overall three key innovations are introduced in the proposed scheme to enhance multi-domain crankback operation, 1) dual intra/inter-domain crankback counters to limit signaling complexity/delay, 2) full crankback history tracking to improve the re-try process, and 3) intelligent per-domain selection. Details are now presented.

3.1 Setup Signaling Overview

Before detailing the proposed solution an overview of RSVP-TE signaling and crankback operation is presented. The basic RSVP-TE signaling protocol follows a backwards reservation model for setting up connections in MPLS networks, termed as *label switched paths* (LSP) [1]. Namely, source "*ingress*" switching/routing nodes first send *PATH* messages along a pre-defined connection route to determine resource availability levels along the end-to-end links, see Figure 3.1a. Here, each receiving node checks its outbound link for sufficient bandwidth resources, and if available, continues to propagate the *PATH* message by sending it to the next downstream node. If, however, resources are not available on a link at an intermediate node, the node terminates the forward propagation of the *PATH* message and instead sends a backward, i.e., *upstream, PATH-ERR* notification message to the source to indicate setup failure. This case is also shown in Figure 3.1b.

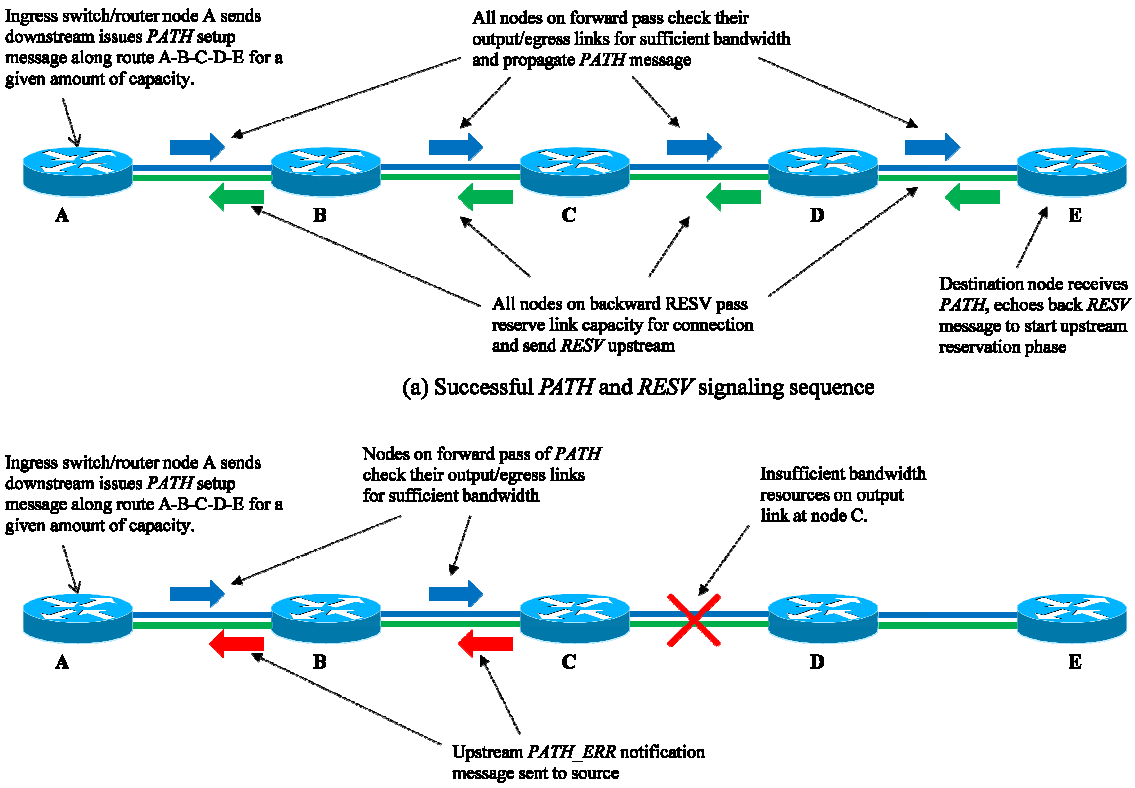

Figure 3.1: PATH and RESV signaling sequence

Now if the forward pass of the *RSVP-TE PATH* message completes all checks at all path links, the destination node initiates a backward reservation phase by sending a RSVP-TE *RESV* message back to the source node along the selected path. Namely, on this pass each receiving node actually performs resource reservation for the request by explicitly removing a free link bandwidth and assigning it to the connection, Figure 3.1a. When the *RESV* message successfully arrives back at the source node, the connection is termed as "established". For more details on RSVP-TE signaling protocol and message types, please refer to [3] and [11].

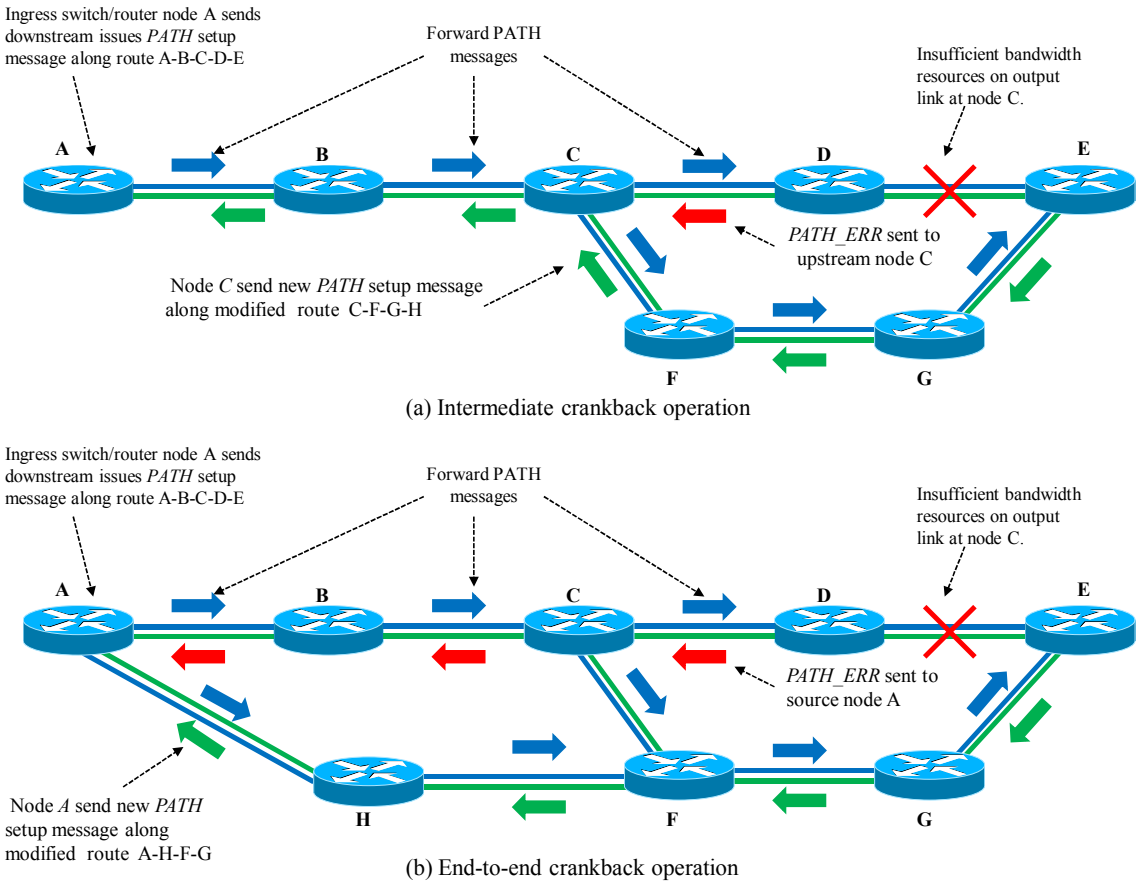

Figure 3.2: Crankback operation

Meanwhile, crankback is the process by which the above-detailed RESV-TE setup signaling mechanism is modified to handle link resource failure events. Overall, the basic aim of the crankback is to improve connection setup success rates, (i.e., reduce connection blocking rates) by acquiring real time information about any link resource failures which may occur, and effectively re-routing around these congestion points. Now recent extensions for crankback in the RSVP-TE protocol have been standardized in RFC 4920 [11]. Specifically this framework uses *PATH-ERR* messages to convey link resource failure information to intermediate upstream crankback points, i.e., not

necessarily source nodes. In response, various actions can be taken by these intermediate upstream nodes, e.g., re-attempting *PATH* signaling setup along other downstream routes that explicitly avoid the problematic links, cranking back and notifying the source node, or dropping/failing the setup request altogether, etc. Specifically to re-attempt *PATH* setup along a new route, an intermediate node simply discards the received *PATH ERR* message and generates a new downstream *PATH* message along a new sub-path route to the destination. This sub-path can be carefully re-routed using domain link-state databases and *PATH-ERR* failure history state to avoid problematic downstream nodes. Here, intermediate nodes can also maintain local crankback tables to share failed downstream link information between multiple user connections [11]. Finally, crankback counters can also be used to link the number of crankback reattempts. Note that in the inter-AS/inter-domain context, crankback re-routing can be done to a variety of upstream points, e.g., ingress border nodes/gateways (intermediate) or source nodes (end-to-end), These two cases are also illustrated in Figure 3.2a (intermediate) and 3.2b (end-to-end), see [11],[12] for more details.

3.2 Multi-Domain Crankback Operation

Using the above detailed crankback framework (in Section 3.1), the proposed solution is next presented. However before detailing the scheme, the requisite notation is introduced. First, consider a multi-domain network comprising of D domains, with the i-th domain having n^i nodes and b^i border/gateway nodes, $1 \leq i \leq D$. This network is modeled as a set of domain sub-graphs, $\boldsymbol{G}^i(\boldsymbol{V}^i,\boldsymbol{L}^i)$, $1 \leq i \leq D$, where $\boldsymbol{V}^i = \{ v^i_1, v^i_2, \ldots \}$ is the set of

domain nodes and $\boldsymbol{L}^i=\{\ l^{ii}_{jk}\ \}$ is is the set of *intra-domain* links in domain i ($1\leq i\leq D$, $1\leq j,k\leq n^i$), i.e., l^{ii}_{jk} is the link from v^i_j to v^i_k with available capacity c^{ii}_{jk}. A physical inter-domain link connecting border node v^i_k in domain i with border node v^j_m in domain j is further denoted as l^{ij}_{km} and has available capacity c^{ij}_{km}, $1\leq i,j\leq D$, $1\leq k\leq b^i$, $1\leq m\leq b^j$. Also, $\boldsymbol{B}^i$ denotes the set of border nodes in domain i. Now consider the relevant RSVP-TE message fields. The path route is given by a node vector, $\underline{\boldsymbol{R}}$. Meanwhile, other fields are also defined for crankback as per [11], and include an exclude link vector, $\underline{\boldsymbol{X}}$, to track crankback failure history as well as dual intra/inter-domain crankback counters, h_1 and h_2 (usage will be detailed shortly). Note that [11] only defines a single counter field but bit masking can be used to generate two "sub-counters".

An overview of per-domain computation is first given for the case of *non-crankback* operation, i.e., no resource request failures. Consider a source node fielding a request for x units of bandwidth to a destination node in another domain. This source queries its PCE to determine an egress link to the next-hop domain, e.g., using the *PCE-to-PCE protocol* [10]. The PCE then determines the next-hop domain to the destination domain (detailed in Section 3.3) and returns a domain egress border node/link to this domain. Note that this information also contains the *ingress* border node in the downstream domain. Upon receiving the PCE response, the source uses its local OSPF-TE database to compute an *explicit route* (ER) [1] to the specified egress border node. This step searches the k-shortest path sequences over the *intra-domain* feasible links (i.e., $c^{ii}_{jk}\geq x$) and chooses the one with the lowest "load-balancing" *cost*, i.e., individual link

costs inversely-proportional to free link capacity, i.e., $1/c^{ij}_{km}$. This method is used as it generally outperforms basic hop count routing, see [21],[25].

Granted that an ER path is found above, it is inserted in the path route vector, ***R***, and RSVP-TE *PATH* messaging is then initiated (along the expanded route) to the ingress border node in the next-hop domain. Here, each intermediate node checks for available bandwidth resources on its outbound link and pending availability, propagates the message downstream. The above procedure is repeated at all next-hop domain border nodes until the destination domain. When the *PATH* message finally arrives at the destination domain, the border node (or PCE) expands the ER to the destination. Upon receiving a fully-expanded *PATH* message, the destination initiates upstream reservation, i.e., by sending a *RESV* message.

Now consider the case of *PATH* processing failure, i.e., due to insufficient bandwidth resources along a route link. Leveraging the crankback framework for RSVP-TE signaling in [11], two strategies are chosen for implementation herein, i.e., *intra-domain* and *inter-domain*. Namely, the enhanced scheme defines dual crankback counters, i.e., h_1 and h_2, to limit the number of re-try attempts at the intra and inter-domain levels, respectively. Specifically, the above counters are initialized to pre-specified limit values (H_1 and H_2, respectively) in the initial *PATH* message and then decremented during crankback to limit excessive searching along longer and less resource-efficient paths. As such, these values effectively bound the number of intra and inter-domain crankback attempts to H_1H_2. Furthermore, crankback failure history is also tracked at both the intra/inter-domain levels.

Using the above counters, two key crankback operations are defined, i.e., *notification* and *re-computation*. The former refers to the (upstream) signaling procedures executed upon link resource failure at an intermediate node, whereas the latter refers to the actual re-routing procedure to select a new route. Now in general, resource signaling (*PATH* processing) failures can occur at *three* different types of nodes, i.e., domain ingress border nodes, domain egress border nodes, and interior nodes. However, in the proposed scheme, only the former performs *re-computation* whereas the latter two simply perform crankback *notification*. These steps are now detailed further in the following sub-sections.

```
if (insufficient resources on outbound link)
    Decrement intra-domain counter h1, extract route vector R and exclude link vector X from PATH
    Add failed outbound link to exclude route vector X
    Remove all nodes in route vector R up to ingress border node, i.e., prune failed intra-domain segment
    Generate PATH_ERR, copy h1, R, X fields and send to upstream ingress border node
```

Figure 3.3: Crankback notification algorithm (at local or egress border node)

Crankback Notification: Upstream notification is done when there is insufficient bandwidth at an intra-domain link (i.e., at an intra-domain node) or an inter-domain link (i.e., at an egress border node) on an already-expanded ER. This overall algorithm is shown in Figure 3.3. Namely, the *PATH* message is terminated and its appropriate fields updated and copied to an upstream *PATH_ERR* message to the domain's ingress border node. Specifically, the intra-domain counter h_1 is decremented and the failed link is noted. Note that if blocking occurs in the source domain, the *PATH_ERR* is sent back to the source.

Crankback Re-Computation: Meanwhile, path re-routing is done by ingress border nodes receiving a *PATH_ERR*. Note that for special case of a source domain (i.e., non-ingress border node), the receiving source node relays the *PATH_ERR* to its PCE for processing. The overall algorithm is summarized in Figure 3.4. Here, two types of crankback re-computations can be done. First consider "intra-domain" crankback. If the intra-domain h_1 counter has not expired in the received *PATH_ERR* message, another next-hop domain/egress border node is selected by the ingress border node (or PCE) for ER expansion. In particular, the exact sequence of next-hop domains tried is pre-computed to try successively longer inter-domain routes (i.e., via multi-entry distance vector table, detailed in Section 3.3). Now the enhanced scheme makes full use of crankback history to avoid any failed intra/inter-domain links. Primarily, all failed inter-domain links in ***X*** that egress from the domain are removed from consideration, i.e., only consider "non-failed" next-hop domain egress links. Additionally, all intra-domain links listed in the exclude link vector ***X*** are also removed from local ER computation. Note that the route vector ***R*** is also searched to make sure that an upstream domain is not traversed twice, i.e., no "domain-level" loops. Regardless, it still may not be possible to initiate/establish a domain-traversing route for various reasons, i.e., h_1 counter expired, LR expansion failure to selected egress node, or all egress border links in exclude link vector ***X***, etc. In these cases, the ingress border node must initiate a more globalized "inter-domain crankback" response via a *PATH_ERR* message to the ingress node in the upstream domain in

the *PATH* route vector ***R*** (or source node if upstream domain is source domain). To improve history tracking in this case, the ingress border node also inserts its own ingress link in the exclude route vector of the *PATH_ERR* message, i.e., in order to avoid future re-tries on this link. Note that "inter-domain crankback" is only initiated if the inter-domain crankback counter, h_2, is non-zero, otherwise the request is failed (i.e., *PATH_ERR* to source, Figure 3.2).

```
/* Attempt intra-domain re-routing */
 if (h1 not expired)
     Select next-hop domain/egress link using multi-entry distance vector table s.t. next-hop domain is  not in
    R and egress link is not in X
    if (next hop egress node found)
        Make copy of local network graph (via IGP database), prune all local failed links listed in X,
        compute new ER to egress border node
        if (LR expansion successful)
            Initiate PATH signaling to new egress node
             intra_domain_crankback_done=1;
/* Attempt inter-domain re-routing */
 if (!intra_domain_crankback_done & h2 not expired)
    Decrement inter-domain counter h2, extract route vector R and exclude route vector X from PATH
    Add ingress inter-domain link to exclude link vector X
    Remove all nodes in route vector R up to previous domain's ingress border node
    Copy h2, R, X fields, reset h1=H1, generate PATH_ERR and send to previous domain's ingress
    border node
    Else
     Copy h1, h2, R, X fields,generate PATH_ERR,send to source
```

Figure 3.4: Crankback re-computation algorithm (at domain ingress border node)

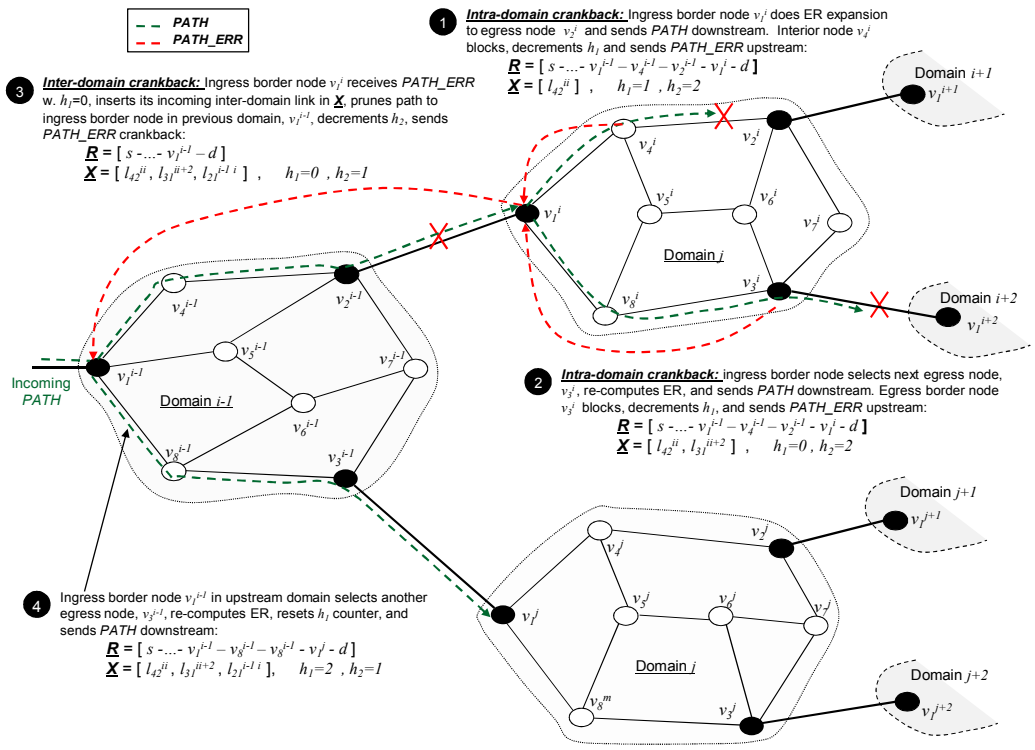

Figure 3.5: Enhanced intra/inter-domain crankback scheme (H_1=2, H_2=2)

An example of crankback notification is shown in Figure 3.5 for interior and egress border nodes (H_1, H_2=2). For example, consider bandwidth blocking on the link l^{ii}_{42}, i.e., step 1. Here, the interior node v_4^i prunes the route vector ***R*** to the domain ingress node, adds the blocked link to the exclude route vector ***X***, decrements the intra-domain counter h_1, and sends all this information back to the ingress node v_1^i via a *PATH_ERR* message. A similar procedure is also shown for blocking at the egress border node v_3^i (i.e., step 2, Figure 3.5). Sample crankback re-computation is also shown in Figure 3.5. For example when blocking initially occurs on link l^{ii}_{42}, the ingress border node v_1^i re-tries intra-domain path expansion to egress border node v_3^i. When this second intra-domain attempt fails at the egress link l^{ii+2}_{31}, ingress node v_1^i receives a *PATH_ERR*

with a zero h_1 counter. In response, it marks its ingress link $l^{i-1\ i}{}_{21}$ as failed, prunes the route to the ingress border node in previous domain *i-1*, i.e., node v_1^{i-1}, and sends a *PATH_ERR* message (step 3, Figure 3.5). The upstream ingress border node v_1^{i-1} decrements h_2, resets the h_1 counter to H_1, and then initiates a re-try to a new egress border node, v_3^{i-1} (step 4, Figure 3.5). Note that if the previous domain is the source domain, the *PATH_ERR* is simply sent to the source.

3.3 Next-Hop Domain Computation

As mentioned earlier, a key provision in the enhanced crankback scheme is the use of existing inter-domain state to improve the search process. This is achieved by pre-computing a *multi-entry* distance vector table at all domain border nodes (or PCE) to list up to K next-hop domains/egress links to each destination domain. Namely, at domain i, the k-th table entry to a destination domain j, $T^i(j,k)$, is computed as the egress inter-domain link (to the next-hop domain) on the k-th shortest "domain-level" hop-count path to domain j ($1 \le i, j \le D$, $i \ne j$, $1 \le k \le K$). Clearly the number of entries to a destination will be upper-bounded by the minimum of K and the maximum number of inter-domain links that egress from the domain.

Now consider the actual computation of this table at a border node (or PCE) in domain i, the algorithm for which is summarized in Figure 3.6. Here a "simple node" [19],[21],[25] view of the global topology is first derived, i.e., $\boldsymbol{H}(\boldsymbol{U},\boldsymbol{E})$, where $\boldsymbol{U}$ is the set of domains $\{\boldsymbol{G}^i\}$ reduced to vertices and $\boldsymbol{E}$ is the set of inter-domain links $\{l^{ij}_{km}\}$, $i \ne j$. At the inter-area level, this graph can be obtained from hierarchical OSPF link-state

databases whereas at the inter-AS level it can approximately be deduced from BGP path vector state (albeit not all inter-domain connectivity may be visible due to policy restrictions). An iterative shortest-path scheme is then used to compute multiple routes to all destination domains over $\boldsymbol{H}(\boldsymbol{U},\boldsymbol{E})$. Namely, the scheme basically loops over all destination domains $j \neq i$ (index j) and computes up to K next-hop egress links (index k) over a temporary copy of $\boldsymbol{H}(\boldsymbol{U},\boldsymbol{E})$, i.e., $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$. At the k-th iteration, the scheme computes the shortest "domain-level" hop-count path to the destination domain using $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$, and if found, stores the egress link from the source domain in $T^i(j,k)$. This link is then pruned from $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$ and the procedure repeated to compute the next shortest "domain-level" hop-count path. The procedure is terminated if all K entries are filled and/or the vertice for domain i in $\boldsymbol{H'}(\boldsymbol{U},\boldsymbol{E})$ becomes disconnected. Hence the next-hop domain selection procedure during crankback re-computation (as detailed in Section 3.2) simply searches these K table entries, $T^i(j,k)$, to a destination domain j in increasing order. This sequentially drives the crankback search along fixed "domain-level" sequences of increasing length, but with provisions to prune "failed" entries (in $\underline{\boldsymbol{X}}$). Overall, these entry tables will be relatively static if inter-domain topology changes are relatively infrequent.

Generate simple-node abstraction of global topology via EGP database information, i.e., $\boldsymbol{H}(\boldsymbol{U}, \boldsymbol{E})$

/ At domain i, loop across all possible destination domains */*

for *j = 1* to *D*

  if ($j \neq i$)

    Make temporary copy of graph $H(\boldsymbol{U},\boldsymbol{E})$, i.e., $H'(\boldsymbol{U},\boldsymbol{E})$

    / Compute up to K table entries */*

    for *k=1* to *K*

      Compute shortest-path from domain *i* to *j* in $H'(\boldsymbol{U},\boldsymbol{E})$

      if (shortest path route found)

        Save route line from domain *i* in *k-th* table entry $T^i(j,k)$, i.e., link from domain *i* vertice in $H'(\boldsymbol{U},\boldsymbol{E})$

        Prune above-selected link from $H'(\boldsymbol{U},\boldsymbol{E})$

        if (domain *i* becomes disconnected)

          break *k*-loop

Figure 3.6: Multi-entry distance vector table computation algorithm (at PCE)

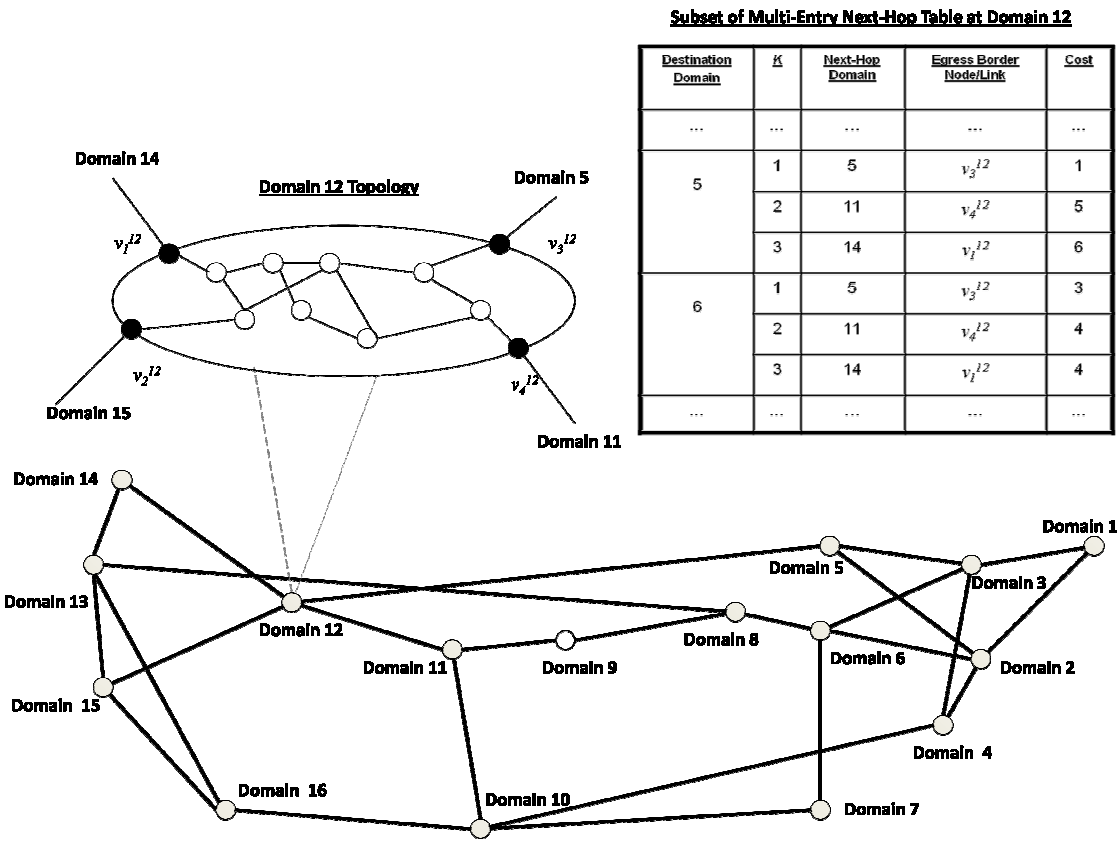

Figure 3.7: Multi-entry next-hop table

A sample computation for the next hop multiplexing distance vector table is presented in Figure 3.7. First, the overall "skeleton" topology of the global network is depicted, with domains represented as nodes. Next some sample table entries are shown for border nodes in domain 12 to destination domains 5 and 6. Namely, domain 12 has 4 border gateway nodes and its table lists K empirical paths to external domains. For example, consider destination domain 5. Here, the shortest path is clearly the single hop path emanating from node v_3^{12}. Hence the first table entry for destination domain 5 lists v_3^{12} as the egress node, the next hop domain as domain 5 itself, and a hop-count cost of 1. Meanwhile, the next shortest path to domain 5 is via egress border node v_4^{12} and along domains 11, 10, 4 and 3. Hence the second table entry for destination domain 5 lists v_4^{12} as the egress border node, with a next hop domain of 11, and a total hop count of 5. Similarly, other entries can also be computed to the other domains. Note that these tables can be periodically recomputed if there are any path vector updates, e.g., BGP updates.

3.4 DWDM Extension (GMPLS Networks)

Carefully note that all of the discussions in Section 3.2 have focused on IP/MPLS *bandwidth* reservation networks only. However, the proposed enhanced crankback scheme can also be modified to support optical DWDM networks. Namely, in these settings the key goal is to achieve lightpath RWA across multiple domains. Now clearly the RWA problem is very dependent upon the availability of wavelengths convertors in the network, both at the intra-and inter-domain levels see [5] and [25]. Along these lines, several assumptions are made to reflect realistic multi-domain DWDM network settings.

First of all, it is assumed that all network nodes are now *optical cross-connects* (OXC) [3] systems. Next, all interior (domain internal) nodes are assumed to be "all-optical" in design, i.e., perform no wavelength conversion. Finally all border OXC nodes are assumed to have full wavelength conversion capabilities on their inter-domain links only. This setup is chosen to reflect real-world settings in which all-optical "islands" are delineated by full *opto-electronic* (OE) conversion to support bit-level service monitoring between domains.

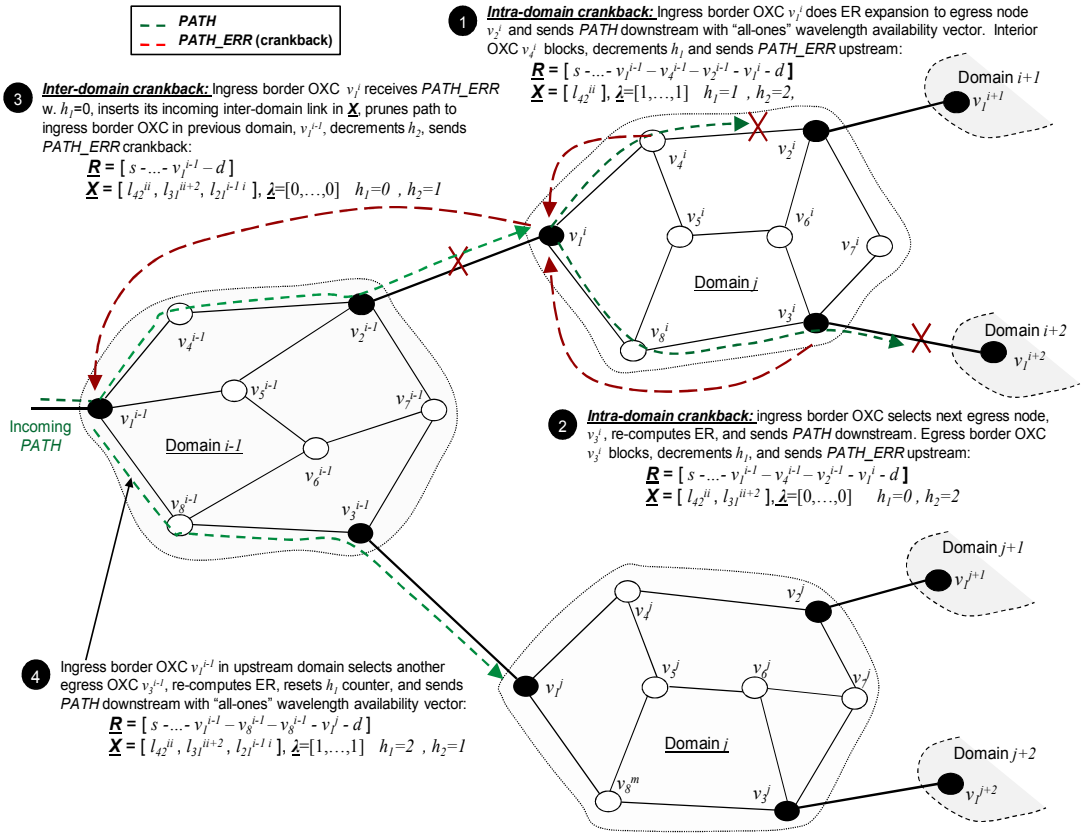

Figure 3.8: Enhanced crankback scheme for multi-domain lightpath RWA (H_1=2, H_2=2)

Now using the above framework, the proposed crankback framework can be extended for multi-domain lightpath RWA. Namely, the only updates required are the

inclusion of an bit-level wavelength availability vector, $\underline{\boldsymbol{\lambda}}$ =[01010...] in the RSVP-TE *PATH* and *PATH-ERR* messages and slight modifications to *PATH* message processing. Again, "per-domain" RWA is again done in an iterative manner starting at the source domain. Here the OXC (or domain ingress border OXC) first consults its PCE to determine the next-hop domain to the destination domain, i.e., identify next-hop domain and egress border OXC/link in the current domain using same next-hop multi-entry next hop table from Section 3.3. Upon receiving this information, the source OXC (or domain ingress OXC) uses its local routing database to compute a local lightpath route to the chosen egress border OXC. Namely, this intra-domain lightpath route is selected as the minimum hop feasible route, i.e., with at least one free wavelength. This path is then inserted into the route field, $\underline{\boldsymbol{R}}$, of a downstream *PATH* message. This message *PATH* also contains the crankback counters (h_1,h_2) and an "all-ones" wavelength availability vector i.e., ($\underline{\boldsymbol{\lambda}}$=[1,…,1]). The latter vector is then AND-ed with the available wavelength vectors of the intra-domain path links on the forward pass of the *PATH* message, i.e., in order to find an "all-optical" intra-domain path.

Now since wavelength conversion is done at domain border OXC nodes, the ingress border OXC nodes received the *PATH* message must also save the availability vectors from the previous domain in $\underline{\boldsymbol{R}}$ and then generate new "all-ones" $\underline{\boldsymbol{\lambda}}$ vector for downstream *PATH* processing. Note that actual wavelength selection for the lightpath segments is done during the upstream *RESV* signaling phase. In particular, *most-used* (MU) multi-domain $\underline{\boldsymbol{\lambda}}$ selection is used as it is shown to give lower blocking in both the intra [5] and inter-domain [25] contexts.

Finally crankback notification is now also done if there is no available wavelength at an intra-domain link (intra-domain OXC i.e., $\underline{\boldsymbol{\lambda}}$=[0,…,0]) or there is no available wavelength or converter at an inter-domain link (egress border OXC). Namely, here the *PATH* message is terminated and an upstream *PATH_ERR* crankback notification is sent to the domain's ingress border OXC. Subsequently, crankback re-computation also performs intra-domain RWA to alternate border egress OXC's, selected using the same next-hop table entries (Section 3.3).

The operation of the proposed crankback scheme in crankback procedure for multi-domain DWDM networks is also shown in Figure 3.8 for H_1, H_2. Namely, when wavelength blocking occurs on link l_{42}^{ii} (step 1, Figure 3.8), OXC v_4^i prunes the route vector $\underline{\boldsymbol{R}}$ to the domain ingress node v_1^i, adds the blocked link to $\underline{\boldsymbol{X}}$, and decrements h_1. This information is then sent to v_1^i via a *PATH_ERR*. The case of subsequent wavelength blocking at an egress border OXC link is also shown, i.e., at link $l_{31}^{i\,i+2}$ at node v_3^i (step 2, Figure 3.8). Further crankback re-computation is also shown here. Namely, when ingress OXC v_1^i receives a *PATH_ERR* with h_1=0, it notes ingress $l_{21}^{i-1\,i}$ as failed, prunes the route to the ingress border OXC in the prior domain, v_1^{i-1}, and sends a *PATH_ERR* to v_1^{i-1} (step 3, Figure 3.8). This upstream OXC then re-tries path expansion to a new egress border OXC, v_3^{i-1} (step 4, Figure 3.8).

Chapter 4

SIMULATION AND PERFORMANCE TOOLS

Introduction Software based network simulation is a widely-used means of evaluating network performance. This is particularly important given the growing complexity of modern networks and user services. In many cases, network simulation offers the only viable means of analyzing such complex systems in a realistic manner, i.e., as analytical modeling becomes too intractable. Along these lines, *discrete event simulation* (DES) [30] has emerged as a very popular technique in network analysis. This approach models network behaviours as a series of responses to events, e.g., such a connection requests arrivals, control messages, link failures, etc. These events are then sorted and queued in a time-increasing buffer, i.e., via event timestamp fields. The simulation engine then loops and processes these events in a sequential manner, further generating new events and/or removing/retiming existing events.

Now over the years, a wide range of network simulation softwares have been developed and some leading examples include *OPNET Modeler*^{TM}, *NS*, *NS2, OMNET++*, etc. However for this study the *OPNET Modeler*^{TM} tool is chosen as it provides the most complete set of features, i.e., *graphical user interface* (GUI), robust DES simulation/list processing routines, a wide range of packet definitions, etc. More importantly, this tool provides a full C/C++ interface to allow users to build and customize their network

models. Overall, this tool has gained very strong traction with many users, both within industry and academia.

4.1 Network Topologies

In order to ensure proper investigation of multi-domain crankback signaling performance, various realistic test topologies are first developed. Now given that there are really no "standard" multi-domain test topologies as such, it is necessary to design different types to cover a good range of realistic scenarios. As a result two network topologies are used here, namely a modified version of the ubiquitous NSFNET backbone and a specially-designed 10 domain topology. In particular, the former topology replaces each node in NSFNET with a domain of approximately 7-10 nodes, see, Figure 4.1. This results in a multi-domain network with 16 nodes and 25 inter-domain links, i.e., approximately 1.56 links/domain. Overall, this topology also has 50 border nodes which act as ingress and egress gateways for inter-domain requests. Conversely the 10 domain topology is shown in Figure 4.2 and has 25 inter-domain links and 2.5 links/domain. In this denser topology there are a total of 43 border nodes.

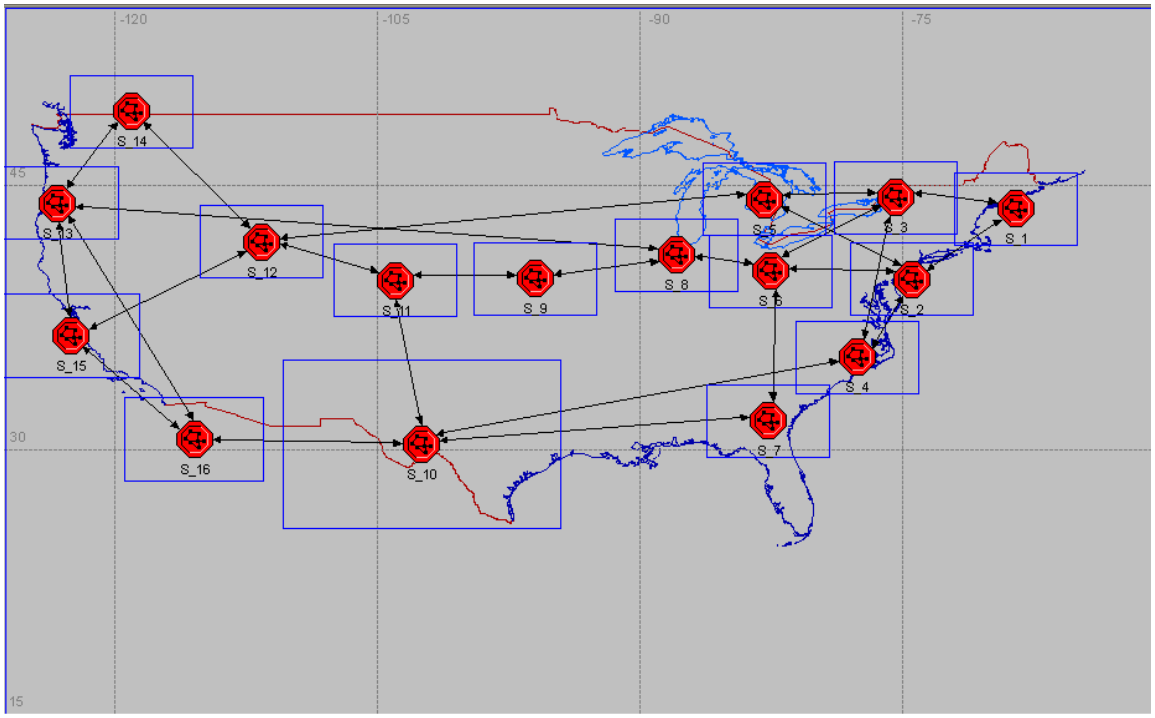

Figure 4.1: NSFNET topology

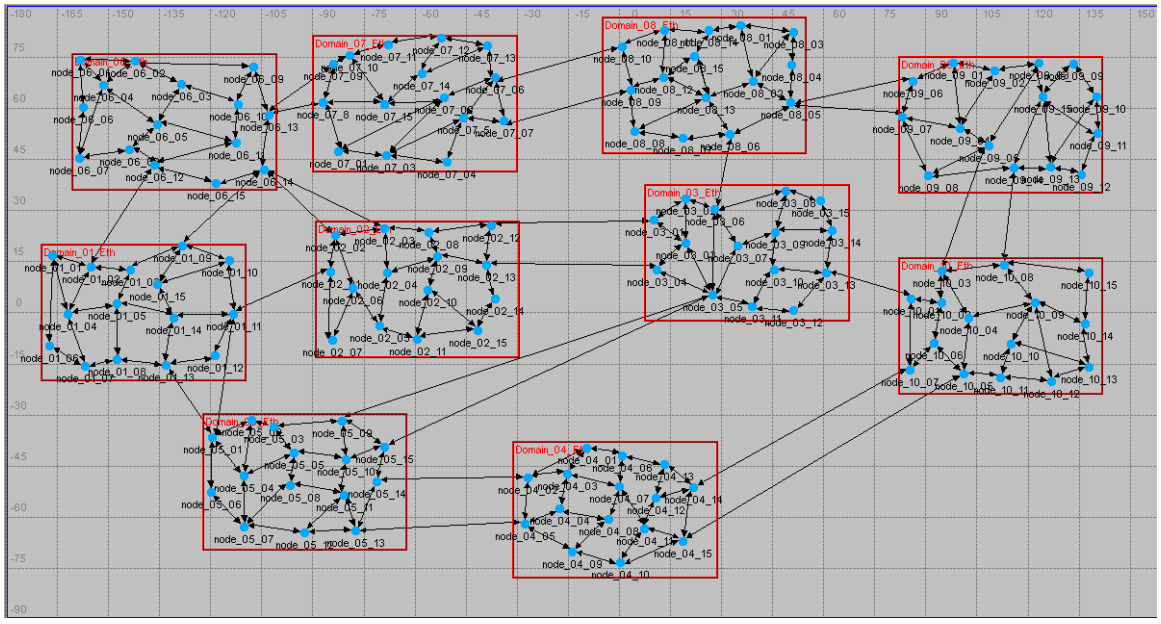

Figure 4.2: 10 domain topology

4.2 Performance Metrics

Various evaluation metrics are used to study the performance of the proposed crankback solution. Foremost the *bandwidth-blocking rate* (BBR) is defined to measure request failure rates. Specifically, first consider the total network capacity requested by all users, i.e., $B_{attempt}$, which is given as the summation of each user request, b_i, where M is the number of attempts, i.e.,

$$B_{attempt} = \sum_{i=1}^{M} b_i,$$

Next, consider the total requested bandwidth of failed inter-domain connections, B_{fail}, given by:

$$B_{fail} = \sum_{i=1}^{N} b_i,$$

where N is the number of failed requests. Hence the BBR is defined as:

$$\text{BBR} = B_{fail} / B_{attempt}.$$

In addition, various other metrics are also used. Namely, the network load is measured using the popular Erlang metric, which is dependent upon the connection request inter arrival and connection holding times as follows:

$$\text{Load (Erlang)} = \frac{T_{hold}}{T_{int}},$$

where T_{hold} is the average connection hold time and T_{int} is the average connection inter-arrival time. In addition, average path length (in link hops) and average connection setup delays (for successful connections) are also used for performance evaluation.

Chapter 5

PERFORMANCE EVALUATION

5.1 Performance Evaluation for Ethernet and IP

Performance of the enhanced multi-domain crankback solution proposed in this thesis is tested by developing specialized models in OPNET *Modeler*^{TM}. Tests are done for the two multi-domain backbone topologies detailed in Chapter 4. Here only inter-domain requests are tested and all connections are generated between random nodes in randomly-selected domains. Each run is averaged over 250,000 connections with mean holding times of 600 sec (exponential). Meanwhile, request inter-arrival times are also exponential and varied with load. Finally, a maximum of K=5 next-hop domain entries are computed in the distance vector table, although the number searched is limited by the H_2 value set in the simulation run.

A key objective in the performance evaluation phase is to compare crankback performance against hierarchical competing inter-domain routing schemes using with topology abstraction, i.e., simple node, full-mesh [19]. Briefly consider the details of these schemes. In full-mesh abstraction, the PCE computes "abstract links" to condense trans-domain routes yielding O ($|B_i|(|B_i|-1)$) state, where $|B_i|$ is the number of border nodes in domain (as introduced in Chapter 3) . The capacity of an abstract link is then derived

as the mean bottleneck capacity of the *k*-shortest paths between the respective border nodes [19], [25]. These links (along with physical inter-domain links) are then advertised using a second level of OSPF-TE routing that runs between border nodes [9]. Namely, link updates are generated using *significance change factors* (SCF) and hold-off timers [31], and the respective values are set to 10% (SCF) and 200 sec (hold-off timer). This inter-domain link state is then used to build a "global" topology for computing/expanding end-to-end *loose-routes* (LR). Meanwhile in simple node abstraction, all domains are condensed to virtual nodes, i.e., no domain-internal state advertised, only physical inter-domain link state. Note, that the exhaustive *per-domain* (PD) crankback scheme of [12] is also tested here for comparison sake. This scheme does not track failed intra-domain links or perform intelligent next-hop domain selection, i.e., next-hop domains selected as those with closest egress border nodes, see Chapter 2. Overall tests are done for multi-domain IP/MPLS and multi-domain DWDM scenarios, and those are now detailed.

5.2 Multi-Domain IP/MPLS Scenarios

For IP/MPLS network settings, all link capacities are set to 10 Gbps and connection requests sizes are varied from 200 Mbps–1 Gbps in increments of 200 Mbps, i.e., to model realistic fractional Ethernet demands. Crankback performance is first evaluated for the case of inter-domain only connections, i.e., no local intra-domain requests. The inter-domain BBR are plotted for the various schemes in Figure 5.1 (for 10 domain topology) and Figure 5.2 (for NSFNET topology). Note that in these "HR" denotes hierarchical routing, "CB" denotes crankback, and "PD" denotes the scheme in

[12]. Moreover, several configurations are tested for the enhanced crankback scheme, including intra-domain only (H_1=0/H_2=3) and joint (H_1=3/H_2=3, H_1=5/H_2=5). First of all, the results for both network topologies indicate that the enhanced scheme gives the best performance when both intra and inter-domain crankback is enabled, i.e., intra-domain-only crankback with H_1=0 gives highest blocking. Next, it is also seen that blocking reduction tends to level off after moderate crankback levels, e.g., the blocking performance for H_1=3/H_2=3 closely matches that for H_1=5/H_2=5 and is notably better than that with the more exhaustive PD crankback scheme [12]. In general, this is due to the fact that excessive crankback attempts yield increased route lengths and higher bandwidth fragmentation.

Additionally, the results in Figures 5.1 and 5.2 also show that the proposed crankback solution (with moderate counter values, i.e., H_1=3/H_2=3) can even outperform the other, more complex hierarchical routing strategies. In particular, resultant BBR values are always lower than those yielded by simple node abstraction, and for the case of NSFNET, even lower than those yielded by more advanced full-mesh abstraction. This is a very significant gain, given the fact that associated crankback messaging overheads (not shown here) are over an order magnitude lower than hierarchical routing message loads.

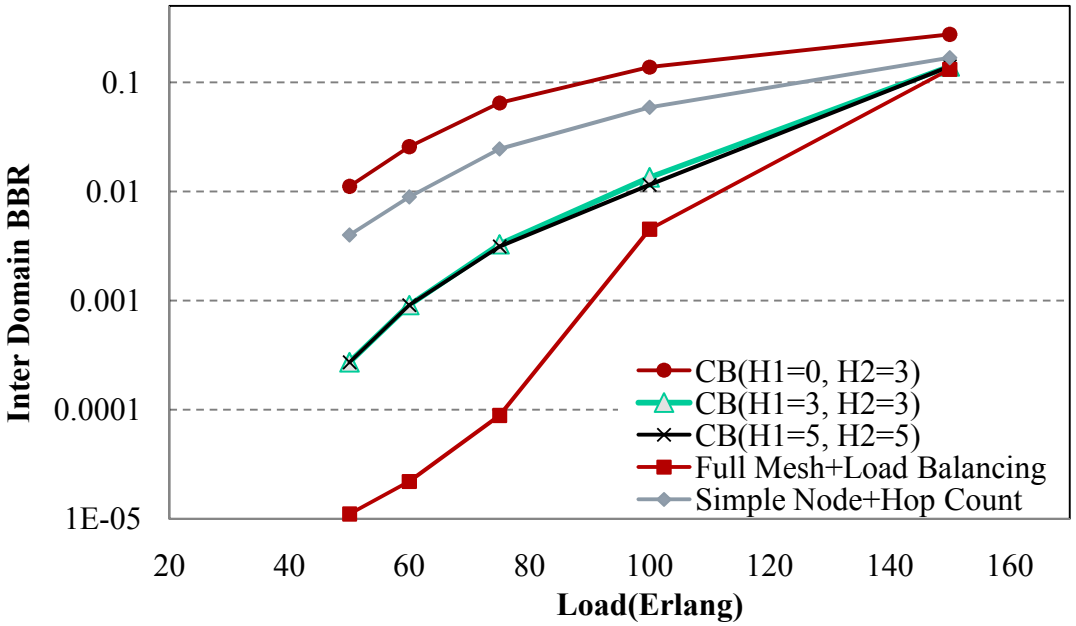

Figure 5.1: Inter-domain BBR performance for 10 domain

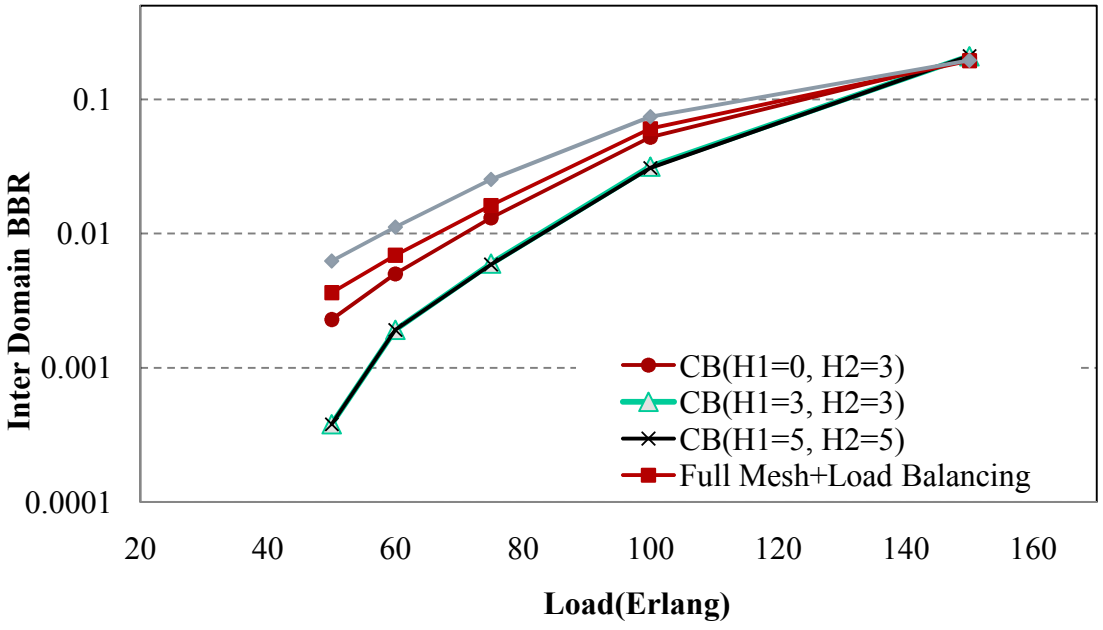

Figure 5.2: Inter-domain BBR performance for NSFNET

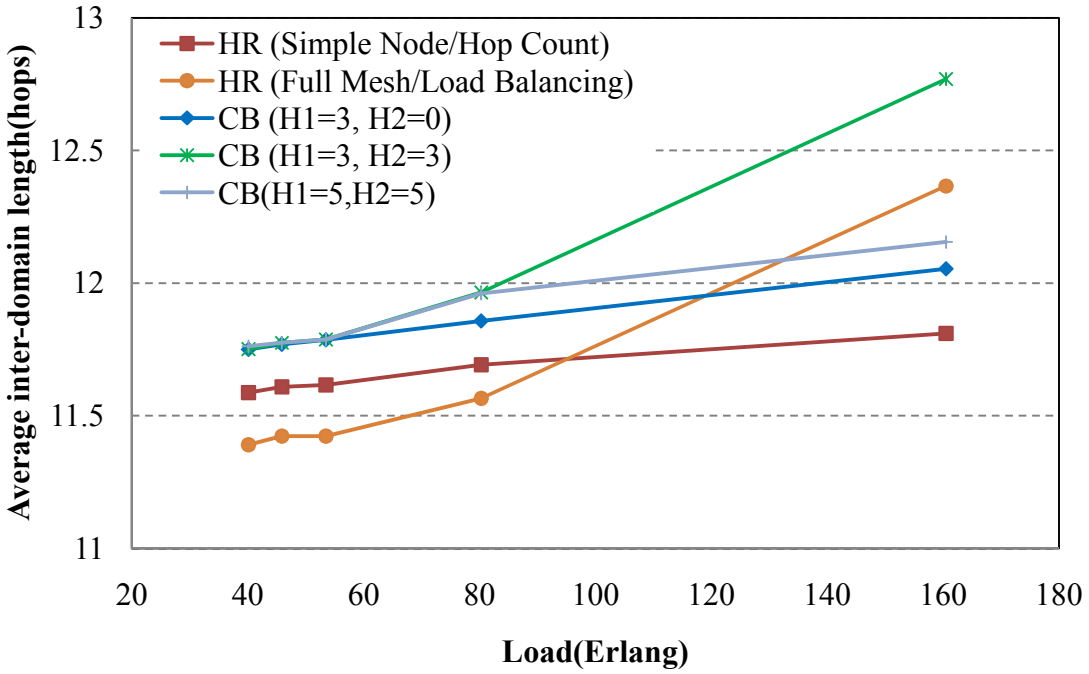

Figure 5.3: Average inter-domain lightpath for 10 domain network

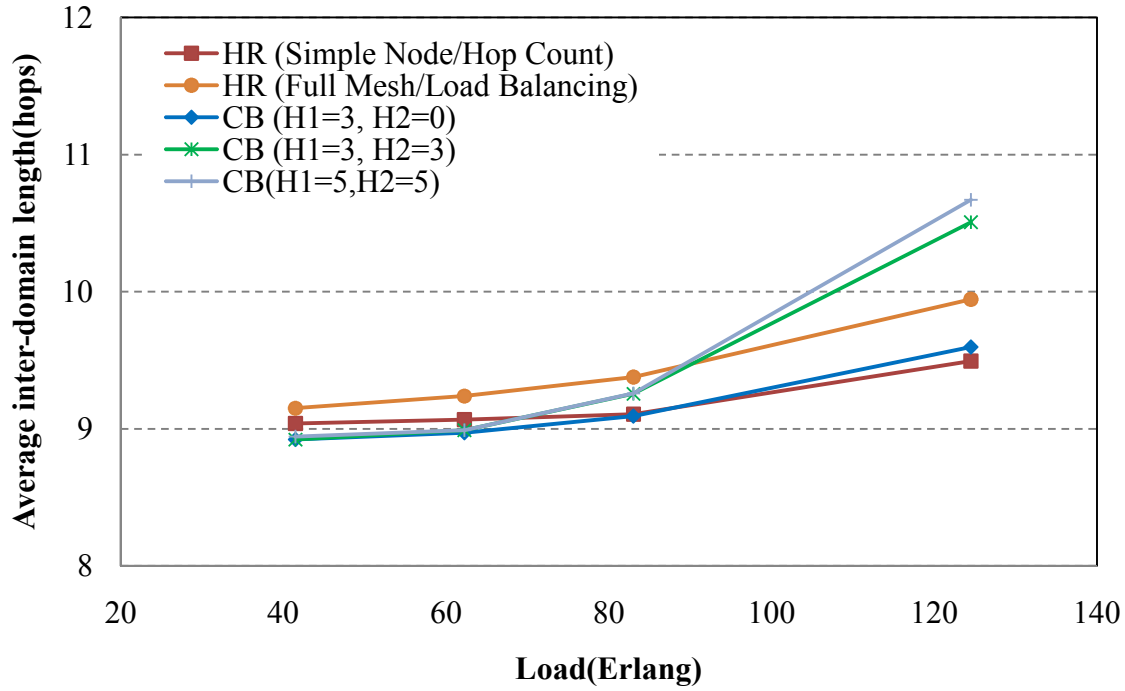

Figure 5.4: Average inter-domain length for NSFNET

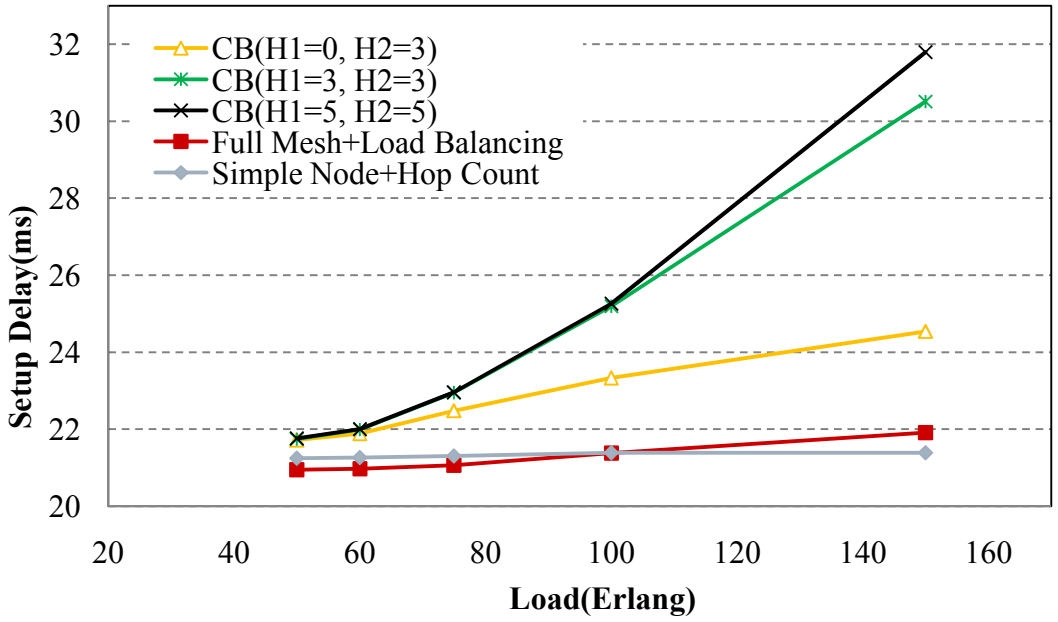

Figure 5.5: Average setup delay for 10 domain network

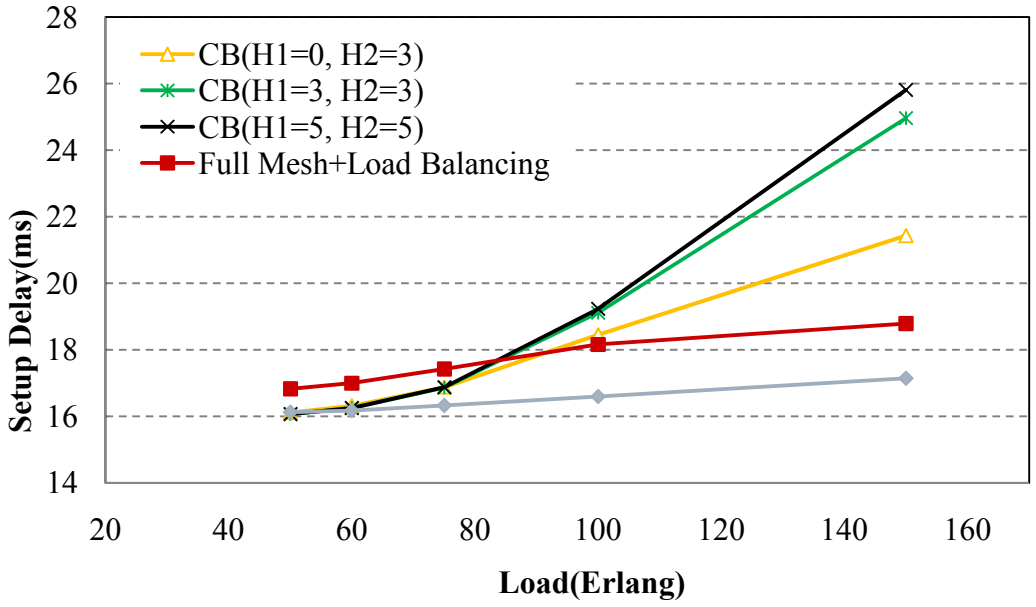

Figure 5.6: Average setup delay for NSFNET

Next, the resource usage/efficiencies of the respective schemes are gauged by plotting the average inter-domain path lengths in Figure 5.3 (10 domain) and 5.4 (NSFNET). In both of these topologies, it is seen that increased inter-domain crankback levels (i.e., H_2=3 or 5, exhaustive PD scheme [14]) result in the highest hop-count utilizations, particularly at higher loads. Moreover, these usage levels are also higher than those for the hierarchical routing schemes running simple node and/or full-mesh abstraction. Nevertheless, such increases are generally expected when performing "per-domain" crankback operation, and the results show that the maximum increases are bounded by 10% even at high loads. In addition, end-to-end setup delays for successful connections are also plotted in Figures 5.5 (10 domain) and 5.6 (NSFNET), assuming 1.0 ms link delays and 0.05 ms node processing delays. Again, these results show that the proposed crankback scheme generally gives higher setup delays when running both intra and inter-domain crankback, i.e., as compared with hierarchical routing. However, these increases are generally bounded in the 15-20% range and are most pronounced at very high loads (over 10% BBR ranges).

5.3 Multi-Domain DWDM Scenarios

The performance of the enhanced crankback scheme in multi-domain DWDM networks is also tested for the two topologies in Chapter 4. Again, lightpath requests are randomly generated between domains/nodes and each run comprises of 250,000 requests with exponential holding times (mean 600 sec). Furthermore, K=5 next-hop domain

entries are computed in the multi-entry distance vector table as well. Finally, the enhanced crankback scheme is also compared against more complex hierarchical inter-domain DWDM link-state routing/RWA solutions that use simple node and full-mesh topology abstractions, see [25].

Inter-domain lightpath blocking is first measured in Figures 5.7 (10 domain) and 5.8 (NSFNET) for varying crankback levels. Foremost, the results indicate that joint intra/inter-domain crankback with moderate counter values again yields the best performance, i.e., lightpath blocking reduction levels off after H_1,H_2=3. Moreover, inter-domain-only crankback (H_1=0) is not effective and yields notably higher blocking. More importantly, the enhanced crankback RWA scheme outperforms hierarchical DWDM routing with simple node abstraction in all cases and even outperforms advanced full-mesh abstraction for the NSFNET topology, i.e., lower inter-domain connectivity. Note that these gains also come with much lower control plane overheads as crankback overheads are over an order magnitude lower than hierarchical routing overheads at mid-to-high loads (not shown).

Next, inter-domain setup delays are plotted in Figures 5.9 (10 domain) and 5.10 (NSFNET), assuming 1 ms backbone link delays and 0.05 ms OXC message processing delays. Here it is seen that the enhanced crankback scheme again yields increased lightpath setup delays, particularly at high loads, and this is most notable in the NSFNET topology with lower inter-domain connectivity. Although, these delays are almost 30% higher in many cases, these values are generally acceptable for long-standing circuit-switched demands. Overall these results show that moderate levels of intra/inter-domain

crankback driven by distance/path-vector state achieve a good tradeoff between provisioning complexity and blocking for inter-domain RWA as well.

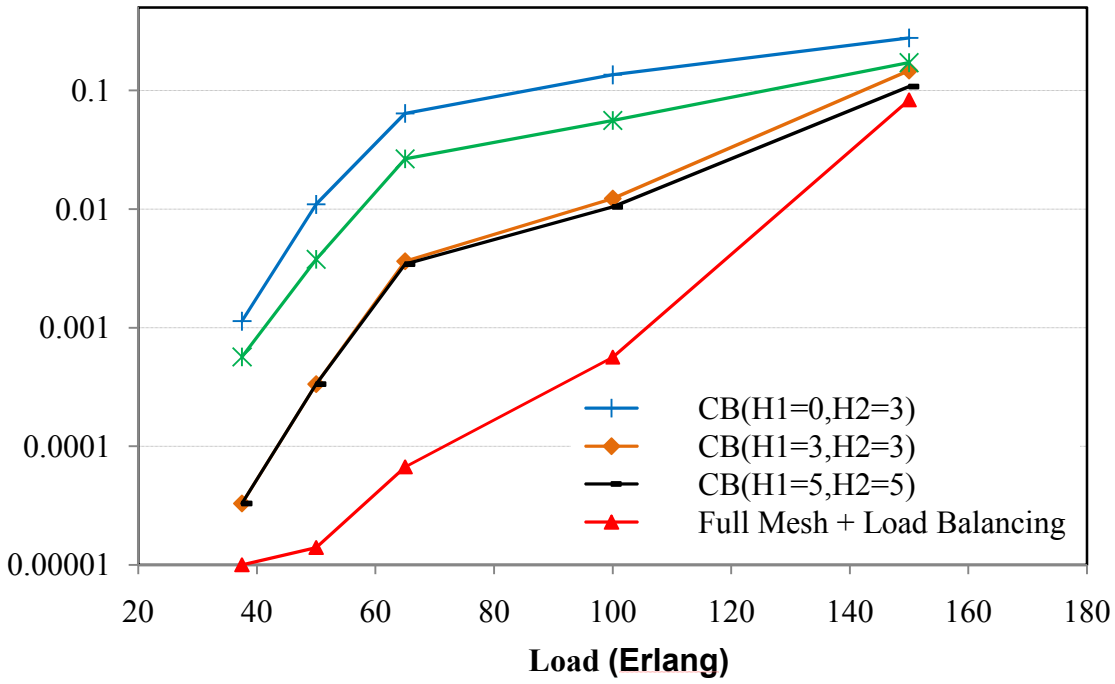

Figure 5.7: Inter-domain lightpath blocking for 10 domain network

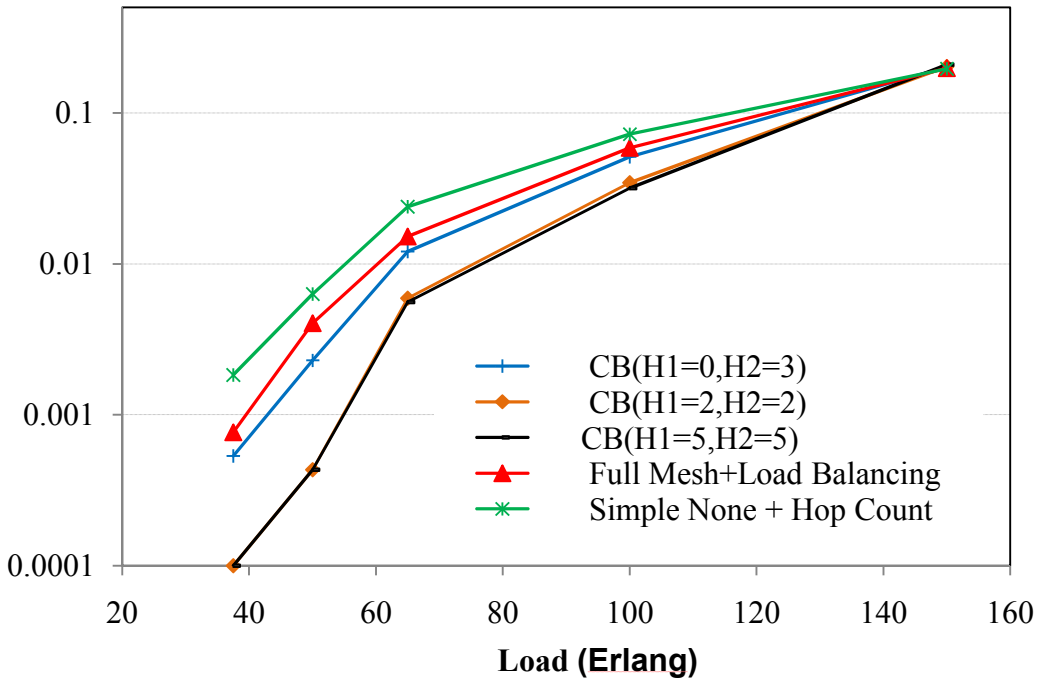

Figure 5.8: Inter-domain lightpath blocking for NSFNET

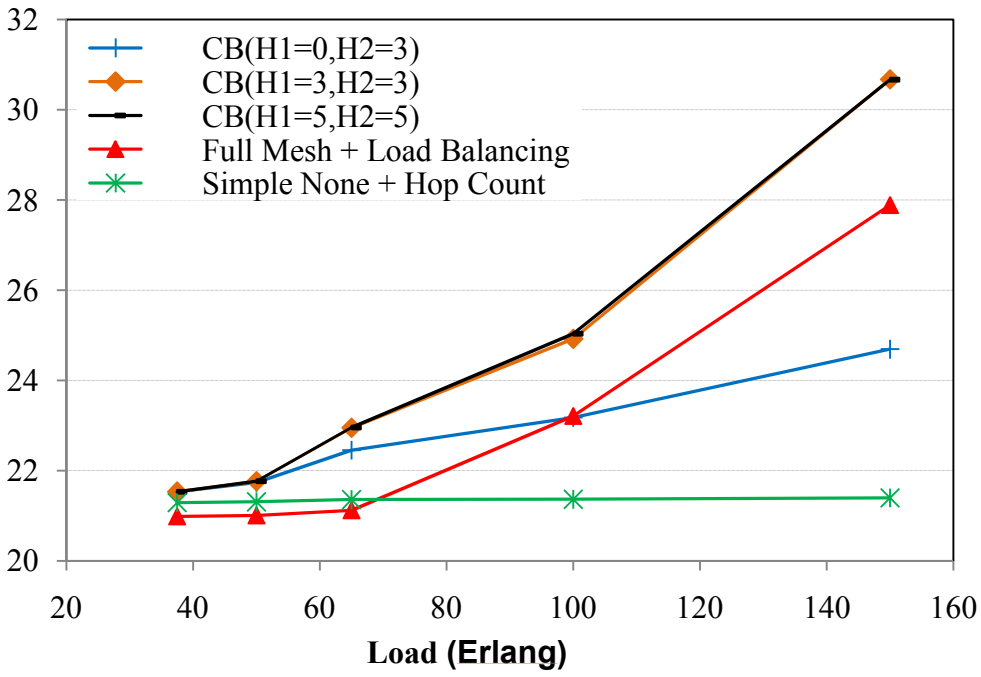

Figure 5.9: Average lightpath setup delay for 10 domain

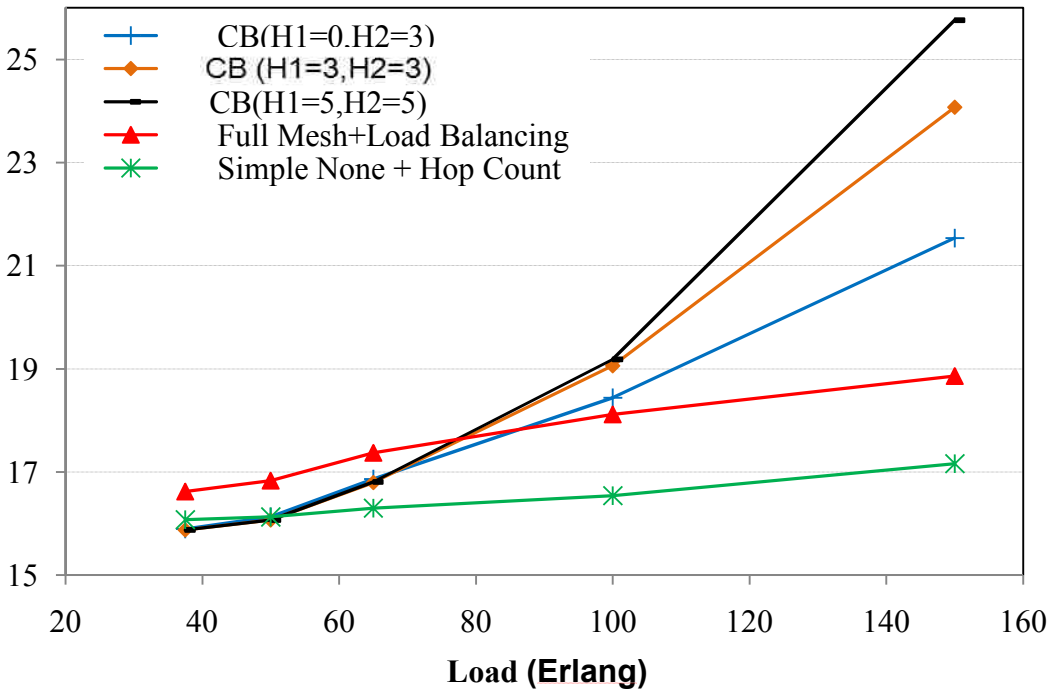

Figure 5.10: Average lightpath setup delay for NSFNET

Chapter 6

CONCLUSIONS AND FUTURE WORK

Multi-domain traffic engineering in MPLS/GMPLS networks is a very challenging problem area and crankback signaling offers a very promising solutions framework. However, there are very few studies the application of crankback in multi-domain networks, and the few existing efforts leave much room for extension. As a result this research project was designed to study realistic IP/MPLS multi-domain networks and develop novel solutions for joint intra/inter-domain signaling crankback. Along these lines this thesis proposed an improved and enhanced crankback solution for multi-domain networks using the standard RSVP-TE protocol. Specifically, two levels of crankback are defined - at the intra and inter-domain levels - and active crankback history (failure state) is also tracked. Furthermore, the proposed solution addresses realistic scenarios where individual domains have full internal visibility via link-state routing, e.g., via OSPF-TE protocols, but generally limited "next-hop" inter-domain visibility, e.g., as provided by BGP or hierarchical OSPF-TE. Moreover, provisions are also introduced to support optical DWDM wavelength routing networks via GMPLS. The performance of the scheme is evaluated using discrete event simulation for different network topologies. The findings are also compared against those yielded by competing hierarchical inter-domain routing strategies.

6.1 Conclusions

This work has developed and analyzed a viable standards-based solution for multi-domain crankback in MPLS/GMPLS networks. The key findings from this effort include:

- The combination of joint intra and inter-domain crankback yields notably better blocking reduction versus just intra-domain or inter-domain only crankback. In many cases, these reductions can approach an order of magnitude.
- The proposed hierarchical crankback scheme gives very competitive performance versus counterpart hierarchical inter-domain routing, i.e., schemes using single node and full mesh topology abstraction.
- Increasing the number of intra/inter-domain crankback counter values yields diminishing impact on blocking reduction. Specifically best results are seen with approximately 2-3 intra and inter-domain crankback attempts.
- Setup delays and average connection hop counts increase with higher loads and crankback counter values. This is generally expected as increased resource contention at higher loading points result in longer path

sequences for successful setups. These increases however are bounded to within 20% of the respective values for hierarchical routing.

- The proposed enhanced crankback performs other "exhaustive" inter-domain crankback strategies. In all cases tested this solution also performs single node abstraction. Furthermore, depending upon the topology, the scheme is also capable of outperforming hierarchical routing.

6.2 Future Research Directions

The solution here has addressed crankback in multi-domain network settings with limited inter-domain state information. Overall, this effort this provides a strong foundation from which to develop more advanced renditions of crankback strategies. Specifically the active tracking of crankback history state at ingress border nodes has not been considered. Along these lines, new solutions can be investigated to share such information between multiple connection setup attempts. Furthermore crankback presents an extremely viable means for post-fault recovery/restoration, i.e., particularly against unstructured multiple failure events resulting in more than one node/link failures. Along these lines, novel crankback extensions can be devised for end-to-end and intermediate fault restoration. Finally, detailed signaling timing and complexity analyses can also be done to characterize and bound the performance of multi-domain crankback strategies in general.

REFERENCES

[1] B. Davie, Y. Rekhter, *MPLS: Technology and Applications,* Morgan Kaufman, 2000

[2] B. Mukherjee, *Optical WDM Networks*, Springer, 2006.

[3] G. Bernstein, *et al*, *Optical Network Control-Architecture, Protocols and Standards,* Addison Wesley, Boston, 2003.

[4] D. Medhi, "Quality of Service (QoS) Routing Computation with Path Caching: A Framework and Network Performance," *IEEE Communication Magazine,*, Vol. 40, No. 12, December 2002, pp. 106-113.

[5] H. Zang, J. Jue, B. Mukherjee, "A Review of Routing and Wavelength Assignment Approaches for Wavelength-Routed Optical WDM Networks", *Optical Networks Magazine*, Col. 1, No. 1, January 2000.

[6] W. Alanqar, A. Jukan, "Extending End-to-End Optical Service Provisioning and Restoration in Carrier Networks: Opportunities, Issues, and Challenges," *IEEE Communications Magazine*, Jan. 2004, pp.52-60.

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

[8] J. Vasseur, "Inter-Area and Inter-AS MPLS Traffic Engineering," *IETF Draft draft-vasseur-ccamp-inter-area-as-te-00.txt*, February 2004.

[9] J. Moy, *OSPF: Anatomy of a Routing Protocol*, Addison Wesley Publishers, 1998.

[10] J. Ash, J. Le Roux, "A Path Computation Element (PCE) Communication Protocol Generic Requirements," *IETF RFC 4657*, September 2006.

[11] A. Farrel, *et al*, "Crankback Signaling Extensions for MPLS and GMPLS RSVP-TE," *IETF Request RFC 4920*, July 2007.

[12] S. Dasgupta, J. C. de Oliveira, J. P. Vasseur, "Path-Computation-Element-Based Architecture for Interdomain MPLS/GMPLS Traffic Engineering: Overview and Performance," *IEEE Network*, Vol. 21, No. 4, July/August 2007, pp. 38-45.

[13] F. Aslam, *et al*, "Interdomain Path Computation: Challenges and Solutions for Label Switched Networks," *IEEE Communications Magazine.*, Vol. 45, No. 10, Oct. 2007, pp. 94-101.

[14] F. Aslam, *et al*, "Inter-Domain Path Computation Using Improved Crankback Signaling in Label Switched Networks," *IEEE ICC 2007*, Glasgow, Scotland, June 2007.

[15] C. Pelssner, O. Bonaventure, "Path Selection Techniques to Establish Constrained Interdomain MPLS LSPs", *Proc. of IFIP International Networking Conference*, Coimbra, Portugal, May 2006.

[16] I. Iliadis, "Optimal PNNI Complex Node Representations for Restrictive Costs and Minimal Path Computation Time", *IEEE/ACM Transactions on Networking,* Vol. 8, No. 4, August 2000.

[17] "User Network Interface (UNI) 2.0 Signaling Specification: Common Part", *OIF-UNI-02.0-Common*, February 2008 (available at http://oiforum.com/public/documents/OIF-UNI-02.0-Common.pdf) .

[18] "OIF E-NNI Signaling Specification", *OIF-E-NNI-Sig-02.0*, April 2009 (available at http://oiforum.com/public/documents/OIF_E-NNI_Sig_02.0.pdf)

[19] F. Hao, E. Zegura, "On Scalable QoS Routing: Performance Evaluation of Topology Aggregation," *IEEE INFOCOM 2000*, pp. 147-156.

[20] K. Liu, K. Nahrstedt, S. Chen, "Routing with Topology Abstraction in Delay-Bandwidth Sensitive Networks," *IEEE/ACM Trans.s on Networking*, Vol. 12, No. 1, February 2004, pp. 17-29.

[21] T. Kormaz, M. Krunz, "Source-Oriented Topology Aggregation with Multiple QoS Parameters in Hierarchical Networks," *ACM TOMACS,* Vol. 10, No. 4, October 2000, pp. 295-325.

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

[23] G. Liu, *et al*, "On the Scalability of Network Management Information for Inter-Domain Light-Path Assessment," *IEEE/ACM Transactions on Networking*, Vol. 13, No. 1, January 2005, pp. 160-172.

[24] S. Sanchez-Lopez, *et al*, "A Hierarchical Routing Approach for GMPLS-Based Control Plane for ASON," *IEEE ICC 2005*, Seoul, Korea, June 2005.

[25] Q. Liu, *et al*, "Hierarchical Inter-Domain Routing and Lightpath Provisioning in Optical Networks", *OSA Journal of Optical Networking,* Vol. 5, No. 10, October 2006, pp. 764-774.

[26] X. Yang, B. Ramamurthy, "Inter-Domain Dynamic Routing in Multi-Layer Optical Transport Networks," *IEEE GLOBECOM 2003*, San Francisco, CA, December 2003.

[27] Y. Zhu, *et al*, "Multi-Segment Wavelength Routing in Large-Scale Optical Networks," *IEEE ICC 2003*, Anchorage, AL, May 2003.

[28] D. Truong, B. Thiongane, "Dynamic Routing for Shared Path Protection in Multi-Domain Optical Mesh Networks," *OSA Journal of Optical Networking*, Vol. 5, No. 1, January 2006, pp. 58-74.

[29] B. Jaumard, D. Troung, "Backup Path Re-Optimizations for Shared Path Protection in Multi-Domain Networks," *IEEE GLOBECOM 2006*, San Francisco, CA, November 2007.

[30] U. Pooch, *Discrete Event Simulation: A Practical Approach*, CRC Press, 1992.

[31] R. Alnuweri, *et al*, "Performance of New Link State Advertisement Mechanisms in Routing Protocols with Traffic Engineering Extensions," *IEEE Communications Magazine*, Vol. 42, No. 5, May 2004, pp. 151-162.