uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Tarek Saad

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Electrical Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

A Framework for Provisioning Reliable Services in Multi-domain Networks

TITRE DE LA THÈSE / TITLE OF THESIS

Hussein Mouftah

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Abdulmontaleb El-Saddik

Michel Kadoch

Chung-Horng Lung

Amiya Nayak

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

A Framework for Provisioning Reliable Services in Multi-Domain Networks

by

Tarek W. Saad

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the Ph.D. degree in
Electrical and Computer Engineering

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-61394-8
Our file Notre référence
ISBN: 978-0-494-61394-8

Dedication

To the loving memory of my beloved father *Wafic Ali Saad*

Abstract

The global Internet is effectively a loose collection of semi-autonomous constituent networks. Each of these operates with its own policies, services, prices, and customers making independent decisions about where and how to secure supply of various components that are needed to create the network service.

The next generation network is about convergence of multiple services onto a single multilayered network. A service or quality offered by the network is envisaged as a concatenation of dissimilar services offered at multiple switching layers and domains. Consequently, control plane mechanisms must support the routing of service requests through a series of regions using dissimilar convergence layers. Generalized Multi-Protocol Label Switching (GMPLS) protocol suite has been extensively deployed to provide consistent quality of service (QoS) within intra-domain boundaries. This process is simple and efficient when a GMPLS Label Switched Path (LSP) involves only one domain, but can potentially become severely resource heavy, complex, and inefficient when multiple domains are involved. The Path Computation Element (PCE) has been introduced as a candidate solution to solve this. A PCE collects link-state information and performs path computation on behalf of other network nodes. Since the PCE is completely devoted to path computation, the visibility of network resources can be extended including a higher number of network nodes and layers.

In this thesis, we propose several novel schemes and heuristics that are suitable for the implementation in multi-domain hierarchical networks. The proposed schemes address the dynamic provisioning of service constrained and/or failure-disjoint paths for point-to-point, as well as point-to-multipoint TE LSPs.

Acknowledgements

It is a pleasure to thank the many people who have made this thesis possible.

First, it is difficult to overstate my gratitude to my Ph.D. supervisor, Dr. Hussein T. Mouftah, for his valuable guidance and ideas. Throughout my thesis-writing period, he has provided encouragement, sound advice, good teaching, good company, and lots of good ideas. I would have been lost without him.

Many thanks to my colleagues Dr. Zafar Ali, and Dr. Siva Sivabalan for the ongoing interesting discussions we have. I am privileged to continue to work with you.

I wish to thank my entire family for providing me a loving environment. Special thanks to my mother, for whom I owe my life, my sister Mona and her whole family, and all my other brothers and sisters who have always been supportive and encouraging.

I wish to thank my wife Hala from the bottom of my heart for her love, endless support, patience, and continuous encouragement without which this work would not at all have been possible. I am very grateful to my children who have inspired me even more to work harder to achieve this goal.

Last but not least, thanks be to God for everything that I have. You have made my life bountiful. May your name be exalted, honored, and glorified.

Contents

List of Tables

List of Figures

Glossary

ABR Area Border Router. 35

AP Alternate Path. 65

AS Autonomous System. 5, 11, 12, 24, 29, 30, 32, 33, 35–38

ASBR Autonomous System Border Router. 35

ASLRT Aggregate SRLG Link Resource Tree. 43, 53–56

ASON Automatically Switched Optical Network. 18, 21

ATM Asynchronous Transfer Mode. 2, 17, 23, 42, 144, 145

BGP Broder Gateway Protocol. 17, 25, 27, 33, 34, 44, 78, 80

BN Border Node. 1, 2, 33, 45, 49, 50, 53, 54, 57, 63–65, 70–72, 78, 99, 119, 135

BRPC Backward Recursive PCE-based Computation. 36, 79, 80, 94–98, 100, 101, 105, 107, 108, 127

CBR Constraint Based Routing. 87, 89

CE Customer Edge. 144, 150

CoS Class of Service. 146, 147

CSPF Constraint Shortest Path First. 124, 135, 148

CWS Computation While Switching. 30, 34

Diffserv Differentiated Services. 151, 161

DSCP Differentiated Services Code Point. 151

DWDM Dense Wavelength Division Multiplexing. 21, 161

EBGP Exterior Border Gateway Protocol. 28

ECMP Equal Cost Multi-Path. 133

ERO Explicit Route Object. 33, 120

EXP MPLS header EXPerimental bits. 151

FA Forwarding Adjacency. 50, 53, 64

FAR Failure Aware Routing. 24

FDR Failure Diverse Routing. 24

FRR Fast ReRoute. 7, 78, 145

GMPLS Generalized Multiprotol Label Switching. 2–5, 8, 11–14, 21–25, 37–39, 41

GP2P Generalized P2P heuristic. 132, 139, 140

GRE Generalized Routing Encapsulation. 7, 155, 156, 158, 161

H-LSP Hierarchical LSP. 29

HDTV High Definition Television. 3, 5

HSPF Hierarchical Shortest Path First. 44, 63, 66, 68

IDRA Inter-domain Routing Agents. 17

IETF Internet Engineering Task Force. 7, 125, 127

IGP Interior Gateway Protocol. 11, 16, 25, 32–35, 39

IP Internet Protocol. 42, 46, 50, 51, 78

IPO IP over Optical. 15

IPTV IP Television. 1, 39, 124

IS Incremental in-progress Sub-tree. 133, 135, 139

ISP Internet Service Provider. 2–5, 8, 11, 27, 77–79, 81, 82, 107

IT Individual SLRT Tree. 54

L2TP Layer 2 Tunneling Protocol. 7, 155, 157, 161

LARAC Lagrange Relaxation Aggregated Cost. 91

LCAB Least Cost Availability Bounded. 90, 91, 98

LDP Label Distribution Protocol. 32, 40

LP Link Protection. 47, 48

LPE Logical Provider Edge. 150

LRS Least-Response PCE selection. 115

LSP Label Switched Path. 6, 8, 11, 16, 78, 86, 87, 98, 126, 127

LSR Label Switching Router. 12, 35, 86, 87

MAC Media Access Control. 144, 150

MCP Multi-Constraint Problem. 31, 70, 72, 76

MLT Multi Link Trunking. 147, 148

MP Merge point. 47

MPEG Motion Pictures Expert Group. 146

MPLS Multiprotol Label Switching Traffic Engineering. 27, 29, 37–39, 127

MPLS Multiprotol Label Switching. 1, 7, 8, 27, 30, 42, 78, 86, 128, 144

MRP Most Reliable Path. 86, 95, 96

MSS Maximum Segment Size. 158

MST Minimum Steiner Tree. 129

MTBF Mean Time Between Failures. 82

MTE Multi-Layer Traffic Engineering. 24

MTTF Mean Time To Failure. 82

MTTR Mean Time To Repair. 82

MTU Maximum Transfer Unit. 158, 159, 161

NHOP Next-Hop. 33, 34, 48, 114

NNHOP Next-to-Next-Hop. 48

NNI Network to Network Interface. 145

NP Node Protection. 47, 48

OBGP Optical Border Gateway Protocol. 16, 17

OCI Optical Channel Identifier. 145

ONRL Optical Networks Research Laboratory. 156

OXC Optical Cross Connect. 16, 20, 23

P Provider core node. 129

P2MP Point-to-Multipoint. 3, 5–8, 124–129

P2P Point-to-Point. 6, 125–127

PATH RSVP Path message. 119

PCC Path Computation Client. 35, 36, 108

PCE Path Computation Element. 4, 5, 7, 8, 12, 17, 18, 22, 25, 27, 29, 32, 35–38, 40, 48, 70, 79, 80, 94, 97, 98, 100, 101, 105, 107–115, 117, 119–121, 123, 127

PCED PCE Discovery. 106

PCEP Path Computation Engine Protocol. 35, 106, 186

PCReq Path Computation Engine Protocol Reply. 36, 97, 98

PCReq Path Computation Engine Protocol Request. 36, 97, 113

PE Provider Edge node. 125, 128, 144, 149

PLR Point of Local Repair. 47, 48

PNNI Private Network to Network Interface. 17, 30

QoS Quality of Service. 1, 4, 5, 13, 32, 48, 52, 70–74, 76–78, 81, 151

QoSR Quality of Service Routing. 12, 31

RESV RSVP Reservation message. 119

RIB Routing Information Base. 80, 101

RoW Right of Way. 51

RRS Round-Robin PCE selection. 114, 120

RSVP Resource Reservation Protocol. 32, 34, 126

RSVP-TE Resource Reservation Protocol - Traffic Engineering. 22, 29, 30, 33, 34, 40, 126

RWA Routing and Wavelength Assignment. 16, 18–22, 41, 162

S2L Source-to-Leaf sub-LSP. 125, 127, 128

SLA Service Level Agreement. 5, 77–79, 81, 82, 99

SLRT SRLG Link Resource Tree. xix, 51–55

SNR Signal to Noise Ratio. 32

SONET Synchronous Optical Network. 20, 23, 42, 144, 145

SPCE Source PCE. 48, 97, 98, 102, 105

SPF Shortest Path First. 56

SRLG Shared Risk Link Group. 4, 8, 23, 25, 30, 43, 44, 46, 48, 51–56, 76

ST Steiner Tree. 124, 127, 132

STP Spanning Tree Protocol. 145

SVLAN Stacked Virtual Local Area Network. 7, 8, 145–147, 155, 157, 161

TA Topology Aggregation. 17, 71

TBS Token-Based PCE selection. 115, 119, 120

TCP Transfer Control Protocol. 158

TDM Time Division Multiplexing. 2, 15

TE Traffic Engineering. 1, 6, 7, 11–13, 43–45, 78, 124, 125, 127

TED Traffic Engineering Database. 4, 35, 48

UDP User Datagram Protocol. 159

UNI User to Network Interface. 22, 149

VCID Virtual Circuit Identifier. 2, 144, 149

VLAN Virtual Local Area Network. 145, 146, 156

VoIP Voice over IP. 77

VPID Virtual Path Identifier. 2

VPLS Virtual Private LAN Service. 8, 145, 150, 151

VPN Virtual Private Network. 1, 3, 5, 7, 8, 13, 77, 144, 148

VSPT Virtual Shortest Path Tree. 97, 98, 108

WDM Wavelength Division Multiplexing. 1, 16, 18, 20, 23, 42, 46, 144

WSON Wavelength Switched Optical Network. 6–8, 18, 32, 41, 44, 46, 71

X-BRPC Extended Backward Recursive PCE-based Computation. 96, 98

List of Symbols

α	weight parameter used to minimize the number of links used by a tree
β_{ij}	over-subscription factor on link (i, j)
Δ	requested availability
δ_i	probability that wp can get backup resource when both wp and other i primary paths in S_{wp} fail
Γ	sub-set of V spanning the destination of the P2MP tree
Γ^k	sub-set of Γ whose destinations all belong to domain k
$\hat{P}(\mathcal{X})$	probability $\mathcal{X}$ of $\mathcal{R}$ requests are directed to a PCE
$\hat{P}^{lsp}_{block}(inter)$	probability an inter-domain LSP is blocked
$\hat{P}^{lsp}_{success}(inter)$	probability an inter-domain LSP is admitted
$\hat{P}^{lsp}_{success}(k)$	probability an inter-domain LSP is admitted in domain k
$\hat{P}^{ij}_{block}$	probability that the LSP is blocked at link (i, j)
$\hat{P}^{lsp}_{block}(intra)$	probability an intra-domain LSP is blocked
$\hat{P}^{lsp}_{block}(pp)$	probability a path-protected LSP is blocked
$\hat{P}^{p}_{block}$	blocking probability for path p
$\hat{P}^{ij}_{success}$	probability that the LSP is admitted at link (i, j)
ι_{ij}	number of LSPs admitted on link (i, j)

λ^p	total traffic arrival rate on path p
λ^r_{Tot}	total arrival rate of requests at node r
$\mathcal{F}_{gJ}$	total traffic demand from border node g destined to domain J
$\mathcal{F}_{iI}$	total incoming traffic demand destined to egress node i in domain I
$\mathcal{F}_{iJ}$	outgoing inter-domain traffic demand from node i in domain I to domain J, where $I \neq J$
$\mathcal{F}_{ij}$	intra-domain traffic demand between nodes i and j
$\mathcal{F}^J_{ij}$	traffic demand from node i to j that j relays to domain J
$\mathcal{K}$	PCE path computation burst size constant
$\mathcal{M}$	number of eligible downstream PCEs
$\mathcal{R}$	set of path computation requests
$\mathcal{R}^J_i$	traffic demand that node i in domain I receives from another domain and relays to domain J
$\mathcal{T}^J_i$	inter-domain traffic from node i in domain I destined to domain J
$\mathcal{X}$	sub-set of path computation requests from $\mathcal{R}$
μ	average path computation service rate, *i.e.*, $\frac{1}{\tau}$
μ^r	average path computation service rate at PCE r
Ω_{wp}	set of primary paths sharing the same backup with wp
$\overline{St}^{\Gamma}$	set of all Steiner trees spanning Γ leaves
$\overline{St}^{o}$	set of all Steiner trees for P2MP LSP o
$\phi^{p(J)}_{ij}$	inter-domain traffic demand between nodes i and j on path p destined to domain J
π^p_{ij}	intra-domain traffic demand between nodes i and j on path p
ρ	simulated traffic load

ρ^r	load of requests at node r
ρ_{global}	global inter-domain traffic load
ρ_{local}	local intra-domain traffic load
σ	number of primary LSPs sharing the same backup with wp
τ	average path computation service time
τ^r	average path computation service time at PCE r
Θ	set of offered protection service classes
θ_{ij}^o	binary variable that indicates LSP o uses protection service class θ on link (i,j)
a_{ij}	availability of link (i,j)
$A_{ij}(1+1)$	availability of client layer link (i,j) given 1+1 lightpath protection at optical layer
$A_{ij}(1:N)$	availability of client layer link (i,j) given 1:N lightpath protection at server layer
A_{lsp}	availability of a multi-layered LSP
$A_{lsp}(pp)$	availability of a path-protected LSP
$A_{lsp}(shared)$	availability of a shared path-protected LSP
B	bandwidth capacity of TE links
b	requested bandwidth
b^o	bandwidth requested by LSP o
b_{ij}	bandwidth on link (i,j)
BN_{en}	set of entry border nodes
BN_{ex}	set of exit border nodes
bn_i	border node i

bp	backup path
c_r	virtual link cost to PCE r
c_{ij}	cost of using link (i, j)
c_{ij}^{θ}	cost of using protection service class θ on link (i, j)
$COST(St)$	accumulative cost of P2MP tree St
d^o	destination of LSP o
D_k	network domain k
E	set of directed links in graph G
e_i^o	variable that indicates whether LSP o initiates or terminates at node i
E_{agg}	set of aggregate links
E_{cent}	centralized PCE TED links: union of inter-domain and aggregate links, *i.e.*, $E_{cent} = E_{inter} \cup E_{agg}$
E_{inter}	set of inter-domain links
$E_{St_{MST}^{\Gamma}}$	subset of E that compose the minimum Steiner tree to Γ leaves
F^{sd}	traffic demand from node s to node d
F_{ij}^{sdw}	indicator function that takes the value 1 if path (s, d) uses wavelength w on link (i, j), and 0 otherwise
F_j^{sdw}	traffic demand from node s to node d using wavelength w on node j
$G(V, E)$	graph that represent V set of nodes and E set of directed links
G_{agg}	graph of meshed border nodes with aggregate links
h	average connection holding time
hp^o	maximum path hops for LSP o
K	maximum number of contacted downstream PCE(s)

k	domain hop
L^f_{bp}	set of fiber links at optical layer along backup path
L^f_s	set of fiber links for other lightpaths sharing same backup path
L^f_{wp}	set of fiber links at optical layer along working path
$L^k_{(agg,t)}$	set of all aggregate links in domain k at layer t
L^k_t	set of all intra-domain links in domain k at layer t
$L_{(agg,t)}$	set of aggregate links in a domain at layer t
L_{1+1}	set of links that are 1+1 protected at the server layer
$L_{1:N}$	set of links that are 1:N protected at the server layer
L_{bp}	set of links at the client layer along the backup path
L_{unprot}	set of links that are unprotected at the server layer
L_{wp}	set of aggregate links at the client layer along the working path
m	residual bandwidth on all links in the network
m_{ij}	residual bandwidth on link (i, j)
n	fraction of simulated inter-domain to intra-domain traffic load
O	set of LSPs
$P'(s, d)$	set of paths from node s to node d bounded by path availability Δ
$P(domain)^{d_1}$	domain path for destination d_1
$P(s, d)$	set of paths from node s to node d
P_{ij}	set of paths traversing link (i, j)
PCE_{dst}	PCE in the destination domain
PCE_{src}	PCE in the source domain
Q'_{wp}	set of aggregate links in Q_{wp} whose SLRTs have leaves in $S_{wp}(t)$

Q_{wp}	set of links at the client layer not along the working path wp
R_{wp}	union of L_{wp} and Q'_{wp}
Rq^r_{exp}	expected number of requests q at PCE r
Rt^S	path computation response times for set of PCEs, S
Rt^r_{exp}	expected response time at PCE r
rt_i	response time for eligible PCE i
rt_{min}	minimum response time from eligible PCEs
s^o	source of LSP o
$S_{wp}(t)$	set of SRLG resources at layer t that are along the working path wp
sp^k_d	sub-path computed in domain k to destination d
St^Γ	Steiner tree spanning Γ leaves
St^o	Steiner tree for P2MP LSP o
St^Γ_{MST}	minimum Steiner tree spanning Γ leaves
t	network layer
tf_l	global traffic on link (i,j)
$Ti^\mathcal{X}$	time interval within which $\mathcal{X}$ path computation requests arrive
Tk^S	tokens for the set of PCEs, S
tk_i	tokens of eligible PCE i
u_{ij}	probability link (i,j) along the LSP path is utilized
V	set of nodes in graph G
v^p	admitted traffic arrival rate on path p
$v^p(m)$	admitted traffic rate on path p in network state m
$V^{\rightarrow i}$	set of nodes j upstream of i, $(j,i) \in E$

$V^{i\rightarrow}$	set of nodes j downstream of i, $(i,j) \in E$
V_{bn}	set of all border nodes
V_{bn}^{k}	set of border nodes in domain k
V_{cent}	centralized PCE TED nodes: set of border nodes from all domains, *i.e.*, $V_{cent} = \cup_{m=1}^{k} V_{bn}^{m}$
V_{cr}	set of core LSR nodes
V_{eg}	set of egress LSR nodes
V_{in}	set of ingress LSR nodes
$V_{St_{MST}^{\Gamma}}$	subset of V that compose the minimum Steiner tree to Γ leaves
W	equalized link capacity
W_{ij}	capacity of link (i,j)
w_{ij}^{a}	available wavelengths on link (i,j)
wp	working path
$wt(p)_q$	accumulative weight of path p using link attribute q
x_{ij}^{o}	binary variable that indicates LSP o traverses link (i,j) (1), or not (0)
y_{ij}^{St}	binary variable that indicates if tree St traverses link (i,j) (1), or not (0)
Z	total network revenue
z^{p}	revenue rate on path p

Chapter 1

Introduction

1.1 Background

Future transport networks are evolving to support the well-known growth of data traffic and the new emerging network requirements such as fast and flexible service provisioning, multiple grades of Quality of Service (QoS), and fast network restoration. Next generation transport networks are also about the convergence of multiple services— examples of which are layer-3, layer-2, and the newly emerging optical or layer-1 VPNs [Take 07]– onto a single unified data communication network.

Extensive research has been devoted for many years to the definition of intra-domain Traffic Engineering (TE) tools and protocols, in order to support advanced services. Nowadays, there is a clear requirement to extend these services beyond domain boundaries, particularly for critical inter-domain VPNs, IPTV transport or voice gateways inter-connections.

Existing transport networks are integrated in a vertically layered architecture consisting of multiple transport technologies such as packet Multi Protocol Label Switching (MPLS), Asynchronous Transfer Mode (ATM), SONET/Synchronous Digital Hierarchy (SDH) and Wavelength Division Multiplexing (WDM). In addition to being vertically multilayered, transport networks are also divided into separate multiple horizontal regions (also known as domains) that have separate management/ownership. In this case, connections traversing multiple horizontal regions are connected by Border Nodes (BNs) that either directly connects adjacent network domains, or use inter-domain links to

connect to other BNs in adjacent network domains.

One way to achieve integration between horizontal/vertical layers is to aggregate information about resources present at one layer/(domain), and present the aggregate information to the upper/adjacent layers. This feature is essential in order to promote cooperation and sharing of spare capacity among layers to meet scalability requirements.

By separating the control plane into path computation and signaling planes, these two functionalities can have different architectures with respect to hierarchy. Specifically, the path computation plane can have a hierarchical structure, allowing the aggregation of traffic engineering information in each domain into compact models suitable for path computation, and the signaling plane can retain its flat architecture and follow the computed path in a sequential manner.

The Generalized Multi-Protocol Label Switching (GMPLS) [Mann 04] [Bane 01] architecture provides a common control plane for managing a variety of different network technologies (see Table 1.1) and for leveraging those technologies to enable high-function services across the network. Currently, GMPLS technology is being deployed within provider boundaries to support real-time and interactive services. The extension of these services into the multi-domain scope requires supporting inter-domain QoS guarantees between providers.

PSC	Packet (switching based on MPLS shim header)
L2SC	Layer 2 (switching based on layer 2 header such as ATMVPID/VCID)
TSC	Timeslot (TDM)
LSC	Lambda
WSC	Waveband (contiguous collection of lambdas)
FSC	Fiber (or port)

Table 1.1: The GMPLS switching types

Today, GMPLS is being deployed as an intra-domain TE tool giving the operator the flexibility and performance of a connection-oriented technology seamlessly integrated with IP. However, the potential of GMPLS across domains in the Internet context is almost unexplored partly due to the way Internet traffic exchange is conceived and architectural constraints of the current routing protocol driving the exchange of traffic among domains.

In practice, Internet Service Providers (ISPs) design their networks so they are resilient against network element failures. An important requirement of survivable network design

is the ability to discover failure-diverse routes for client connections. This characteristic of path computation can become increasingly complicated when multiple layers and/or domains are considered. GMPLS Traffic Engineering (GMPLS-TE) offers the ability to provision, a-priori backup LSPs that protect the end-to-end path or portion of (*e.g.*, local recovery for link or node protection).

Transport services are also usually ranked into classes, each entailing certain availability requirements— usually negotiated and agreed upon by the user and the ISP. To enhance the availability of such services, several failure handling methods can exist that vary in their spare resource requirements, and re-routing time. GMPLS-TE allows the provisioning connections in the network with specific path attributes (*e.g.*, end-to-end availability and bandwidth constraints). The path computation process is responsible of selecting or determining such paths, either at the time of, or ahead of service provisioning. If all paths for the service are computed on one node, such path computation is called centralized path computation. A distributed path computation, on the other hand, is performed by several cooperating computation entities either to provide a single complete path in response to a single request, or when a series of requests are issued by controllers along the path as the service is established. The GMPLS path computation entity is expected to consider all user preferences regarding the selection of paths, and to determine one or more optimal paths that have a good likelihood for successful service establishment and that will be operable even when some network resources fail to perform their functions.

The success of multimedia transmissions over the Internet have recently increased the interest in Point-to-Multipoint (P2MP) transmission services. In Point-to-Multipoint (P2MP) transmission, the root node sends the same information to many receivers, called leaves of the tree. Multipoint transmissions allow optimizing network resources, which is of strategic importance for bandwidth consuming flows such as HDTV. Most video transmissions take place within a single domain, but a demand for inter-domain multipoint services is increasing. Inter-domain multipoint capabilities are also interesting for VPNs. Within this context, there is an opportunity for transport service providers to sell very complex and valuable services that could deliver significant revenues.

1.2 Motivation

A new generation of optical networking services and technologies is rapidly changing the world of communications. ISPs are faced with the challenge to support a variety of reliable, secure, and flexible network services and applications to their end-users on a common network infrastructure.

The evolution and maturity of GMPLS and TE have enabled the planning and support of advanced network services by providers worldwide. GMPLS is a versatile solution addressing current problems at the network level such as scalability, quality of service, traffic engineering, and fast recovery by means of local protection techniques such as Fast Reroute or end-to-end protection schemes [Lang 05]. Constraint-based path computation is another fundamental building block for traffic engineering in GMPLS networks. Specifically, path computation in large, multi-domain, multi-region or multi-layers networks is highly complex and may require special computational components and cooperation between the different network domains.

Hierarchical path computation schemes have been regarded as a promising scalable approach [Guo 98b] [Lui 00] [Maie 08]. However, there is little research on the integration of the hierarchical QoS path computation schemes in a GMPLS multilayered and multi-domain network. Additionally, there has been little research investigating mechanisms for supporting QoS/SRLG aggregation schemes suitable in large scale GMPLS networks.

The Path Computation Element (PCE) [Farr 06a] that is present in each node or centralized in each domain is capable of computing optimal and/or diverse TE LSP paths, and providing dynamic inter-layers resource optimization (*e.g.*, optical, and packet layers) of the network's primary and backup capacity.

In the last few years, QoS routing has been recognized as a strong need both at the intra-domain and the inter-domain level, and it is certainly expected that this need will also be present in the future optical-based Internet. The PCE represents a highly appealing and flexible approach to address the issue of constraint path computation within the context of GMPLS optical networks for a number of reasons. The PCE allows to entirely decouple the complexity and CPU demanding operations of solving the multi-constraint problem from optical LSRs. As well the PCE gathers information about the current state link metrics from the TED, and based on that runs heuristics to find feasible paths (or part of end-to-end paths if loose hop computation between PCEs is in use) in polynomial time.

ISPs typically introduce Service Level Agreements (SLAs) to handle the varying requirements of their customers' traffic. The ability to guarantee SLAs across multiple carrier networks in a truly end-to-end manner is a challenging task. By offering end-to-end SLAs, ISPs could collectively profit from higher revenue traffic. At the same time, a customer whose sites belong to different carrier networks could benefit from the same kind of performance assurance a single-carrier customer receives.

Another driver for this thesis lies in the recent success of multimedia transmissions over the Internet which has increased the interest in P2MP transmission services. Multipoint transmissions allow optimizing network resources, which is of strategic importance for bandwidth consuming flows such as HDTV. Most video transmissions take place within a single domain, but a demand for inter-domain multipoint services is increasing. Inter-domain multipoint capabilities are also interesting for VPNs. Recent extensions have enabled TE for P2MP LSPs in GMPLS networks. ISPs can now offer their customers high-quality multicast VPN services by using P2MP TE-LSP tunnels. Given the complexity of multipoint route computation with QoS constraints, the use of PCE seems particularly appropriate. As TE information is not shared between ASs for scalability and confidentiality reasons, a single PCE is unlikely to be able to compute a full inter-AS path, and has to collaborate with the PCEs of other ASs. Hence, there is a need to define efficient computation algorithm capable of computing P2MP trees across multiple domains.

1.3 Objectives

The thesis addresses the above raised issues, and investigates design constructs and algorithms that are suitable for large scale hierarchical layered networks.

In particular, the thesis aims at designing an inter-domain traffic engineering system, schemes and algorithms that dynamically react to traffic changes while at the same time fulfilling QoS requirements for different classes of traffic. Within this context, the problems addressed in this thesis feature:

- The development of a hierarchical aggregation scheme suitable for the path computation of failure-diverse inter-domain paths across horizontal and vertical layers of the network hierarchy.

- The routing and wavelength assignment of ligthpaths that span multiple Wavelength Switched Optical Network (WSON) domains.
- The path computation of availability-constrained Point-to-Point (P2P) LSP paths that span multiple carrier network domains.
- The tree computation of bandwidth-constrained Point-to-Multipoint (P2MP) remerge free sub-LSP paths that span multiple carrier network domains.
- The evaluation, and comparison of several inter-testbed networking services, technologies, and interoperability mechanisms using experiments carried out on R&D testbeds.

In achieving the above, the thesis will:

- Study, design, and formulate analytical mathematical models for the studied problems and propose algorithms and heuristics to achieve feasible solutions.
- Design and implement the proposed algorithms in a multi-domain and multi-layered extensible discrete-event simulation tool that replicates closely the behavior and timing of the several inter-domain TE LSP path computation schemes and protocols.
- Analyze and compare the obtained results (*e.g.*, average LSP request call blocking, path computation time, *etc.*) extracted from simulations runs over realistic multi-domain topologies, and compare against simulation runs for a hypothetical single-domain flat topology network.

1.4 Contributions

The thesis surveys existing literature and related state of the art on hierarchical network aggregation, the topic of establishing disjoint paths across multiple domains, current techniques for inter-domain TE LSP path computation, as well as the computation of inter-domain P2MP trees for P2MP TE LSPs.

We outline the main research contributions of this thesis as:

- A novel hierarchical scheme for routing end-to-end wavelength-routed lightpaths across multiple WSON domains. Two heuristics are presented to solve the single and multi-constraint routing problem.
- A novel link resource tree aggregation scheme that conveys link's SRLG and QoS information across horizontal and vertical layers of the network hierarchy, and is suitable for the path computation of failure-diverse inter-domain paths.
- A novel PCE-based scheme and algorithms/heuristics for the computation of availability-bounded LSP paths that span multiple domains.
- A novel PCE sequence selection scheme and heuristics for the TE LSP collaborative inter-domain path computation that minimizes time and LSP blocking probability.
- A novel PCE-based inter-domain Point-to-Multipoint (P2MP) tree computation scheme and algorithms for bandwidth constraint sub-LSP paths. The proposed extensions to the PCE Backward Recursive Path Computation (BRPC) mechanism were submitted to the IETF as an Internet Draft [Ali 09].
- An experimental evaluation of several layer-2 and layer-3 provider-based VPNs over SVLANs, and IP/MPLS core networks. A comparison of several protection techniques for ensuring service-continuity, including TE Fast-ReRoute (FRR), layer-2 bundling, and 1+1 DWDM optical protection. In addition, an evaluation of protocol overhead for different tunneling techniques connecting layer-mismatched networks, including layer-2 virtual circuits using GRE tunnels, SVLANs, L2TPv3, and transparent bridging using GRE and L2TPv3 tunnels.
- A design and implementation of an inter-domain discrete-event simulation tool that closely replicates the dynamics and the timing of the overall inter-domain path computation and signaling processes for P2P and P2MP TE LSP paths. The tool has been used to evaluate the performance of the proposed schemes and algorithms and is described in more details in Appendix A.

1.5 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 presents a survey of the existing literature on network resource aggregation schemes and their applicabil-

ity to light path provisioning in multi-domain Wavelength Switched Optical Networks (WSON). The chapter also present a review of the work done with regards to constraint and diverse inter-domain path-computation in GMPLS networks. Within the context P2MP communication, a review of recently published research on the topic of inter-domain P2MP tree computation is also presented. Chapter 3 presents a novel scheme for the aggregation of multilayered SRLG resource information for links established over multiple switching layers, and a heuristic for finding a pair of failure-disjoint working and protection inter-domain paths. Within the context of inter-domain WDM networks, the chapter also presents a novel optical link aggregation scheme and a heuristic for the provisioning of lightpaths spanning multiple WSON domains. Chapter 4 presents extensions to inter-domain recursive path computation schemes to enable availability bounded inter-domain LSPs. The chapter introduces a novel scheme for the Multi-constraint path computation problem within the context of inter-domain network. Chapter 5 studies the effect of PCE selection for inter-domain path computation. Through probabilistic analysis, it demonstrates that by distributing information about PCE state congestion, the end-to-end path computation can be made more efficient, and proposes a novel heuristic for distributing the path computation requests among the available candidate PCEs. Chapter 6 investigates extensions to BRPC for inter-domain P2MP tree computation. It introduces three heuristics to solve this problem by extending the existing BRPC inter-domain path computation. It presents results of simulations that we run to demonstrate the applicability and efficacy of proposed algorithms. Chapter 7 presents results of the implementation of several case-studies for providing customer layer-2 inter-VLAN connectivity over a public ISP network infrastructure based on SVLANs and IP/MPLS core network. In the latter case, two cases-studies are presented: 1) layer-2 VPN inter-connectivity using P2P MPLS LSPs, and 2) layer-2 VPN inter-connectivity using VPLS MPLS LSPs. Finally, Chapter 8 concludes this thesis and proposes some future research work.

In Appendix A, we present the design and implementation of a multilayered multi-domain discrete event simulator that we used to run our simulations.

List of Publications

[Ali09] Z. Ali and T.Saad. "BRPC Extensions for Point-to-Multipoint Path Computation". Internet-Draft draft-ali-pce-brpc-p2mp-ext-00.txt, Internet Engineering Task Force, March 2009. Work in progress. This work was completed in partial fulfillment of the PhD degree.

[Saad04a] T. Saad and H. T. Mouftah. "Constraint-based Routing Across Multi-domain Optical WDM Networks". *In proceedings of IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'2004)*, pp. 2065–2068, Niagra Falls, Ontario, May 2004.

[Saad04b] T. Saad and H. T. Mouftah. "Inter-Domain Wavelength Routing in Optical WDM Networks". *In proceedings of the 11th IEEE International Telecommunications Network Strategy and Planning Symposium (Networks Š04)*, pp. 391–396, Vienna, Austria, June 2004.

[Saad04c] T. Saad, H. Naser, and H. T. Mouftah. "A Hierarchical SRLG Organization Scheme for Multi-domain Multi-layered Transport Networks". *In proceedings of the 7th International Symposium on Communications Interworking (Interworking '04), Ottawa*, pp. 301–338, Ottawa, Canada, September 2004.

[Saad04d] T. Saad and H. T. Mouftah. "A Multidomain Differentiated Resilience Scheme Using SRLG Aggregation in Multi-layered Transport Networks". CITO Innovators Showcase, Meet Tomorrow's Technology Leader's Award, Toronto, November 2004.

[Saad05a] T. Saad, B. Alawieh, S. Gulder, and H. T. Mouftah. "Tunneling Techniques for End-to-End VPNs: Generic Deployment in an Optical Testbed Environment". *In proceedings of the 2nd IEEE International Conference on Broadband Networks (BroadNets'05)*, pp. 924–930, Boston, MA, October 2005.

[Saad05b] T. Saad, B. Alawieh, and H. T. Mouftah. "Inter-VLAN VPNs Over a High Performance Optical Testbed". *In proceedings of the 1st International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, TRIDENTCOM '05*, pp. 221–229, Italy, February 2005.

[Saad05c] T. Saad and H. Mouftah. "End-to-End Inter-domain Routing and Signaling for Wavelength Routed WDM Optical Networks". *In proceedings of the IEEE/OSA Optical Fiber Communication and the National Fiber Optic Engineers Conference (NFOEC/OFC '05)*, pp. 293–298, Anaheim, California, March 2005.

[Saad06] T. Saad, B. Alawieh, S. Gulder, and H. Mouftah. "Tunneling Techniques for End-to-End VPNs: Generic Deployment in an Optical Testbed Environment". *IEEE Communications Magazine*, Vol. 44, No. 55, pp. 124–132, 2006.

[Saad07] T. Saad, J. Israr, S. Sivabalan, and H. Mouftah. "An Evaluation for PCE Selection Schemes for Inter-Domain Path Computation". *In proceedings of the 9th IEEE International Conference on Transparent Optical Networks, 2007 (ICTON '07)*, pp. 187–187, July 2007.

Chapter 2

Survey of Related Work

2.1 Introduction

Today's Internet is composed of many competing Autonomous Systems (ASes) which have to cooperate with each other to provide end-to-end services across and through them. Typically, each AS is controlled by a separate administrative entity, and therefore encodes various economic, business, and performance decisions in its routing and Traffic Engineering (TE) policies. Due to security concerns, when different carriers administer networks, they will not disclose any sensitive internal information about their network topology to neighboring carriers (who may be competitors). This makes it much harder to design efficient path computation algorithms that can compute optimal paths spanning multiple carriers.

Several techniques in the literature have been proposed to allow Internet Service Providers (ISPs) to control their outgoing inter-domain traffic, with the objective of facilitating efficient and reliable network operation while simultaneously optimizing network resource utilization and traffic performance. GMPLS-TE has been used within ISP networks for various purposes, including the its ease of forwarding process, as well as the variety of tools it offers for implementing protection and restoration techniques. Within the context of hierarchical GMPLS networks, a Label Switched Path (LSP) can:

- lie within the boundary of a single IGP area (also known as intra-area LSP span),
- traverse inter-area boundaries owned by the same carrier (also known as intra-carrier, inter-area LSP span),

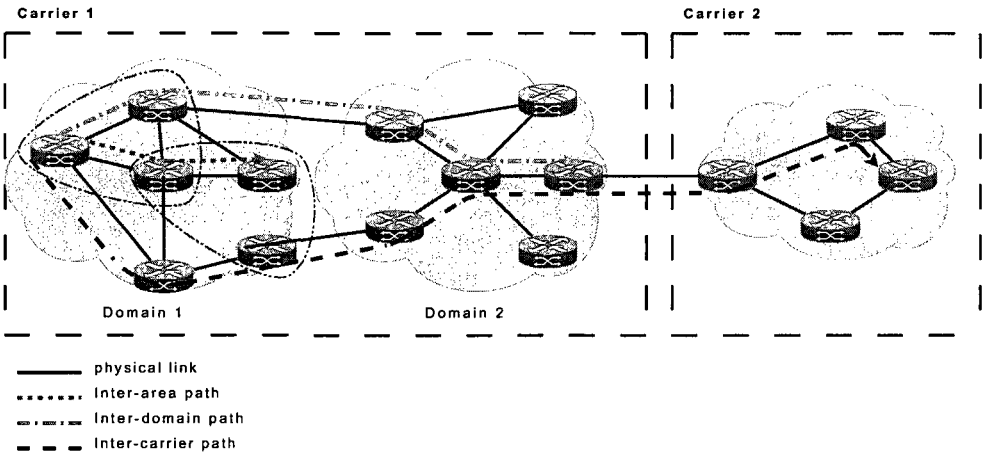


Figure 2.1: Intra-area, Inter-area and Inter-carrier paths

- cross multiple transport switching layers,
- traverse inter-carrier boundaries (also known as inter-carrier or inter-AS LSP span)

Examples of these types of path are shown in Figure 2.1. Recently, there has been expressed interest in extending GMPLS deployment across inter-domain boundaries. Most of the intra-domain GMPLS-TE solutions in the literature do not work when the end-to-end path to the egress LSR leaves the routing domain or AS of the ingress LSR. Hence, having efficient path computation schemes that obey QoS requirements and are able to compute diverse paths for inter-domain LSPs is a yet unsolved research topic. It is challenging since, to make routing scalable, the information that a domain advertises about itself and that learnt from other domains must be rather limited.

One approach to cope with this is with the aggregation of network state information. Aggregation reduces the amount of data to be sent over the network and provides efficient means of hiding confidential details about the domain.

Quality of Service Routing (QoSR) is also expected to play an important role in the future optical-based Internet. The PCE represents a highly appealing and flexible approach to addressing this issue within the context of GMPLS optical networks.

The success of GMPLS-TE for point-to-point TE, has also attracted researchers to study its extensions and related problem withing the context of P2MP communication. Several proposal have been presented, among them the PCE is also being presented as an attractive solution for inter-domain networks P2MP tree computation.

In this chapter, we survey existing literature on the proposed solutions to the above problems within the context of inter-domain and multi-layered GMPLS networks.

2.2 Quality of service

The notion of QoS and network performance are defined in ITU-T Recommendation E.800 (ITU94) as follows:

> *"The Quality-of-Service is the collective effect of service performances that determine the degree of satisfaction of a user of the service. The Network performance is the ability of a network portion to provide the functions related to communications between users."*

End-to-end QoS ordains that all network layers from top-to-bottom, as well as every network element from end-to-end work collaboratively to provide the desired level of QoS. Any QoS assurances are only as good as the weakest link in the chain between sender and receiver. End-to-end QoS service can be achieved by signaling and reserving the network resources in advance before transferring data.

Often, QoS is seen as the need for networks to provide performance bounds on offered services. In other instances, QoS is measured in terms of network survivability or availability in the presence of failures. The traditional best-effort Internet was not designed to support a specified, desired, or consistent level of QoS to network traffic. Nonetheless, customers now require mission critical services– such as corporate VPNs and radio access networks that demand high levels of network availability and guaranteed levels of service under heavy traffic loads. Service providers now use sophisticated TE mechanisms to manage their networks to meet these demands.

2.3 Hierarchy in transport networks

Transport networks are evolving into a complex inter-connection of circuit-switched domains and layers (or granularities), with the delineations being driven by many factors; for example, geographic, administrative, client requirements, economic cost, entrenched infrastructure, and so on [Alan 04].

The integration of the GMPLS architecture with optical transmission networks is considered an important requirement in the effort towards an optical Internet. Concurrently, the scale and reach of high-bandwidth applications continues to grow, mandating services delivery across heterogeneous optical domains. In all, these trends are driving the need for multi-domain, multi-layer optical control plane integration, termed vertical-horizontal hierarchy integration.

Hierarchy [Lai 02] is a method used for creating scalable and complex systems based on the abstraction, at each level, of the most significant of details to the levels further away. Specifically in communication networks, hierarchy is an abstraction of parts of the network's topology, routing or signaling mechanisms. Here, abstraction or aggregation of information is used as a technique to achieve scalability in large networks, or for enforcing administrative, topological, or geographic boundaries. For example, network hierarchy might be used to separate the metropolitan and long-haul regions of a network, to separate the regional and backbone sections of a network, or to interconnect service provider networks.

The vertical hierarchy, however, partitions the network functions into a series of functional or technological layers with clear logical, and sometimes physical separation between adjacent layers. Figure 2.2 shows interactions between adjacent vertical layers of a network element, as well how vertical network layers form adjacencies between adjacent peering network elements.

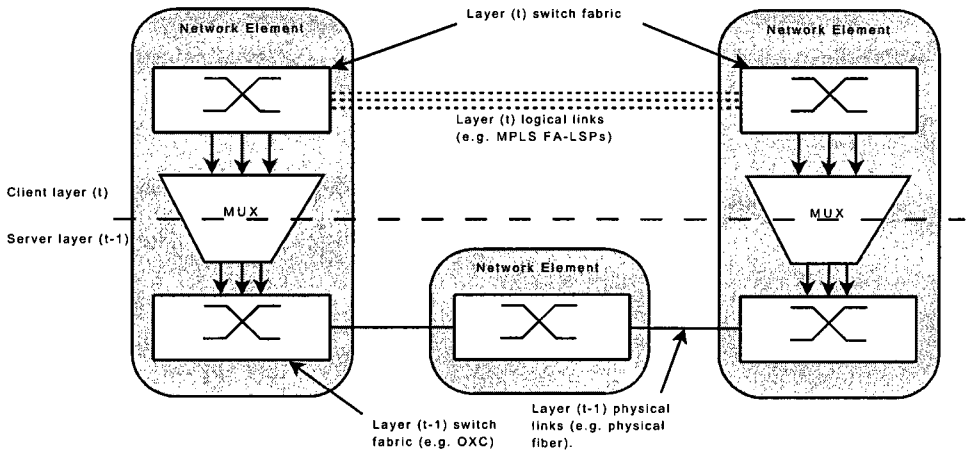


Figure 2.2: Client-server relationship between vertical layers of network hierarchy.

2.3.1 GMPLS control plane designs

Despite continuous work in-progress within standards bodies, the overall area of multi-domain (multilayer) transport networking has not seen significant research focus. Most results reported in the literature for wireline multi-domain networking have focused largely on homogeneous packet networks. To get a better sense of the key challenges herein and survey the related areas, a generic taxonomy is presented.

GMPLS extends the MPLS protocol— originally designed to accommodate IP traffic– to accommodate other types of traffic, such as TDM traffic and WDM traffic. GMPLS provides a common control platform for managing and provisioning these different transport networks. Today, GMPLS is being deployed as an intra-domain traffic engineering tool giving the operator the flexibility and performance of a connection-oriented technology seamlessly integrated with IP. However, the potential of GMPLS across domains in the Internet context is almost unexplored partly due to the way Internet traffic exchange is conceived and architectural constraints of the current routing protocol driving the exchange of traffic among domains.

Several architectural models for the control plane of GMPLS multilayered networks are in existence today. Namely, the overlay, the augmented, and the peer models [Sanc 05]. One of the key differences among these models is how much, and what kind of network information is exchanged between individual layers.

In the overlay network model [Swal 05, Papa 05], the nodes in each layer maintain network information about nodes and links residing in the same layer— such as residual capacity on existing logical links and number of available ports– which makes it more suitable in the case with different management/ownership in each layer. The upper layer only receives a response from the adjacent lower layer on whether or not the requested connection can be set up. There is no specific network information exchanged between individual layers, since the routing in each layer is done separately with each layer's own signaling and control plane. Hence, in this model, each network layer has to decide whether it will use the existing logical links in its topology or try to create new connections/logical links, and how to route the new request over the existing logical topology without any network information from the lower layer. Each layer's control plane is strictly separate from its adjacent one and runs its own routing and signaling instance protocols,and no information is shared among the two. RFC3717 [Raja 04] describes a framework for IP over Optical (IPO) networks that relies on this model.

In the peer model, the topology and other network information (*e.g.*, routing and link state) are shared among all network elements across all the layers by a unified signaling protocol and control plane (*e.g.*, single control plane instance for packet and optical layers). Such a model is appropriate when the transport and service networks are operated by a single entity. As often visualized, a network in the peer model can be seen as one graph with both LSRs and OXCs interconnected with physical and logical edges. For example, an integrated routing scheme can decide routing over logical links (MPLS links), and routing and wavelength assignment (RWA) in the WDM layer at the same time. In the overlay and augmented models, on the other hand, a distinction is present between the RWA in the WDM layer, and the scope of the LSP provisioning problem in the IP/MPLS layer.

The augmented model provides a compromise between the two extreme cases of peer and overlay models by running separate routing protocols in each layer but still allowing the exchange of partial network information between them— for example, reachability and/or summary of link state information and residual capacity. In this case, the IP/MPLS switching layer may utilize a limited network state information passed from the WDM switching layer– for example, the number of lightpaths that the WDM layer can further provide between every LSR pair in the current state of the WDM network.

Within this context, *vertical integration* refers to the collaborative mechanisms within a single control plane instance driving multiple data planes (also referred to as switching layers). *Horizontal integration*, on the other hand, refers to the collaborative mechanisms of control planes extending over several partitions (*e.g.*, IGP areas or ASes) within one data plane instance or vertical switching layer. In this case, the relation between the various horizontal partitions constitutes a peer-to-peer relationship as opposed to a client-server relationship as in the case of vertical integration model.

Szegedi *et al.* in [Szeg 07] discuss the different distributed control plane architectural models in optical networks including peer-to-peer, overlay and augmented models. They investigate the different inter-domain control plane failure problems and their effect on connection provisioning delays. Using simulations carried out on two networks topologies, namely *congruent* and *dual-star* control planes, they present results for the amount of control messaging needed to provision a connection, as well as the provisioning delay for the different models.

Yannuzzi *et al.* in [Yann 08] study different initiatives proposed for control plane model for multi-domain optical networks, including Optical BGP (OBGP). They argue, however,

that future optical networks offer the opportunity to avoid inheriting the limitations of BGP, especially in terms of routing and traffic engineering. They present a route control model that replaces BGP/OBGP that is based on Inter-domain Routing Agents (IDRAs)— closely resembling the functionality of a PCE– based on splitting the control and data planes, where independent connections that connect nodes within the control plane. They present simulations to confirm that their proposed route control model is able to reduce the blocking experienced with OBGP, without increasing the number or the frequency of routing updates exchanged between domains.

2.3.2 Hierarchical topology abstraction

The problem of finding a suitable topology abstraction that enables efficient computation of the minimum weight paths across the network is well studied in the literature. Topology Aggregation (TA) as a way for achieving scalability and security in global networks has been extensively studied in the literature [Awer 98] [Iwat 98] [Lee 03]. The ATM PNNI standard [Ahme 96] defines a hierarchical structure and a flexible representation mechanism for performing topology aggregation. In [Awer 98], Awerbuch *et al.* compared using simulations the performance of several aggregation schemes in terms of network throughput, and network controlled load. They showed that representing the internal structure of a domain as a "star" graph resulted in higher number of cranckbacks and required more frequent re-aggregation. They found that the random spanning tree aggregation was very sensitive to network topology, and showed that with the minimum spanning tree representation with "2-spanner" aggregation the results of their simulation performed close to optimality. In [Iwat 98], Iwata *et al.* suggest several heuristics to path selection accounting for inaccuracy due to state aggregation. In [Xie 97], Xie *et al.* propose a hierarchical reduced load approximation to evaluate PNNI routing. However, no simulation results are presented.

Subsequent efforts attempted to extend these concepts to IP networks as well. For example, [Ulud 07] applies topology abstraction in multi-domain IP QoS networks using star, mesh, tree, and spanner graphs. The schemes are tested with various path computation strategies (*e.g.*, widest-shortest, shortest-distance) and show strong improvements in routing scalability and route fluctuation reduction. Further work in [Lui 00] studied aggregation in directed graphs with delay and bandwidth metrics using information-theoretic and line segmentation techniques. Overall results here showed good gains with

aggregation, that is, higher success, lower crankback, and so on.

Other multi-domain studies, *e.g.*, [Ulud 07] [Tang 07] [Lui 00], considered *topology abstraction* for routing. These schemes proposed graph transformations to condense resource state via virtual graphs with fewer abstract vertices and edges. Typically, this is performed by a designated domain-level entity— *e.g.*, a routing controller in Automatically Switched Optical Network (ASON) [Sanc 05], or a PCE– which propagate abstract link state to other domains to build a global aggregated graph.

In [Maie 08], Maier *et al.* investigate the effectiveness of different topology aggregation methods to the representation of network domains on multi-domain routing in ASON networks. They compare their results against the hypothetical case of no topology aggregation in their comparative analysis. They present the three homogeneous-aggregation cases they refer to as *Simple Node*, *Full Mesh* and *Symmetric Star*. They also propose a new non-homogeneous-aggregation technique they call *Hybrid*. Using simulations performed over testebd topology under both static and dynamic traffic conditions, they present results for blocking probability, average inter-domain connection cost, and dynamic channel occupation for the different abstraction schemes.

2.3.3 Inter-domain RWA in WSON

Although extensive research is found in the literature about the inter-domain connection provisioning problem in packet/cell-switched networks, little effort has been invested to date in the research for the connection provisioning across inter-domain Wavelength Switched Optical Networks (WSON) [Bern 08]. In a WSON network, connections— also referred to as *lightpaths*– are established using all-optical WDM channels. Given a set of connections, the problem of setting up lightpaths by routing and assigning a wavelength to each connection is usually referred to as the Routing and Wavelength-Assignment (RWA) problem.

RWA involves selecting the best combination of route (path) and a wavelength for each of the connections in a given demand matrix such that number of connections established is maximized and no two connections sharing a common link use the same wavelength. A lightpath must occupy the same wavelength on all the fiber links through which it traverses; this property is known as the wavelength-continuity constraint.

The different variants of solutions to the RWA problem can be broadly classified into

two categories: the 1) static RWA, whereby the traffic requirements are known a priori, and 2) the dynamic RWA, in which a sequence of lightpath requests arrive in some random fashion, and depending on the state of the network at the time of the request, the available resources may or may not be able to establish the lightpath.

Routing relies on knowledge of network topology and resource availability. First step towards network-wide link state determination is the discovery of the status of local links to all neighbors by each light switch. Each network node creates an inventory of local resources: *e.g.*, active links, link attributes including delay, cost, *etc.*, and state of available channels (wavelengths) on each link. Routing protocols then distribute this information to the rest of the network.

The dynamic path selection process may take into account the network changing link-state; in this case it is referred to as adaptive routing. Once a routing decision is made, a wavelength assignment scheme can be used to select a wavelength for setting up the lightpath. The performance of a dynamic RWA algorithm is typically measured in terms of the call blocking probability which is the probability that a lightpath cannot be established in the network due to lack of resources.

Some of the reported wavelength assignment schemes in the literature– *e.g.*, the First Fit (FF), Random Fit (RF), *etc.*– select a wavelength along a physical route according to a predefined rules without considering the dynamic link-state [Subr 97] [Kara 98] [Barr 97]. The Least-Used (LU) and Most-Used (MU) schemes [Spat 00] use custom-defined cost function and link-state metrics to select one of the feasible wavelengths for a connection request.

Our works in [Saad 04b], [Saad 05c], and [Saad 04a] study the problem of provisioning lightpaths across multiple optical WSON domains. In [Saad 05c], we examine the general problem of dynamic lightpath provisioning that involves neighbor discovery, topology discovery, route computation, wavelength assignment and lightpath setup signaling functions. In [Saad 04b], we present heuristics to solve the inter-domain RWA problem using two models: 1) a single-domain abstraction model where domains are abstracted to single nodes with abstract optical links connecting them, and 2) a border-to-border abstraction model where domains are abstracted with border nodes with abstract optical links connecting them. Using simulations, we evaluated the two heuristics for request blocking probability under different dynamic traffic load. In [Saad 04a], we considered the problem of routing a multi-constraint lightpath request across multiple WSON domains. We adopted a similar approach to that presented in [Jaff 84] to combine multiple link costs

and a heuristic that finds a feasible path. Using simulations run over a generic multi-domain network, we evaluated the proposed heuristic against request blocking probability against dynamic traffic load.

Several multi-domain WDM RWA studies have been presented in literature— *e.g.*, [Yann 06a] [BS A 01] [Yang 03] [Sanc 05] [Liu 06b]– with the primary focus on light-path selection and resource reservation. In [Yann 06a] and in [Yann 05], Yannuzzi *et al.* present an RWA strategy that is based upon stochastic estimation of the effective number of available wavelengths (ENAW) along the inter-domain paths. Particularly an adaptive Kalman filter is devised to further refine the estimation. However, detailed topology aggregation schemes are not considered.

In [BS A 01], Arnaud *et al.* propose route advertise/withdraw messaging between domains using modified BGP along with proxy lightpath route arbiters to compute routes between border optical cross-connect (OXC) nodes. Nevertheless, the detailed algorithmic study of inter-domain WDM routing/provisioning algorithms has not been considered and this is only now gaining attention.

[Yang 03] details a modified BGP routing and signaling scheme in which domain gateways maintain complete (alternate) route state. Nevertheless related resource propagation issues are not considered and hence this setup is more favorable to BGP-type implementations.

Other proposals also suggested extending topology abstraction to multidomain WDM settings, where optical TE links encompass constraints such as wavelengths, timeslots, converters, colors, in addition to the packet specific link attributes such as basic delay and bandwidth metrics.

In [Liu 05] Liu *et al.* present a theoretical analysis of state aggregation in multi-domain WDM networks with border node conversion. Here various information models are studied and lightpath assessment is modeled as a Bayesian decision problem. However, this treatment only considers (unrealistic) bus topologies thus the crucial issue of aggregated topology computation between domain border nodes is averted. Furthermore, no routing protocols or distributed lightpath setup are detailed. Although these schemes present many saliencies, they are premised upon the availability of global state, *i.e.*, "flat/single-domain" network. Hence the extension of these schemes into distributed multi-domain WDM/SONET networks with no/limited global state is not straightforward.

In [Liu 07c] and [Liu 07a], the authors propose a GMPLS-based hierarchical OSPF-TE

routing approach to facilitate inter-domain service provisioning. Specifically, they propose adapting inter-area OSPF-TE to resolve routing scalability and security issues in WDM networks. Performance analysis results are presented for several inter-domain lightpath RWA and signaling schemes to demonstrate the effectiveness of their proposed mechanisms.

Several other related efforts have considered the next-hop domain routing approaches. For example, [Yang 03] details a domain-by-domain RWA scheme where gateways maintain complete alternate routes across all-optical and opto-electronic networks. Simulations show the overall effectiveness of this approach, although path dissemination is not studied. In [Zhu 03], the authors study RWA for multi-segment DWDM networks and develop three schemes, namely, end-to-end, concatenated shortest path, and hierarchical routing. The end-to-end scheme assumes a flat globalized graph, the hierarchical routing scheme assumes a hierarchical graph with segments summarized as nodes, and the concatenated shortest path scheme simply uses local information for segment-by-segment routing. Results for a specialized mesh-torus topology show significant blocking reduction with the end-to-end and concatenated shortest path schemes. However, no intra or inter-domain routing is performed here.

King *et al.* in [King 08] also investigated the computation of end-to-end paths across multiple optical domains. In their work, they describe key features of the efforts in the International Telecommunications Union (ITU-T) and Internet Engineering Task Force (IETF) to achieve path computation for end-to-end connectivity in multi-domain optical transport environments. Key issues related to these schemes, such as topology aggregation, inter-domain paths construction, path computation algorithms and constraints are discussed. They also outline the technical areas that require continued development to enhance existing inter-domain path computation techniques.

Velasco *et al.* in [Vela 08] propose a mechanism for connecting two or more MPLS islands belonging to the same MPLS domain through one ASON/GMPLS domain based on the overlay model where client and server networks do not exchange routing information. The inter-connection is firstly done at the control plane level allowing the OSPF-TE flooding mechanism to advertise the existence of a link between two MPLS islands. No optical resources are used in the transport plane of the ASON/GMPLS network. New LSPs can be routed end-to-end triggering, if necessary, the establishment of LSPs in the ASON/GMPLS domain.

The work in [Liu 06a] incorporates many of the key components of a multilayer solution.

Foremost, hierarchical routing is performed using two-level OSPF-TE to disseminate physical inter-domain link state (simple node abstraction). Novel k-shortest path RWA schemes are developed to condense wavelength/converter resources along routes between border nodes and generate abstract links. The propagated link state then is used by source PCE entities to build aggregated DWDM graphs for LR sequence computation (minimum hop or load-balancing). Then, these lightpaths are signaled and expanded using RSVP-TE with wavelength trace vectors.

2.3.4 Path diversity in multilayered networks

Survivable network design is a subject of many past and present studies. In the past, designing a survivable network involved capacity planning and demand forecast within a single network layer only. Due to the dynamic nature of transport network traffic, design of survivable IP/MPLS networks usually manifests itself in forms of dynamic survivable connection provisioning. As for the design of survivable multi-layer networks, it is a topic that recently attracts both industry and academic interests. For example, AT&T looks at the integrated design of IP-over-optical networks based on the UNI signaling [ATT 09]. Other researchers consider integrated design of reliable GMPLS networks.

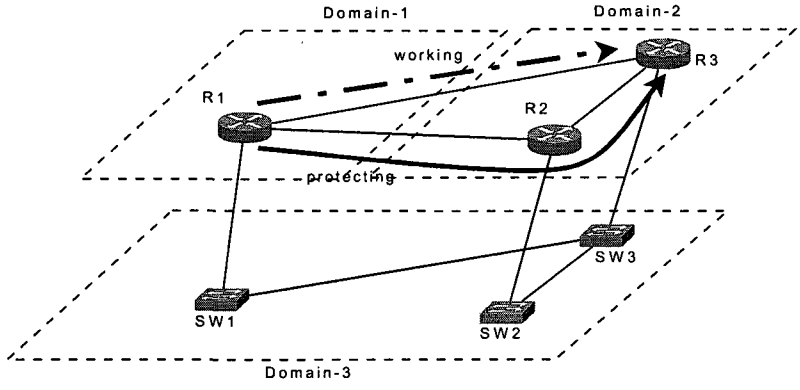


Figure 2.3: Diversity in multilayered networks

In multi-layer networks based on overlay or augmented model, the client is not aware of the routing information in the server layer. Finding diverse protection paths in a layer in a multilayered network is a challenging problem. It is imperative to note that path failure-disjointness at one layer does not necessarily infer physical path disjointness at all lower layers— see Figure 2.3. For example, in a hierarchical optical network an MPLS

packet LSP is tunneled inside a lambda-LSP tunneled inside a fiber-LSP. Multiple fibers are placed into conduits, which are buried along the right of way (ROW). For economic reasons, service providers rent ROWs from third parties, such as the railroad companies. As a result, two diverse fiber links at the OXC layer may be placed into the same conduit at the conduit layer and are subject to a single point of failure. Such links cannot be regarded as diverse links when being used to compute working and protection path pairs. Shared Risk Link Group (SRLG) was proposed to address this problem [Raja 04]. An SRLG is a group of links that are subject to a common risk. The SRLG concept can be applied to physical as well as logical resources at all layers of the network

Therefore, finding a pair of diverse paths at the optical layer involves computing a pair of SRLG-diverse paths. Although the concept of SRLG was originally proposed to deal with conduit cuts, it can be extended to include general risks. For example, all the fiber links located in a geographic area may be assigned the same SRLG considering the risk of earthquakes.

GMPLS networks can be segmented into several areas, such as, wavelength, waveband, and fiber areas in order to optimize the resilience and scalability of the networks as shown in Figure 2.3. The achievable availability of a connection depends highly on the proper identification and selection of the protection layer that containing the client connection.

The practical requirements for multi-layer survivability were examined in RFC3386 [Lai 02] and in [Papa 08], in which the use of nested hold-off timers was recommended. In [Prin 05] Prinz *et al.* present a multilayer protection mechanism that enables a client layer to protect its connections against dual failures in the server layer. The authors present an overview of the different multilayer protection models for client links being protected at server layer (*i.e.*, Optical Transport Network). Specifically they study the case where the client link is protected by optical transport network with path restoration at client layer after second failure, the client link is protected by the optical transport network with additional client layer path protection after first failure, the client end-to-end path is protected using SRLG information provided by the optical transport network. To route the traffic in the client network, an algorithm to route traffic along path with smallest sum of cost. The cost c of a client link is dependent on its current state.

Demeester *et al.* in [Deme 99, Pick 06] studied survivability in multi-layer transport networks. They provided guidelines for coordination of recovery actions in WDM, SONET, and ATM layers. Quantitative comparisons were given in terms of recovery time and investment cost. Fumagalli *et al.* in [Pand 06, Fuma 00] also envisioned the cooperation

of the IP and optical layers in providing network resiliency. A heuristic based on simulated annealing was proposed to choose the optimal protection/restoration scheme for each link of an IP-over-optical mesh network.

Puype *et al.* in [Puyp 05, Puyp 04] propose Multi-Layer Traffic Engineering (MTE) schemes based on two main strategies: a "reactive" one where MTE actions are triggered only by the detection of network congestion and a "proactive" one that tries to keep the network optimal at all times, triggering a reconfiguration whenever optimizations are possible.

Sabella *et al.* in [Sabe 03] propose an offline multilayered solution for the global path provisioning in GMPLS multilayered optical networks. They propose a new heuristic and to solve an optimization problem with help of the CPLEX solver optimizing network configuration and traffic routing considering both the optical and the electrical layers.

Iovanna *et al.* in [Iova 04] propose a hybrid approach for the routing of GMPLS LSPs over a WDM layer which takes advantage of a combined use of off-line and on-line routing strategies to optimize the use of network resources. The proposed heuristic approach is composed of two main phases: (i) an initial paths set-up (performed off-line) by means of a successive shortest-path algorithm (ii) an on-line local search procedure (triggered by network congestion detection) based on the deletion of the lightest loaded lightpath and used for improving the resource utilization to allow the accommodation of incoming LSP requests. Compared to the previous scheme, no details are included here to describe when a link can be considered congested.

In [Larr 05], Larrabeiti *et al.* give an overview and classification of issues related to resilience in a multi-layer and multi-domain network environment. The work shows that, within the context of inter-AS, network resilience based on fast re-routing can be achieved complementing existing interdomain link protection strategies, by using alternative disjoint multi-domain backup paths through other domains. They also propose a specific method to apply MPLS label stacking to make such backup paths scalable to larger clusters of inter-protected domains by means of MPLS transit. They argue that this transit can be used as a temporary solution before BGP routes are stable.

In [Vela 06], Velasco *et al.* propose three routing strategies which differ in the use of failure information, namely *Failure Independent Routing* (FIR), *Failure Driven Routing* (FDR), and *Failure Aware Routing* (FAR). The algorithms take into account the failures in the network at the moment the request is triggered. The the three different routing

strategies are evaluated through simulation experiments run on two reference networks ring and triangular topologies. Performance results are compared in terms of request blocking probability and availability.

2.3.5 Path diversity in multi-domain networks

The problem of finding primary and backup paths subject to QoS constraints in the context of label-switched networks has been widely studied at the intra-domain level. As the network relies more on multiple carriers, issues of survivability across domain/area boundaries are becoming increasingly important. With the advent of the PCE-based architecture, a few recent works have started to extend the study of this problem to LSPs spanning multiple domains. In the current IGP/BGP routing context, a major issue is that the PCE in the source domain has to compute inter-domain LSPs based on a very limited visibility of the topology and state of the network, yielding solutions that are far from optimal. To cope with this, several works suggested aggregating topological and path state information and making this available at the PCE of the source domain [Yann 06b].

Papadimitriou *et al.* in [Papa 05] analyzed suitability of using the GMPLS control plane in multi-region networks. RFC 3386 [Lai 02] studied interoperable survivability approaches in a multi-provider environment. Criteria that trigger protection mechanisms at domain boundaries, as well as requirements on the interaction of protection mechanisms on both sides of a boundary, were suggested.

Sprintson *et al.* in [Spri 07] present a distributed routing algorithm for finding two disjoint primary and backup paths that run across multiple domains. In their proposal, they assume routing decisions in each domain are made by the PCE. The PCEs run a joint distributed routing protocol and run a distributed routing algorithm over an aggregated representation of each domain to enable them to compute disjoint QoS paths across domains. The proposed aggregated representation of the network, however, is suitable in the context of a single layer (*e.g.*, the MPLS packet layer) and hence is only able to compute disjoint paths at that layer without consideration to link SRLG membership. The study also lacks simulation results. Another contribution in [Spri 07] is that we show that the standard approach of representing a multi-domain network by a graph is inadequate for finding disjoint paths subject to the export policies. However, the export policies can be efficiently represented by employing the concept of the line graph.

We show that the distributed routing scheme that we developed for the general setting can be easily extended for finding optimal disjoint paths that satisfy the export policy constraints by using the line graph.

Farkas *et al.* in [Fark 05] and Cinkler *et al.* in [Cink 07b] investigate the use of p-cycles in a multi-domain network to achieve higher availability and acceptable resource without full knowledge intra-domain level internals. In [Fark 05], the authors decompose the multi-domain resilience problem into two sub-problems, a the higher level inter-domain protection, and the lower level intra-domain protection. By doing so, they propose handling shared link protection failures at the higher level (for the inter-domain links) and at the lower level (for the intra-domain links) separately. They present a scheme for building p-cycles at the higher level by straddling link connection, capacity allocation, and path selection at the lower level. Using simulation performed on two different network topologies, the authors present results of average connection availabilities using different schemes.

Truong *et al.* in [Truo 07] consider the dynamic routing problem for Overlapping Segment Shared Protection (OSSP) in multi-domain networks. They propose a two-step routing solution in multi-domain networks based on a topology aggregation scheme and link cost. A greedy and a dynamic programming algorithms, GROS and DYPOS, with and without "crackback" option are proposed. The proposed heuristics are evaluated by comparing performance results for simulation runs over several multi-domain network topologies against a single-domain optimal solution. The blocking probability is examined under dynamic traffic for several network topologies.

Gao *et al.* in [Gao 08] present two distributed end-to-end shared restoration algorithms in a multi-domain network environment. The first proposed algorithm yields a pair of "link-disjointed" paths between a pair of nodes. The second algorithm computes a pair of "domain-disjoint" paths by abstracting each domain by its border nodes interconnected by point-to-point virtual paths. The performance of both algorithms are evaluated and compared through simulation experiments. However, since both algorithms assume exchange of summarized link state information between domains, they are mostly applicable to inter-domain intra-carrier environments where security and scalability of exchanged information is not a concern.

In [Ricc 05] Ricciato *et al.* compare the performance of some recently proposed distributed schemes for disjoint path computation of inter-domain LSPs. They assume that the AS-level path was previously computed by BGP at the source domain and that both

disjoint paths belong to the same "chain" of domains. This approach has two major limitations. First, solving the two disjoint paths problem restricted to the AS-path selected by BGP will frequently return paths that are far from optimal. This is because BGP does not offer any guarantee about the quality of the chosen AS-path. Second, when several disjoint LSPs need to be established following the same (or part of the same) AS-path, crankback [Asla 07b], or even blocking might occur, even though the paths could have been established along the alternative AS-paths available at the source domain

They propose a PCE-based solution to dynamically compute multidomain diverse paths in MPLS networks. Their distributed schemes are based on collaboration between PCE in each domain. Three alternatives are proposed, and are quantitatively assessed for their performance using simulations over real ISP topologies.

The authors in [Cnod 03] present a technique to compute two link or node-disjoint paths for LSPs protection. The proposal does not assume existence of a per-domain entity (or multiple entities per-domain) responsible for the path computation on behalf of all the ingress ASBRs of the domain along the disjoint LSPs. Thus, they do not consider the simultaneous computation of the disjoint paths and we do not require that the disjoint paths follow the same AS path. We note that the consecutive computation of the disjoint paths is subject to the trapping problem [1]

Huang *et al.* in [Huan 04a] present a solution for the setup and maintenance of independent protection mechanisms within individual domains and merged at the domain boundaries. The proposed solution to solve the inter-domain MPLS recovery problem is based on the establishment of primary and backup LSPs relying on basic information provided by neighboring domains and makes no assumption on protection mechanisms of other domains and level of cooperation. Each primary LSP is protected by three or more distinct and independent protection regions merged at their respective boundaries. Those protection regions are the Source Protection Domain, the Domain Interface Protection and the Destination/Transit Protection Domain. They present results of simulations runs using two models: a simple BGP network consisting of three independent domains, and an MPLS recovery using an end-to-end path protection mechanism. They show through simulations the superiority of the MPLS proposed scheme over BGP recovery schemes.

[1]The trapping problem concerns the fact that the path chosen for an LSP may not allow to find a disjoint path even if two disjoint paths exist.

2.4 Review of inter-domain path computation schemes

Unlike in the single domain context, path computation in the inter-domain scenario has to deal with the lack of routing information from other domains. This limitation has two main dimensions: lack of knowledge of AS inter-connection graph and lack of knowledge about the neighbors' intra-domain routing.

The Border Gateway Protocol (BGP) is the inter-domain routing protocol used in the global Internet [Rekh 94]. BGP is a path-vector routing protocol, which only handles and exchanges reachability information between Autonomous Systems (ASs). In other words, BGP routers do not exchange network "state" information. BGP routers only handle destination prefixes, and the Next-Hop to reach a destination prefix. The approach of handling solely reachability information in BGP has proven to supply a highly scalable inter-domain routing framework, but unfortunately, it hinders the deployment of QoS across domains.

Several alternatives to enhance BGP to control the path of the inter-domain traffic have been proposed in the literature. For example, Agarwal *et al.* in [Agar 03] propose changes to the Internet architecture by using an overlay— called Overlay Policy Control Architecture (OPCA)– to BGP to transmit control information. The objective is to be able to control the flow of the incoming traffic of an AS by negotiating the selection of inter-domain paths with remote ASs. The authors also consider the problem of reducing fail-over time in case of failure of an interdomain path.

The solution proposed in [Quoi 05a] also relies on cooperation between ASs. The authors propose *Virtual Peerings* to establish IP tunnels between a border router of a source AS and a border router of a destination AS. An EBGP session is used over the peering link to advertise the prefixes that are reachable via each AS. These tunnels can be negotiated by using backward compatible modifications to the BGP. By using *Virtual Peerings*, the source and destination ASes can achieve various traffic engineering objectives such as traffic balancing or reducing the latency. The authors argue that their solutions does not require cooperation of the intermediate ASes and that it can be incrementally deployed in todaySs Internet. They show by simulations that in a load-balancing scenario, a multi-homed AS only needs to request a few dozens of *Virtual Peerings* to balance its incoming traffic.

In [Yann 05] and [Ho 04] the authors define an architecture with a centralized entity inside each domain. They propose to define a new inter-domain routing protocol to be used between the entities, either called Inter-Domain Routing Agent or Routing Control Platform. They propose to exchange QoS information with this routing protocol. A mechanism for the negotiation of Service Level Specification (SLS) is also defined in [Howa 05] as a support for the provision of QoS-based services.

Several proposals have been proposed within the Internet Engineering Task Force (IETF) in order to deal with the computation of inter-AS paths. Recently, two inter-AS path computation methods have been standardized [Vass 08b] [Farr 06a] [Dasg 08], namely: 1) PCE-based computation, and 2) per-domain path computation. Depending on the service provider requirements and functionality available on nodes, one may adopt either one of these techniques.

Within this context, there are three methods defined for inter-AS MPLS tunnel establishment/provisioning [Asla 07a]:

Contiguous LSP A contiguous LSP is a single end-to-end LSP established across multiple domains. The actual setup requires RSVP-TE signaling procedures described in [Awdu 01]. With this approach a single end-to-end LSP is signaled from ingress label switched router (LSR) to egress LSR, across AS boundaries. Reoptimization is entirely under the control of the head-end LSR and requires a new RSVP-TE procedure defined in [Vass 06].

Hierarchical LSP LSP hierarchy (also known as nesting [Komp 05]) allows one or more LSPs, along a common part of their path, to be carried within a single hierarchical-LSP (H-LSP). A pre-established H-LSP— also referred to as Forwarding Adjacency (FA-LSPs)– that traverses many links may be advertised as a single link. When nodes advertise an H-LSP as a TE link, other nodes can include this TE link in their path computations, as they would for any other link, without knowledge of the existence of the underlying H-LSP. In this case, an LSP using one or more H-LSPs appears as a single end-to-end LSP in the data plane. The number of LSPs to be maintained in the core is minimized thanks to hierarchy.

Stitched LSP With this approach, an end-toend inter-AS LSP is stitched, in each AS, with an intra-AS LSP, also called an LSP-segment, following LSP stitching mechanisms

defined in [Ayya 08]. The main difference between LSP stitching and LSP hierarchy is that LSP stitching relies on one-to-one mapping between inter-AS LSPs and LSP-segments, while LSP hierarchy can use the same FA-LSP to transport several inter-AS LSPs.

Few papers have discussed solutions to allow the establishment of LSPs across AS boundaries. In [Okum 01], a solution based on the utilization of a Specialized Bandwidth Broker agent relying on the inter-domain signaling protocol is proposed. Extensions for local link, node, SRLG protection of inter-AS LSPs has been proposed in [Pels 03], as well as extensions to RSVP-TE to enable it to establish inter-AS LSPs with local protection while still preserving the confidentiality requirement of network SPs. As study for the performance analysis for the mentioned path computation have been presented in [Dasg 08].

2.4.1 Hierarchical path computation

As mentioned in Section 2.3.2 aggregation as a way for achieving scalability and security in global networks has been studied in the literature. The benefit of a larger, more accurate representations is better path selection. For example, a connection setup slated to travel through some domain might discover upon arriving in the domain that the intended transit is actually not feasible. When this type of blockage occurs (due to a stale information) the connection setup is blocked and an error usually retraces its steps backwards to either the border node entry of the domain, or towards the originator of the request. Once, a connection has been blocked, a new path can be computed and another signalling attempt can be retried.

In [Guo 98b], Guerin and Orda suggest several heuristics to path selection accounting for inaccuracy due to state aggregation and distribution. In [Xie 97], Xie *et al.* propose a hierarchical reduced load approximation to evaluate PNNI routing. However, no simulation results are presented. Some graph theoretic studies examined compact graph representation

Apart from routing, other efforts also studied signaling crankback in multidomain IP/MPLS networks. For example, Alsa *et al.* in [Asla 07b] detailed a compute while switching (CWS) scheme in which per-domain computation is used to set up an initial feasible route. Although data transmission is started on this initial route, crankback is used to search for more optimal routes (requiring new RSVP-TE attributes). Results show

very high setup success. Others (*e.g.*, in [Quoi 05b]) also have looked at multidomain survivability by using per-domain primary/back-up LSP routing using BGP routing tables. However, this approach is very suboptimal and can easily overload heavily-traversed interdomain links.

To address these limitations, [Spri 07] develops a more advanced scheme to capture domain-level diversity. Commensurate dedicated protection schemes also are developed using Surballe's algorithm for trap topologies. However, generated state is quadratic in nature, that is, $O(N^4)$ for N border nodes, and must be flooded at the inter-domain level. Moreover, detailed performance results are not presented.

In [Truo 06], Truong *et al.* study the problem of multidomain shared-path protection. Specifically, an aggregated graph is defined with (full-mesh) virtual links along with primary/back-up loose-route computation and sequential signaling. Results show that the schemes are very competitive with an idealized flat routing, but no details are presented on virtual link computation. More recently, the authors also extended the above to consider back-up path reoptimization, yielding moderate blocking reductions in the 5 percent range.

In [Szig 04] and [Szig 05] the effects of the delay while flooding information and the period of and the trigger-threshold for starting information flooding in multi-domain networks are investigated. In [Loja 05] a game theoretic approach is proposed to analyze what effects the pricing policy of certain operators has on the blocking and the income of other operators in a multi-provider/operator environment.

QoS with hierarchical approaches One of the main components of a TE system is the ability to compute and solve the problem of finding primary and backup paths, wherein each of these paths simultaneously satisfies a set of independent QoS constraints. This Multi-Constrained Path (MCP) problem is a complex problem (typically NP-hard) and is the focus of what is known as Quality of Service Routing (QoSR).

In the last few years, QoSR has been recognized as a strong need both at the intra-domain and the interdomain level, and it is certainly expected that this need will also be present in the future optical-based Internet.

In general, the provision of certain QoS requirements between two end-points in a network depends upon the performance properties of individual network elements such as links and nodes (*e.g.*, delay, loss rate, error rate, *etc.*).

QoS routing tries to select a feasible path that satisfies the set of required constraints, while also achieving overall network resource efficiency. It has been found that computing a path that is subject to multiple additive constraints is an *NP*-complete problem with complexity of finding a solution, in the worst case, growing exponentially with the size of network.

Several approaches have been presented in the literature to tackle this problem. These usually fall under two categories: per-domain path computation, and the PCE-based approaches. Depending on the service provider requirements and functionality available on nodes, one may adopt either one of these techniques. In the next sections, we review the different works presented in the literature relating to each of the two techniques.

For example in WSON, QoS-based routing affects the routing decision (*i.e.*, choice of traversed links), as well as the selection of dedicated wavelengths. The objective of a QoS routing scheme, in this case, is to select network paths with sufficient resources to satisfy a connection's QoS request. In general, the provision of certain QoS requirements between two end-points in a network depends upon the performance properties of individual network elements such as links and nodes (*e.g.*, delay, loss rate, error rate, *etc.*). QoS routing tries to select a feasible path that satisfies a set of required constraints, while also achieving overall network resource efficiency. It has been found that computing a path that is subject to multiple additive constraints (*e.g.*, delay, SNR degradation) is an *NP*-complete problem that cannot be exactly solved in polynomial time [Gare 79].

2.4.2 Per-domain path computation

Inter-domain next-hop routing The simplest technique to establish an inter-domain TE LSP is for the LSP path to follow the inter-domain next-hop (NH) routing at each hop along the path. In this case, LSPs are hop-by-hop routed allowing each node to independently choose the next hop for a destination.

Using this scheme, the RSVP Path signaling messages used to signal the LSP are routed at each hop by performing a IP route lookup for the preferred next-hop to the destination end of the LSP (*e.g.*, using the IGP routing table for intra-area nodes). This path is usually chosen by Label Distribution Protocol (LDP) LSPs. Another solution is the utilization of the BGP extension defined in [Rekh 01] to distribute MPLS labels and thus establish inter-AS LSPs. For an inter-domain LSP, the BGP routing table can be consulted so the RSVP Path signaling message can take the preferred next-hop domain

path towards the destination LSP's prefix (usually referred to as the hot-potato path). One drawback of this approach is that with BGP, the inter-AS LSPs are established without being able to specify bandwidth or fast restoration constraints. Moreover, at times, these shortest paths may become overloaded or may not respect the desired QoS.

[Bouc 05] propose extensions to BGP in order to advertise the QoS of the interdomain routes. These extensions have however not been evaluated nor deployed, even though they are likely to generate a lot of signaling messages due to the dynamics of the QoS information [Ho 04] already considers the availability of such extensions for the selection of an egress router for the provision of a service with bandwidth guarantees.

The per-domain path computation scheme defines a technique for establishing inter-domain LSPs where the path is computed during the signaling process on a per-domain basis. It is notably a combination of the inter-domain IP NHOP routing heuristic, and the intra-area path expansion scheme.

The mechanisms for inter-domain TE LSP computation described in [Vass 08b] are applicable when the full "strict path" of an inter-domain LSP path is not pre-determined prior to initiating the signaling of the LSP. This usually arises from the lack of TE topology visibility of neighboring TE domains when network state information is not exchanged across domain boundaries. In this case, a set of (loose and strict) hops can be either statically configured at the head-end LSR or dynamically computed at signaling time.

The RSVP-TE extensions [Awdu 01] makes it possible to indicate the path or a portion of the path to be followed by the LSP inside an object called the Explicit Route Object (ERO). The ERO expansion technique relies on this object and consists of completing, at the entry BN of a domain, the path computation up to the next IGP or BGP hop, *i.e.*, last reachable hop toward the destination. This node is usually either the first hop inside the downstream domain or the last hop inside the current domain. The computed path segment is then stored inside the ERO of the RSVP-TE Path message and the message is forwarded along the path inside the ERO.

It is worth noting that such path computation technique does not always guarantee finding a feasible constrained path since it relies on heuristics to choose an appropriate NHOP among the available NHOPs announced for the destination (*e.g.*, among the BGP preferred NHOP). Furthermore, it cannot be efficiently used to compute a set of inter-domain diversely routed TE LSPs. For example, a head-end LSR computes a set of strict path hops up to the first BN visible in its TE database (TED) topology and then appends

the path with loose hops for boundary or domain exit nodes in remote domains. When dynamically computed, the loose hops can be learnt through discovery mechanisms such as IGP, BGP, or policy routing information.

The per-domain path computation scheme mentioned above chooses the first available interdomain path that fulfils the TE constraints. Thus, the resulting path could potentially be the worst possible available path. This limitation of selecting the first path in standard per-domain path computation can be overcome by using Computation While Switching (CWS), proposed in [Asla 07b]. Unlike the standard perdomain path computation scheme, the CWS path computation scheme continues the quest for a better path instead of terminating the search at the first available path, by making use of a few additional crankback signaling attributes. The CWS scheme uses simple extensions to crankback signaling attributes while maintaining RSVP-TE scalability. Furthermore, it provides a mechanism to select from a set of candidate paths, each of which traverses the minimum number of domains. Thus, the CWS scheme can guarantee that the resulting path traverses the minimum number of domains. Finally, the CWS scheme exhibits path setup latency similar to that of standard per-domain path computation given in [Vass 08b].

The crankback capability of RSVP-TE to stop the establishment of an LSP when a node cannot compute a path toward the destination and attempt different Next Hop (NHOP) BN . The ABRs can store the list of NHOPs that have already been tried for an LSP and lead to an unfeasible path with regard to the constraints. When no NHOP is found that can complete the path with a segment respecting the constraints, a "crankback" is performed by generating an RSVP Path Error message and sending it upstream. The upstream ABR in turn will attempt the expansion of a new segment avoiding the NHOPs that have already been tried.

Figure 2.4 illustrates the RSVP-TE ERO expansion technique with path computation that takes place at BNs. In this example, an LSP with delay constraint of 100 ms has to be established from S to D. Therefore, the source of the LSP S sends out a Path signaling message to D. Inside this message (a), the source specifies the tail-end and the constraints of the LSP. When the BN receives the Path message, it attempts to expand the a path segment respecting the constraints based on its knowledge of the internal topology and the BGP routes for the destination. Then, it forwards the Path message along the computed path segment.

2.4.3 PCE-based path computation

One of the proposed solutions to aid the establishment of interdomain LSPs is the utilization of a Path Computation Element (PCE). The notion of a PCE, originally named Path Computation Server, was initially introduced in order to solve the specific issue of inter-domain path computation for TE LSPs. The basic idea was to rely on the collaboration of BNs to compute a TE LSP spanning multiple IGP areas or domains.

The PCE may be any of LSR, ABR, ASBR or any dedicated server that participates in the computation of constrained paths. It is usually assigned to each domain and can compute constrained paths segments within its domain. Paths computed by individual PCE are usually referred to "local" path segments. The PCE can also compute an end-to-end path based on the local path segment as well as path segments received from other PCEs. When the head-end and the tail-end of the LSP do not belong to the same domain or, if the LSP has to cross different domains, computation of the path of the LSP is distributed, and multiple PCEs may contribute to the computation of the end-to-end path.

The PCE computes path segments respecting given QoS and diversity constraints based on its own TED. The content of the TED depends on the domain of the PCE and contains at the least the topology of the domain and the TE attributes for links belonging to the domain. In addition, it may contain the TE attributes of the links at the border of the domain, for example the inter-AS links.

The current PCE architecture document [Farr 06a] mentions several solutions for the synchronization of the PCE TED. A first solution is that the PCE will participate in the IGP of the different ASes involved in the interdomain path. This solution is applicable in limited environments, for example when two domains belong to the same company, but we do not expect that it will become widely used, notably due to the confidentiality requirements. The second solution proposed in [Bita 08] is to use an out-of-band TED synchronization. In this case, each PCE will regularly obtain topological information from the neighboring domains by using a mechanism or a protocol that is still to be defined.

PCC/PCEs use the Path Computation Engine Protocol (PCEP) [Oki 08] to comunicate with other PCEs. Nodes requesting a path computation from a PCE are referred to as Path Computation Clients (PCCs). Thus, a PCE asking for a computation from another PCE acts as a PCC.

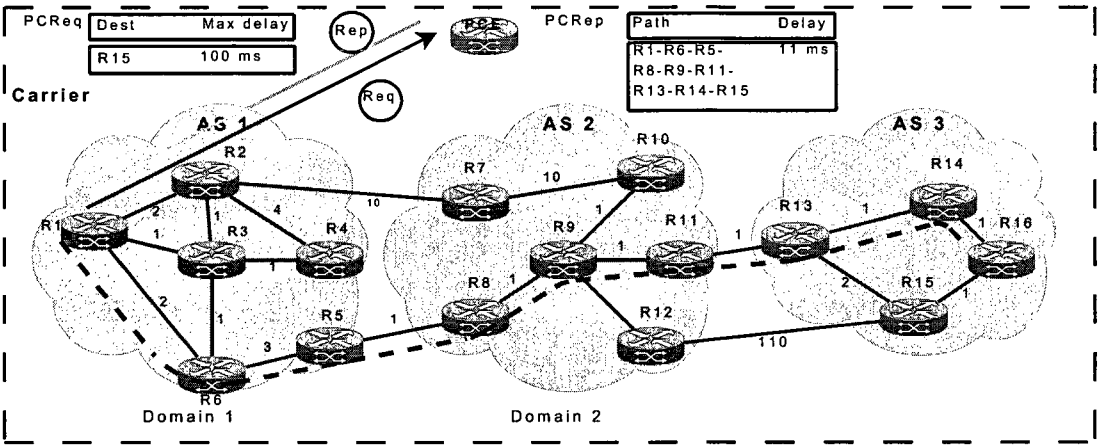


Figure 2.4: Inter-domain centralized path computation with global PCE

Within a single carrier network, a single PCE instance can oversee over several domains/areas. With this approach, also referred to as centralized PCE, the PCE is capable of computing the end-to-end path on its own. For example, Figure 2.4 illustrates this procedure. The PCC initiates by sending a PCReq with the path computation constraints to the centralized PCE. The centralized PCE computes the full end-to-end path and returns it back in a PCReq to the head-end PCC. PCEs may also cooperate with other PCEs in order to request or delegate the computation of path sub-segments contained other domains, regions or layers that requesting PCE does not possess topological information over. The Backward Recursive PCE-based Computation (BRPC) [Vass 08a] has been presented as a collaborative technique for the computation of constrained shortest for inter-AS TE paths– see Figure 2.5.

The PCE selection process is crucial in determining a feasible path as we shall demonstrate in Chapter 5. In [Saad 07], we studied the problem of PCE selection when multiple candidate next hop PCEs are possible. We presented a number of schemes that can be considered to elect the preferred PCE from a set of candidates: a selection scheme using round-robin scheduling, a least-response delay selection, and an adaptive approach based on the individual path computation response times received from each of the candidate PCEs. The third scheme assumes that an average response time is preserved for each of the candidate PCEs and requests arriving at the source are partitioned among the candidate PCEs depending on the ratio of the average response times recorded. Using simulations, we have shown that last scheme results in an improved load balancing

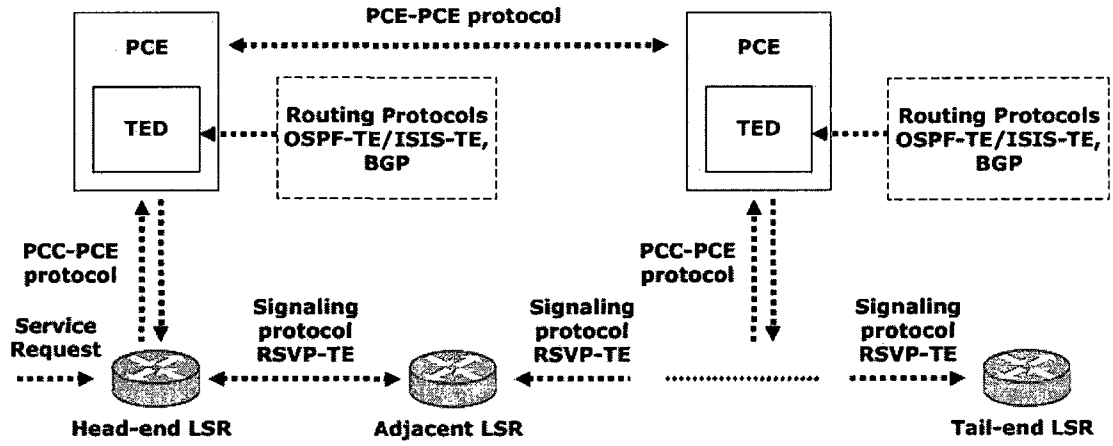


Figure 2.5: Collaborative PCE Architecture

of the path computation requests among the candidate PCEs, and hence minimize the overall path computation time of the inter-area or domain LSPs and the LSP blocking probability.

In [Pels 06] Pelsser *et al.* propose two heuristics so that the PCEs can solve the problem of finding inter-domain LSPs with low end-to-end delay. However, this work addresses the computation of only a single path (without a disjoint counterpart). In addition, the availability of inter-domain paths is inferred directly from the BGP routing information. Accordingly, the authors do not need to address the issue of finding an AR that captures path diversity and the internal structure of the domains.

Two heuristics in this direction have been recently proposed in [Pels 06], but the focus there is on the selection of a single path. The resulting paths under these routing schemes are expected to be of higher quality than those that can be obtained with the current PCE-based approach. Still, these routing schemes cannot guarantee to find optimal QoS paths (*e.g.*, , the shortest paths) across domains. Another alternative that can be used as an interim solution (before the deployment of the PCEs) was proposed in [Rome 06]. This proposal exploits the multi-connectivity between peering ASs in order to find disjoint LSPs along a chain of domains.

Several works in the literature studied the implementation of PCE for inter-domain path computation. For example, in [Moha 05] Mohammad *et al.* discuss the requirements for the inter-AS MPLS and summarize the PCE architecture. In [Gram 07], Grampin *et al.* discuss a management framework for the connection provisioning problem in GMPLS

inter-domain networks.

Douvill *et al.* in [Douv 08] concentrate on the automated provisioning of inter-AS services based on GMPLS technology. They consider a provider alliance where TE connections are established between the members of the alliance, and propose an architecture for the automatic provisioning of inter-AS GMPLS service based on a service plane coupled with PCE. They define the notion of inter-AS GMPLS service as a composition of service elements. They then proposed a comprehensive architecture based on three planes: service, management, and network planes. They outlined the roles of each plane and showed how they can interact. They concluded by showing how inter-AS provisioning can be automated within this framework with an emphasis on service composition and activation.

Torab *et al.* in [Tora 06] study the cooperative path computation scheme where PCEs exchange path information in the context of a specific end-to-end path computation instance. They differentiate between two types of cooperation between PCEs, namely model-based and ad hoc. They show that an end-to-end path computation can use both models to complete the path. They model cooperation mechanism between PCE as a multistage decision problem, and offer a probabilistic analysis. In particular, they argue that having an estimate of the blocking probability in each domain is helpful in determining the path computation effort needed to find an end-to-end path.

Dasgupta *et al.* in [Dasg 07] presents analysis and results that compare performance of the PCE architecture with the traditional "per-domain" approach to path computation. Through detailed simulations undertaken on several realistic scenarios using the two approaches, they present a performance comparison with respect to several critical metrics between the PCE-based and the per-domain path computation deployment. They deducee that several performance benefits from the deployment of PCE can be achieved.

Cugini *et al.* in [Cugi 07] study the cooperative deployment strategies for the PCE in multi-layer GMPLS networks. They discuss the *horizontal approach* which assigns different PCEs to control different layers and highlight the related problems with either the excessive amount of exchanged information through PCEP and/or its inefficient network resource utilization. They present the *vertical approach* as an alternative to the horizontal approach approach's limitations. They carry out simulation for multiple scenarios. Their simulations included results for the blocking probability of LSPs against dynamic traffic load with both the horizontal and vertical approaches. They conclude based on their results that their proposed approach avoids excessive PCEP exchange of

information and the waste of resources due to inefficient Traffic Engineering.

2.5 Review P2MP TE path computation schemes

As mentioned, previous work within the context of inter-domain (AS) mostly focused on point-to-point MPLS. The success of IPTV and, more generally, of multimedia transmissions over the Internet have recently increased the interest in Point-to-Multipoint (P2MP) transmission services.

In the literature, the Border Gateway Multicast Protocol (BGMP) [Thal 04] and Decentralized-Core-based Tree (DCBT) [Kim 99], which are used for multidomain multicasts, are extensions of a corebased tree [Ball 97]. However, they are not supposed to be utilized to find the cheapest multicast tree for traffic engineering purposes. In [Mats 06] extensions are proposed to BGP protocol to allow signaling of P2MP LSPs. The multicast source discovery protocol [Fenn 03] was proposed to find multicast routers in multiple domains. However, the protocol that is used for the route selection between domains is based on the hop number between the domains, and link costs and bandwidth between domains are not considered.

Within the context of GMPLS label-switched networks, operators are considering the deployment of P2MP TE LSPs in order to improve QoS, service availability, and reduce core complexity. Requirements and extensions for signaling P2MP TE LSPs have been developed recently and published in RFC4461 [Yasu 06], RFC4875 [Agga 07], [Mats 08], and [Chai 08a]. Other works concentrated within the intra-domain level where ingress nodes has the full topology visibility. For example, in [Taka 05] two packing algorithms for TE-LSP are proposed in order to efficiently accommodate P2P and P2MP TE-LSPs in a MPLS network. The proposed heuristic considers the residual bandwidth of each link in the network and the number of hops from the ingress node to the egress node(s). They show using simulations that these heuristics calculate more efficient routes and accommodate more LSPs than a traditional routing algorithm, namely the interior gateway protocol (IGP).

In [Secc 08a], Secci *et al.* study the problem of inter-domain AS tree selection for multipoint tunnel set-up within an alliance of ASs. They first describe the framework based on the introduction of a service plane for automatic multi-domain service provisioning. They introduce an abstract representation of domain relationship by means of directional

metrics which are applied to a triplet (ingress point, transit AS, egress point) where the ingress and egress points can be ASs or routers. Then, they focus on the multipoint AS Selection problem that arises in such an architecture. They propose an original approach that allows one to reach almost optimal solutions with tractable computation times for the constrained Steiner problem. Another contribution in that work is some steps of the proposed heuristic can be precomputed, independently of the tunnel demands. By extensive tests on random topologies derived from the Internet, they show that the proposed heuristic is often equal or a few percent close to the optimal, and that, in the case of precomputation, its time consumption can be much lower than other well-known algorithms.

Matsuura *et al.* in [Mats 07] and Kumaki *et al.* [Kuma 08] propose a hierarchically distributed PCE schemes to cooperatively create appropriate multicast trees for multidomain users. In [Mats 07] three different algorithms to create the cheapest multicast trees in individual domains are used. One of them is a new multiplex-aware-route selection algorithm. Evaluation of the applicability of the three algorithms to various types of domains depending on network conditions is later presented by running simulations.

Vijayalakshmi *et al.* in [Vija 08] propose a multiple-constrained multicast routing algorithm based on hybrid genetic algorithm. Artificial immune based method is used to handle the constraints and it removes the difficulties faced by penalty factor method. The artificial immune algorithm simulates the interaction between antigens and antibodies. Experimental results show that the proposed algorithm yields constrained least-cost solutions for various sized random networks.

The most common solutions are to apply direct path selection heuristics to construct the static multicast tree step by step [Zhu 95]. Other Steiner-tree heuristic algorithms in the literature [Chen 95], [Chai 08c], [Chai 08b] that create minimum-cost multicast trees in a network. For example, in [Chai 08c] and [Chai 08b] Chaitou *et al.* proposes extensions to LDP aiming to achieve better resource optimization. In both extensions, a new leaf is provided a partial tree knowledge, by involving either all the nodes of the tree or only its leaves. The leaf joins the tree by connecting to the closest node among the known ones. Valuable comparisons with RSVP-TE are performed, and they represent an important background to decide when and how to use each protocol. In [Chai 08b], they propose the concept of multi-point to multi-point TE-LSP that connects a group of nodes called leaves, acting as senders and/or receivers, with potentially distinct bandwidth needs. They illustrate that proposed method allows for important scalability improvements due

to the high reduction in the dependence between the number of states and the number of leaves. Simulation results show that the number of states to be maintained on a router is drastically reduced thanks to multi-point to multi-point TE-LSPs, and this leads to significant reductions in memory consumption.

2.6 Conclusions

In this chapter, we presented a survey of the state of art on the path computation of GMPLS LSPs that spans multiple inter-domains. As well, we surveyed different techniques presented in the literature that have dealt with topology aggregation in large scaled networks as a way to achieve scalability and security between neighboring competing domains. Within the context of diversity and path protection, we surveyed recent work presented on this topic of establishing disjoint paths across multiple domains and/or switching layers. As well, we have researched the current state of art on inter-domain RWA in WSON networks. Finally, within the context of P2MP communication, we reviewed several recent works presented that have dealt with the problem of computing inter-domain P2MP trees for establishing P2MP TE LSPs.

Chapter 3

Centralized Path Computation with Topology Aggregation

3.1 Introduction

Multilayered transport networks run over a variety network switching layers including IP, MPLS, ATM, SONET, and the WDM optical layer. Each technology layer may support a variety of protection and restoration schemes. As well, transport networks also rely on multiple physically or logically divided domains to provide end-to-end services. The typical approach to dynamically provision protected services or connections in a multi-layered transport network is to assume the overlay network model. In the overlay model, each layer is treated as a separate entity with a separate control plane, with lack of ability of one layer to identify or use resources available in other layers. Although this model is simple to implement, it leads to redundant reservation of protection bandwidth. Hence, the coordination between the several transport network layers via an integrated system becomes imperative in order to achieve survivability and efficient use of network resources.

Network survivability relates to the ability of a network to recover from network element failures. Typically, service providers (or carriers) provision backup or redundant network resources to protect their services against network failures. However, providers are equally faced with the objective to efficiently utilize their network resources since they usually run on a limited budget.

To design a reliable network, protection techniques [Rama 03] that reserve backup resources in advance during connection provisioning can be employed to combat network failures. Such techniques typically requires a link-disjoint primary and backup path pair for a connection to survive from a single-failure scenario. It may not be efficient for a connection to survive from failures when connection provisioning is based on a static link state parameter that is unaware of the current network slate or network failure characteristics. The resilience mechanism which uses pre-calculated backup paths and pre-assigned resources, that effects shorter re-routing time, is defined as protection. The mechanism that calculates and sets up backup paths based on topology information update after a failure has taken place is defined as restoration.

One important requirement of survivable network design is the ability to discover diverse paths for connections. Multiple connections are often setup between end-points for primary and backup paths with a constraint of being diversely routed. Establishing a primary working path disjoint from the secondary or backup one at the same layer reduces the chances of losing or dropping traffic for longer time. This can include paths that are link-diverse (*i.e.*, do not share any common link), or node-diverse (*i.e.*, which do not share any node). This characteristic of path computation can become complicated when multiple layers (physical and logical) are considered. With the limited visibility of topology in other domains, or when topology aggregation is used, the computation of diverse paths remains a challenging task. In this case, an intelligent aggregation scheme is required to represent diversity information across vertical and horizontal layers.

Path computation for end-to-end paths across multiple domains or layers is the next step towards wide-scale deployment of a distributed control plane that supports Traffic Engineering (TE). By separating the control plane into path computation and signaling planes, these two functionalities can have different architectures with respect to hierarchy. Specifically, the path computation plane can have a hierarchical structure, allowing aggregation of traffic engineering information in each domain into compact models suitable for path computation (with aggregation applied at each level of hierarchy), and the signaling plane can retain its flat architecture and follow the (maybe loosely) computed path in a sequential manner.

In this chapter, explore the limitations of computing primary and backup interdomain paths for LSPs spanning multiple vertical/horizontal domains. We propose a hierarchical link resource tree aggregation scheme that we refer to as Aggregate SRLG Link Resource Tree (ASLRT) that embeds the SRLG information in each layer and permits the compu-

tation of diverse working and protecting paths with differentiated protection levels along multiple vertical and/or horizontal domains.

In the second part of the chapter, we study the problem of provisioning a lightpath that spans multiple WSON domains. We present a hierarchical approach to routing the end-to-end lightpath across the WSON domains that we refer to as Hierarchical Shortest Path First (HSPF).

3.2 Problem definition

A key aspect in the design of distributed path computation algorithms is to find an adequate aggregate representation at each layer/domain that can capture the availability of diverse paths across multiple domains and layers. The information exchange between different domains at the control plane level is conveyed by the inter-domain routing protocol, that is, the Border Gateway Protocol (BGP). Although BGP supports the distribution of some limited TE information (for example, the use of BGP communities attribute [Rekh 94]), in practice, BGP only advertises reachability information between domains. BGP routers do not exchange network "state" information— such as path bandwidth utilization, SRLG information, path delays, or wavelength availability– crucial for the purpose of finding feasible paths. Moreover, the nature of relationship between different carriers (*e.g.*, customer-provider, or peer-to-peer) and its implication on route export policies in each scenario, makes interdomain routes not easily inferred directly from the topology. These set of rules turn inter-domain routing into being more policy-driven rather than topology-driven or network-state driven, so finding disjoint paths across domains is strictly limited by these rules.

Without the full knowledge of the network topology and state, a head-end LSR cannot reliably choose the "best" path for an LSP or lightpath to take. Instead, the node must try to approximate this function. The goal is to select a path that with a high confidence satisfies the specified constraints without incurring an unacceptable cost in terms of complexity and/or scalability. Generally, there are two ways in which a node can attempt to provide this function.

- push mechanisms where a summary of information about all other domains is pushed down to the nodes in each area.

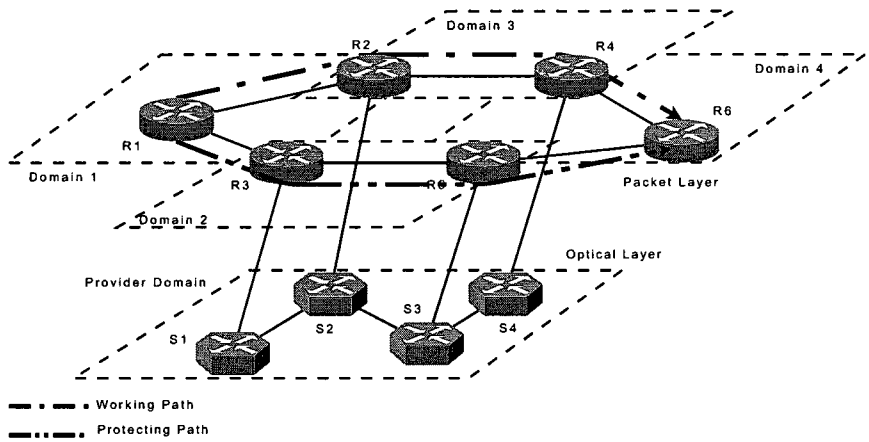


Figure 3.1: Diversity across switching layers

- pull mechanisms where nodes do not need maintain a view of the rest of the network. Instead, triggered with arrival of a path request, they query remote entities (*e.g.*, collaborating PCEs that have a better global view of the network) for an end-to-end path.

Furthermore, there are other complications to setting up protected LSPs spanning different domains. When an end-to-end service is provided across a single provider's network, the LSR has full TE topology visibility over resources within that domain, and can efficiently compute both working and backup paths. However, where an end-to-end connection spans multiple provider networks the full topology visibility is no longer readily available. Due to lack of knowledge of all nodes and links along the LSP's working path, it is a challenging tasks for a Border Node (BN) that perform the sub-path computation to find backup LSP paths that protect against node and/or link failure. Furthermore, the lack of coordination between adjacent layers also makes the computation of end-to-end paths across layers more challenging.

Another important observation to note is that it is not sufficient for two LSPs paths using links in completely different domains or even different carrier networks to imply that their paths are failure diverse. This is because different links at the server layer may use the same physical conduit– for example different optical fibers may share the same cable and consequently share the same fate. For example, consider the two paths between $R1$ and $R6$ *working*=$R1, R2, R4, R6$ and *protecting*=$R1, R3, R5, R6$ in Figure 3.1, one can easily infer that the two paths are domain diverse as well as link/node diverse at

the packet layer. A further inspection of the path along the network hierarchy, however, shows that the IP link $(R2, R4)$ and $(R3, R5)$ share link $(S2, S3)$ at the optical layer. In this case, a failure of link $(S2, S3)$ at the optical layer will affect both the working and protecting paths at the IP layer, and in this case, failure-diversity of the two paths does not hold.

Hence, the computation of failure-diverse paths in a multilayered network involves diversity be satisfied with respect to all underlying switching layers (*e.g.*, packet, wavelength, fiber, duct, *etc.*). The problem, in this case, is that the client does not know anything about the routing and diversity of links (physical or logical) in the server layer.

To solve this problem, the Shared Risk Link Groups (SRLGs) has been introduced within carriers networks. The SRLG is a group of links that share resources whose failure affects all the links in the group. In this case, different links within the carrier network may be grouped together as belonging to a particular SRLG. However, links owned by different carriers can still share the same lower layer infrastructure network(*e.g.*, physical conduits). The concept of SRLG still needs to be extended for the inter-carrier environment to enable nodes across all domains to be able to understand which SRLG is being referred to– *i.e.*, SRLG identifiers need to be standardized across the whole network. This, however, is challenging especially in the inter-carrier environment. A key issue is how to compute paths across multiple areas and/or carrier networks that satisfy the specified constraints.

The degree of availability of an LSP highly depends on the proper identification and selection of the protection layer through which the connection traverses. In general, a connection can be protected against failures that occur at its uppermost containing layer as well as layers that fall below. For instance, in a traditional multilayered IP transport network, a protection layer for a connection can be any of the layers from IP down to WDM.

Due to the failure inheritance properties, a connection's availability would be higher if the protection layer is selected further below the layer containing the connection. Hence, assuming a multilayered transport network, the problem can be addressed by provisioning a failure-disjoint working and backup paths for a client connection at layer t against an SRLG resource failure at server layer $t-1$ below the client layer.

For example, within the context of WSON networks, the following set of protection services may be offered to a generic logical connection crossing a WDM domain based

on the availability requested:

- *Low-resilience requirement*: for traffic protected against failure at the fiber layer,
- *Medium-resilience requirement*: for traffic protected against failure at the conduit layer, or
- *High-resilience requirement*: for traffic protected against failure at the domain layer.

Protection and restoration schemes. In general, there are two types of fault management techniques: protection and restoration. With dynamic restoration, the backup is computed and signaled after the failures happen. The main disadvantage being larger recovery time, mainly due to its dependency on the slow convergence of routing protocols. In protection, however, backup paths are pre-provisioned, and spare capacity is reserved for them at the time the working path is set up. Generally, protection is more resource costly, since it requires pre-allocation of spare capacity for the pre-established backup paths. On the other hand, restoration may take longer to restore the connection, since the restoration process happens in real-time. Protection and restoration have traditionally been addressed using two techniques: end-to-end path switching and local span/link switching.

Local versus end-to-end recovery. End-to-end recovery refers to the recovery of an entire LSP from its source (ingress) to its destination (egress). On the other hand, local/segment recovery refers to the recovery over a portion of an LSP segment. In an end-to-end path protection scenarios, a pre-established backup LSP is set up from ingress to the egress LSR. In order to guarantee maximum availability, the backup LSP path is typically routed over a physically disjoint path from the working LSP. When the ingress LSR is notified of failure on the working LSP (*e.g.*, reception of a signaling RSVP Path Error due to the failure of a network component), the ingress LSR immediately starts switching traffic over to the backup LSP.

For local link protection, a backup LSP can protects against failures on a single link (Link Protection (LP)) or node (Node Protection (NP)). The backup LSP originates at the protection switch LSR– also known as the Point of Local Repair (PLR)– and terminates at the protection merge LSR or Merge Point (MP) where the backup LSP intersect with the working LSP. It is important to note that if a working LSP spans several links, one

backup LSP has to be set up for each protected link along the working LSP path in order to protect the entire working LSP. In the case of LP, the backup terminates at the PLR next-hop LSR, and is referred to as NHOP backup. In this case NP, the backup terminates at the PLR next-to-next-hop and referred to as NNHOP backup. Both LP and NP backups can protect against the failure of an SRLG.

In normal situations, the backup LSP does not normally carry any protected traffic as long any resources along the working LSP have not failed. For efficiency, and to make best available use of network resources, it is common practice to allow backup paths to carry traffic (referred to as low-priority traffic or extra traffic). This low-priority traffic gets preempted by the high-priority protected traffic when the working LSP fails.

3.3 PCE hierarchical-based solution

As described in Section 2.4.3, this approach proposes a decoupled architecture in which path computation tasks are performed by a PCE router present in each domain. This approach is based on the exchange of aggregated advertisements between domains. The network state information of the domain is gathered into a Traffic Engineering Database (TED) on each the PCE or on a border node in each domain.

Domains become capable of exchanging aggregated topology and state information, which can be used to compute the "entire loose" LSP path from the Source domain PCE (SPCE). Hence, a key aspect to this approach is finding an adequate aggregate representation that captures the available path diversity at multiple layers as well as the QoS attributes of the aggregated link(s) across a small group of domains. Certainly, a trade-off exists between the optimality of the resulting QoS paths and the size of the aggregate representation. The advantage of aggregation is that it facilitates the computation of entire (primary and backup) shortest paths directly from the source domain PCE.

Since the source domain PCE only knows an aggregated state of the whole network, the resulting paths are still a mix of strict and loose hops. The list of hops could include the source node, the list of border nodes to be traversed across the different domains, and the destination node. Thus, approaches like this still need to rely on the loose-hops expansion, but with the advantage of increasing the number of strict hops conveyed in the signaling messages.

The extent of aggregation of information and what to aggregate play an important role

in determining feasible paths and improving the chances of establishing an LSP request. Yet typically, very little information is conveyed to non-local areas of the hierarchy, *e.g.*, in the previous example only prefix reachability information was provided by the BGP protocol. However, for some networks, this could cause significant problems when attempting to set up diverse-path LSPs or LSPs that guarantee bandwidth for traffic across multiple domains.

Moreover, for security reasons, an aggregation scheme must not represent any information about a domain's internal structure. The concept of levels can be used to help aggregate properties of a particular domain. One example of such aggregation would be to represent the Internet as a *level-1* graph with vertices representing abstract nodes and edges representing the inter-domain links— we refer to this as the simple-node representation. Another, would be to abstract the domain by its border nodes connected by abstract links— we refer this scheme as the complex-node representation. This can be viewed as a complete weighted graph whose vertices are the border nodes of domains and whose edges represent the costs of the corresponding transits as shown in Figure 3.2. Such a representation may be viewed as a complete weighted graph whose vertices are the BNs of the domain whose edges represent the costs of the corresponding transits. Unfortunately, the size of such a representation grows quadratically in the number of nodes, which implicates scalability problems.

Both aggregation models either require considerable routing updates to be passed to PCEs in each domain (after dynamic aggregation computation within the lower layer or adjacent networks) or serve as approximations that source domain PCE can use without any certainty that a computed path will be satisfied by neighboring or lower-layer domains that will be traversed.

The performance benefits of having higher-fidelity aggregation scheme need to be balanced against real-world burdens imposed by the use of larger representations. Such burdens include having greater space requirements for the topology database within each domain, increase background traffic between domains due to topology updates, and longer computation times for determining the least-cost paths.

As an illustrative example, consider the example of a 2-level hierarchical network in Figure 3.2. In this example, the physical nodes and links are at the lower level of the hierarchy. The higher level consists of abstract nodes and links that are summarization of the lower level topology. Every node in each area discovers link state information about the links and nodes inside that domain. At the higher level, each domain is represented

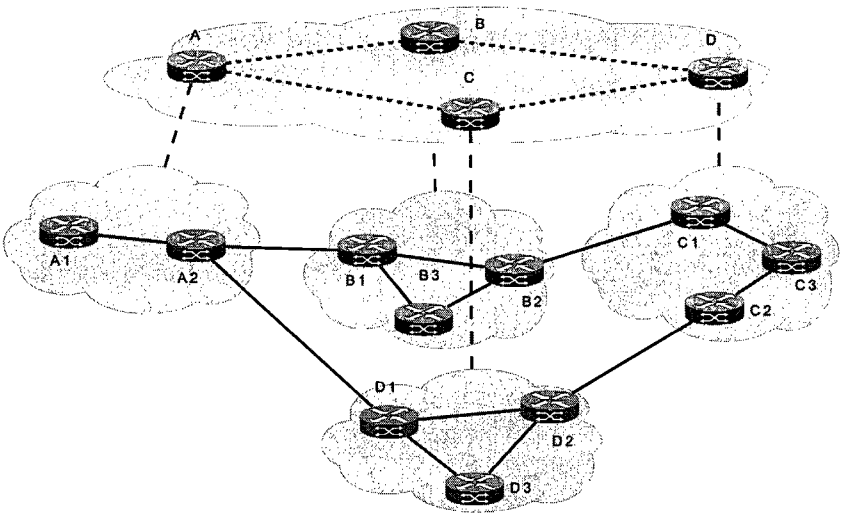


Figure 3.2: Path computation in two level hierarchy networks

by an abstract node. If there is at least one physical link between two domains, then the corresponding abstract nodes have an abstract link between them. The set of abstract nodes at the higher level form a virtual domain, and distribute information about each of the an abstract links amongst themselves. Each abstract node "feeds down" summarized information about these abstract links to the lower level nodes in the area it represents. From the perspective of the lower level nodes, there is now a link between node A2 and domain B.

It is important to note that the above aggregation scheme equally applies to horizontally adjacent domains at the same layer (*e.g.*, peer routing domains), as well as between vertically adjacent domains (*e.g.*, IP packet and optical layer domains). In the latter case, the border model scheme discussed, can be used to flood logical link representations (also referred to as forwarding adjacencies) to higher layers (*e.g.*, packet layer). Consequently, these Forwarding Adjacency (FA) links can be used in the path computation for requests that are originate at packet layer.

The natural representation of a routing domain D_k that is modelled with a graph $G = (V, E)$ is an array that stores, for each pair of BNs bn_i and bn_j of D_k, minimum weight of a path between the two BNs. This representation allows a PCE to compute loose paths that has the space complexity of $O(|V_{bn}^k|^2)$, where $V_{bn}^k = \{v : v \in V \text{ and is a BN in } D_k\}$.

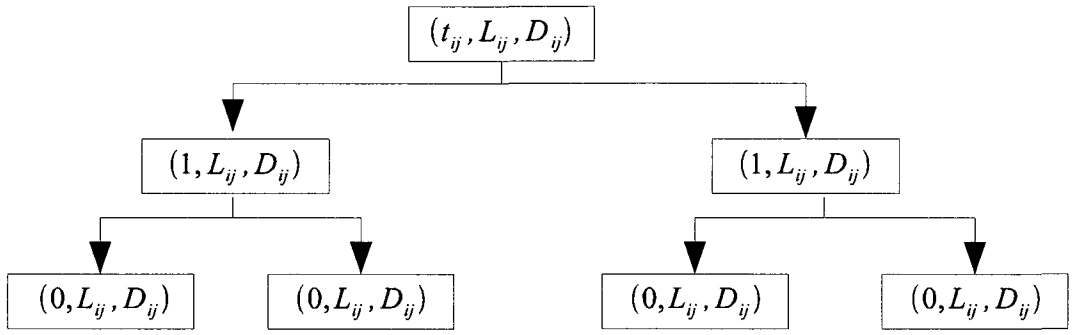


Figure 3.3: Hierarchical link resource tree organization

3.4 Proposed aggregation using SRLG trees

As described earlier, there is a need to for an intelligent aggregation scheme with the necessary SRLG information being present per layer so diverse paths can be computed between end nodes of different domains. To solve this, we propose an inter-domain and inter-layer hierarchical SRLG representation of link. In our proposal, every link at a given layer t is associated with a SRLG Link Resource Tree (SLRT), as shown in Figure 3.3. The root of this tree is the SRLG and resource parameters of the link at that layer, and the leaves are the SLRTs of aggregate or physical links that the given link is tunneled through at lower layers.

The concept of SLRTs generalizes the notion of link diversity to take into account multiple layers and domains. For example, to provide failure-diversity at a client layer, we propose an aggregation scheme for diversity at the server layer. Here, the upper layer is referred to as the client layer, and lower layers are referred to as the server layers. In such a topology, a link at the client layer can traverse many nodes and links in the server layer (for example, optical and fiber layer). For example, multiple IP flows tunnels inside same MPLS LSP, multiple lightpaths can be tunneled inside a wave-band link, multiple wave-band links into a physical fiber link, multiple fibers into one cable, multiple cables into one fiber-conduit, and multiple fiber-conduits into a Right of Ways (RoW). Lastly, multiple RoWs may run through the same risk domain, which usually is defined geographically; the hierarchical structure of the SRLG in this case is depicted in Figure 3.4.

Figure 3.3 shows how a link at client layer t can be associated with a SLRT (t, L_{ij}, D_k) where L_{ij} represents the link identifier, D_k the domain identifier that the link belongs to, and the triplet (t, L_{ij}, D_k) signifies the link's SRLG attribute at layer t or SRLG type-t.

The SRLG type-1 becomes the head of a SLRT tree whose leaves are at the layers below. The above procedure essentially creates a hanging SRLG tree from every link at each layer.

The SLRT tree conveys to the client layer topological information changes for links in lower layers. For example, if a link at a lower layer fails or is taken out of service, links at the client layer that are tunneled inside of it will also be affected– *e.g.*, failed, or notified of the change in state. These links can easily be identified if information about the failed link is propagated upward along the SLRT. In addition, the below are other inherent properties we infer about SLRTs:

> ***Multiple inheritances***: a child resource can have more than one ancestor resource. In this case, for example, a failure of a given link may be caused by a failure of one of many links at the lower layers through which the given link is tunneled in.
>
> ***Failure propagation***: the SLRT conveys to the client layer the topological changes in the lower layers. If a link in a lower layer fails or is taken out of service for maintenance/upgrade, all the links in the client layer (upper layer) tunneled inside the server link are directly affected and are also taken out of service.
>
> ***Protection bandwidth propagation***: A resource may carry the working paths of several connections established at any particular client layer. If the resource fails, the connections are restored (if protection is requested) by redirecting them to their backup paths in the client layer. In order to restore these connections, some amount of backup bandwidth will be required on the links along the backup paths. A failure of a child of the above resource will require at least the same amount of backup bandwidth on these links as well. Hence, when an amount of bandwidth is reserved on a link at any given layer it must also be reserved on all links at lower layers through which the link is tunneled in.
>
> ***QoS parameter propagation***: A link typically carries quantitative performance attributes that various classes of service normally are required to meet. An example of such attributes are packet-layer specific (*e.g.*, delay,

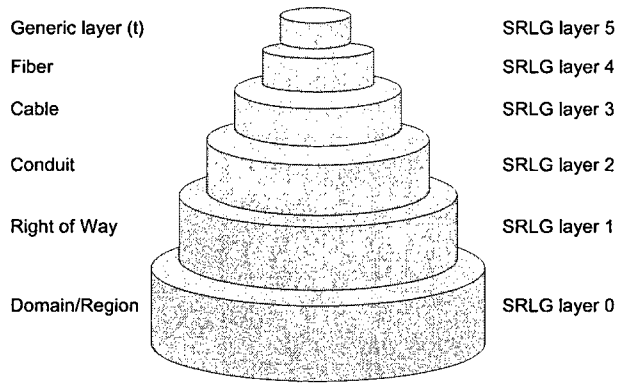


Figure 3.4: Hierarchical SRLG tree for multi-layered transport network

delay-jitter, packet loss-rate), or optical-layer specific (*e.g.*, available wavelengths, *etc.*). In this case, each node in the SLRT carries link attribute information about at that specific layer. Based on this metric, if needed the ancestor quality attribute can be derived from the children's parameters by either recursively adding (*e.g.*, for additive metric like delay), multiplying (*e.g.*, for multiplicative attributes like availability), or choosing the minimal value (*e.g.*, available bandwidth or wavelengths).

The applicability of the SLRT tree organization per link presented earlier can also apply to logical or aggregate links— defined as sequence of physical links at the same switching layer and included in the same domain. The domain, in this case, can be a group of resources (nodes and links) that provide similar capabilities and share the same set of risk(s) (refer to Section 3.3). This aggregation of SRLG resources in a domain can be useful in summarizing and reducing the amount of information propagated in the routing protocols across layers and in hiding the topology of the domain for the sake of loose path specification, and distributed diverse path calculations.

The logical or abstract links that connect BNs within a domain can be represented by Aggregated SRLG Link Resource Trees (ASLRTs) and utilized in provisioning of inter-domain failure-diverse working and backup paths. The relevant layers of a multi-domain transport network can be represented as shown in Figure 3.4.

The ASLRT can be formed between Border Nodes (BNs) in each domain as follows. For each BN-pair in the domain, one or more FA paths are computed (*e.g.*, an FA per class of service supported by the domain). Each FA LSP is composed of a concatenation of a

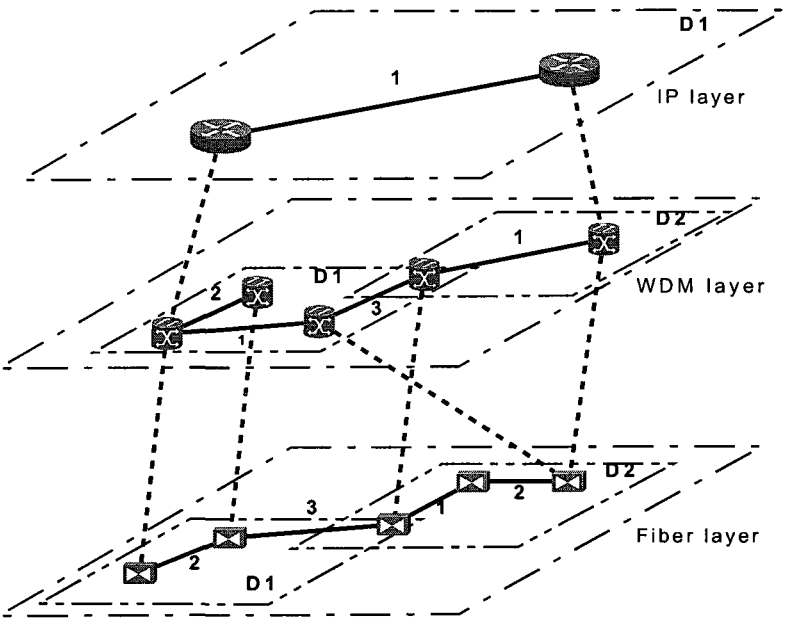


Figure 3.5: Example of a hierarchical ASLRT organization

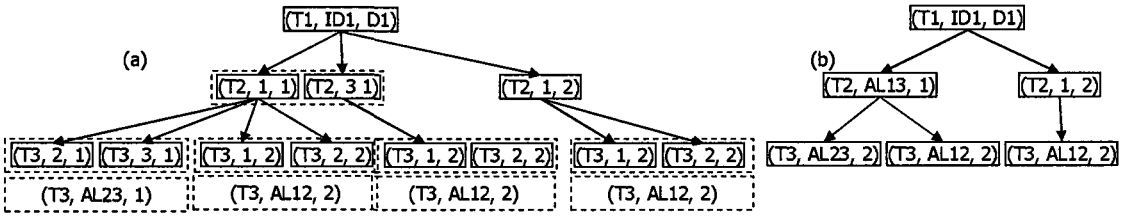


Figure 3.6: ASLRT formation: (a) relaxed individual trees, and (b) aggregated tree

number of individual links, each link is associated with an Individual SLRT (IT) consisting of ancestor and children resources as described in previous section (see Figure 3.3). The ASLRTs of the links between the BNs become supersets of all the information contained in ITs along the path as shown in Figure 3.5 and Figure 3.6. Each ASLRT is defined by one SRLG type-1 head and several type-n leaves underneath. The following steps can be followed for the formation of such ASLRT:

- Replace all type-1 heads (at switching layer 1) by one head in the ASLRT.
- If any of type-1 SLRTs in any of the ITs is connected to a t ($t > 1$) resource, a type-t resource is also connected directly to the head in the ASLRT
- Starting with the head of the formed tree, for all successors replace all individual

links that share the same domain D_k.

- The trees are traversed a step lower, and above steps are repeated for the type-2 resources, and so on.

The use of ASLRT for logical links offers benefits since only a few links need to be communicated to the routing hierarchy and can be updated as the resources are used. Moreover, at a certain switching layer need only flood partial ASLRT to neighboring PCEs in domains of the same switching layer. Since availability and capacity of potential links can change continuously as resources are used or freed within the domain, the aggregation can re-computed and new information flooded to participating PCEs in each domain.

3.5 Hierarchical backup path heuristic using ASLRTs

In this section we describe a heuristic for computing a backup inter-domain multilayered path after a working path has been setup. We restrict our study to the end-to-end path-protection against single point failures. Assuming a connection with demanded bandwidth b, and requested dedicated protection level at layer t. We start by defining the following notations:

L_{wp}: set of aggregate links at the client layer that are along the working path of the newly arrived connection. This set is created when the working path is computed

Q_{wp}: set of links at the client layer that are not along the working path P_{wp}

$S_{wp}(t)$: set of SRLG resources at layer t that are along the working path wp

Q'_{wp}: set of aggregate links in Q_{wp} whose SLRTs have leaves in $S_{wp}(t)$

R_{wp}: union of L_{wp} and Q'_{wp}; that is $R_{wp} = L_{wp} \cup Q'_{wp}$

L_{bp}: set of links at the client layer that are along the backup path. This set is created when the backup path is computed

After the path computation process computes a feasible working path wp, $S_{wp}(t)$ is identified by accessing each member of L_{wp} in turn, and retrieving and placing the elements from ASLRT at row t in $S_{wp}(t)$. The elements in $S_{wp}(t)$ represent the SRLG resources that are not failure-disjoint with the working path and should not be traversed by the failure-diverse backup path bp.

An aggregated link in Q_{wp} may have some SRLG leaves that are in $S_{wp}(t)$. These leaves can be identified by comparing the elements at row t of that ASRLT tree with the elements in $S_{wp}(t)$. If the ASRLT has at least one common SRLG element with $S_{wp}(t)$ the link will be placed in Q'_{wp}. The union of Q'_{wp} and L_{wp} constitutes all the links at the client layer that must not be traversed by the backup path. For the purpose of backup path computation, these links can be assigned a large, or *infinite* cost so they can be avoided the computation of the backup path.

For the computation of the inter-domain backup path, we assume the presence of a hierarchical fully-meshed topology at the source domain PCE. The backup, can be computed using a two-step sequential heuristic that computes a pair of diverse working and backup paths for a newly arrived connection. In the first step, a working loose path is computed by the source-domain PCE based on the available aggregated topology. In the second step, the SPF algorithm is executed again, with the logical links that are sharing the same SRLG in the aggregate link costed-out (*i.e.*, their cost set to infinity) as discussed earlier. This step identifies a feasible backup path that is failure-disjoint from the path computed in step 1. The centralized PCE can set the weight of each link according to:

$$c_{ij} = \begin{cases} \infty, & \text{if link } (i,j) \in R_{wp} \\ \infty, & \forall w_{ij}^a < 1,\ (i,j) \in E \\ -log(1 - \frac{1}{w_{ij}^a}) & \forall w_{ij}^a > 1,\ (i,j) \in E \end{cases} \tag{3.1}$$

where w_{ij}^a is the available wavelengths on link (i,j), and the term $-log(1 - \frac{1}{w_{ij}^a})$ – the penalty function– denotes the measure of willingness a link offers to accept a call request.

3.5.1 Complexity considerations

We consider the scalability of the proposed aggregation scheme from computation complexity in the context of the backup path provisioning. The computation complexity relates mainly to the number of basic operations required to compute set R_{wp} for an end-to-end global connection. In order to compute this set, $S_{wp}(t)$ and Q'_{wp} must be computed, assuming that L_{wp} has already been known. The process is separated into two steps. First, an inter-domain backup path is computed. In the second step, a domain local sub-backup path is computed across each transit domain along the global connection's path.

The basic operations required to compute $S_{wp}(t)$ for the interdomain aggregate links connecting domain BNs is similar to the case of a single domain and is in the order of $|L_{wp}|.|L_{(agg,t)}|$, where $|.|$ denotes set cardinality, L_{wp} the set of aggregate links across the working path, and $L_{(agg,t)}$ the set of all aggregate links at layer t. Within each domain, the backup connection demands a complexity of $|L^k_{wp}|.|L^k_t|$, where L^k_{wp} is the set of expanded links for the working path in domain k, and L^k_t is the set of all intra-domain links in domain k at layer t.

The operations required to compute Q'_{wp} include searching for the SRLG resources at layer t in the SRLG tree associated with every aggregate link, checking if these are in $S_{wp}(t)$, and if so, inserting the link in Q'_{wp}. The number of such operation is in order of $|Q_{wp}|.|L_{(agg,t)}|.|L_{(agg,t)}|$. In each domain k, Q'^k_{wp} is further computed with number of operations in the order of $|Q^k_{wp}|.|L^k_t|.|L^k_t|$ with a computation complexity of $O(|L^k_t|^2)$.

3.6 Analysis of LSP blocking probability

The reliance on the aggregation of information as well as periodic network state dissemination to determine an appropriate path for a LSP connection request may sometimes result in infeasible paths being used— ultimately leading to signaling failures at call admission time at some intermediate midpoint node along the path. This type of a blockage (due to a discrepancy between actual and advertised information, or due to changes that might have occurred since the last advertisement) causes the LSP connection request to retrace its steps from the point of blockage back to its original entry point into the domain (*e.g.*, BN)– a process described in previous chapter as a crankback. Once the request has returned to the point where it first entered the domain or layer, a new path through

the domain can be computed. Otherwise, if no path can be found, the request cranks back further upwards towards its origin as shown in Figure 3.7. The target in general when designing efficient resource and topology aggregation schemes is to minimize such blockage at signaling time.

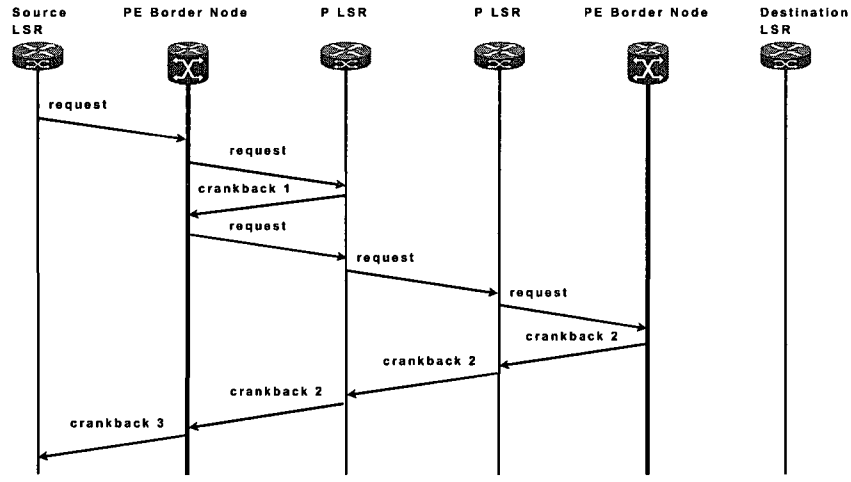


Figure 3.7: Crankbacks with inter-domain signaling

It is typical to define an objective function to evaluate network performance and optimization for different proposed techniques. One natural choice to be considered is the total *network revenue*. According to this approach, LSP requests accepted on a certain path p generate a revenue at a rate z^p, and the total expected network revenue can be written as:

$$Z = \sum_{\forall p} z^p v^p = \sum_{\forall p} z^p v^p (1 - \hat{P}^p_{block}) \tag{3.2}$$

where v^p is the carried traffic (admitted bandwidth for established LSPs) on path p, λ^p is the total offered traffic on path p, and $\hat{P}^p_{block}$ is the end-to-end blocking probability for path p.

If z^p is proportional to the bandwidth requirement of each LSP request then Z signifies the total carried traffic in the network.

Consider a network with E directed links connected in an arbitrary topology. Denote W_{ij} the capacity of link (i, j) (*e.g.*, maximum reservable bandwidth). At a given instant of time, some of the bandwidth on link (i, j) will be allocated while the remainder free. Let

m_{ij} denote the residual bandwidth on link (i, j), and let m denote the residual bandwidth on all links in the network.

Denote by P_{ij} the set of paths that traverse link (i, j). In order for a LSP connection to be set up on path $p \in P$, at least b units of requested bandwidth must be available at each link $(i, j) \in p$. Denote the rate at which LSP connection requests are *set up* on path p when the network is in state m by $v^p(m)$. In this case, $v^p(m)$ satisfies:

$$v^p(m) = 0 \text{ (blocked) if } m_{ij} < k \text{ for some } (i, j) \in p. \tag{3.3}$$

Consider the situation where LSP connection requests arrive to path p at rate λ^p and an LSP request is setup on path p if and only if $m_{ij} \geq k$ for all $(i, j) \in p$. Thus,

$$v^p(m) = \begin{cases} v^p, & \text{if } m_{ij} \geq k \ \forall (i, j) \in p. \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

3.6.1 Unprotected LSP Blocking

Consider the situation where an ingress LSR attempts to signal an unprotected LSP over a path p with n hops. Assuming that an LSP request will be established if at least 1-unit of bandwidth is available on each link, and links have equal capacity W.

Let u_{ij} be the probability that a link (i, j) along the path of the LSP is busy (*i.e.*, $m_{ij} < W$, at least 1 unit of bandwidth is in-use). The expected bandwidth allocation on link (i, j) can be written as $u_{ij}W$. An LSP from s to d is blocked if at least one of the links along the path has reached its full capacity (*i.e.*, fully utilized).

The probability that the LSP request is *blocked* at any link (i, j) is:

$$\hat{P}^{ij}_{block} = (u_{ij})^W \tag{3.5}$$

and the probability the LSP is *admitted* at link (i, j):

$$\begin{aligned} \hat{P}^{ij}_{success} &= 1 - \hat{P}^{ij}_{block} \\ &= 1 - (u_{ij})^W \end{aligned} \tag{3.6}$$

Since all links have the same capacity W, the probability that an intra-domain LSP request taking path r gets blocked can be written as:

$$\hat{P}^{lsp}_{block}(intra) = 1 - \prod_{\forall(i,j)\in p} \hat{P}^{ij}_{success}(intra) \tag{3.7}$$

3.6.2 Path-protected LSP Blocking

Given a path-protected tunnel (*e.g.*, 1:1 or 1+1), for an LSP request not to be blocked both the working and backup paths have to successfully signaled on the working and backup paths. The probability $\hat{P}^{lsp}_{block}(pp)$ that such a request gets blocked can be written as:

$$\begin{aligned} \hat{P}^{lsp}_{block}(pp) &= \hat{P}^{lsp}_{block}(work) \cup \hat{P}^{lsp}_{block}(backup) \\ &= \hat{P}^{lsp}_{block}(work) + \hat{P}^{lsp}_{block}(backup) - \prod_{\forall(i,j)\in L_{wp}\cap L_{bp}} \hat{P}^{ij}_{block} \end{aligned} \tag{3.8}$$

Note, for link-disjoint paths the second item can be eliminated and Equation 3.8 can be written as:

$$\hat{P}^{lsp}_{block}(pp) = \hat{P}^{lsp}_{block}(wp) + \hat{P}^{lsp}_{block}(bp) \tag{3.9}$$

3.6.3 Inter-domain LSP Blocking

Considering the case where an inter-domain LSP path crosses k domains. We first consider the case where the blocking probabilities for each sub-LSP portion crossing a domain is determined independently (*i.e.*, the blockage probability of links belonging to a domain $k-1$ are independent of state of links in another domain k, we can write:

$$\hat{P}^{lsp}_{success}(inter) = \hat{P}^{lsp}_{success}(1) \cap \hat{P}^{lsp}_{success}(2) \cdots \cap \hat{P}^{lsp}_{success}(k) \tag{3.10}$$

where $\hat{P}^{lsp}_{success}(k)$ is the probability LSP succeeds in domain k. Assuming that the probability of successes of an LSP across the crossed domains are independent, Equation 3.10 can be rewritten as:

$$\begin{aligned} \hat{P}^{lsp}_{success}(inter) &= \prod_{\forall k} \hat{P}^{lsp}_{success}(k) \\ &= \prod_{\forall k} (1 - \hat{P}^{lsp}_{block}(k)) \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} \hat{P}^{lsp}_{block}(inter) &= 1 - \hat{P}^{lsp}_{success}(inter) \\ &= 1 - \prod_{\forall k} (1 - \hat{P}^{lsp}_{block}(k)) \end{aligned} \tag{3.12}$$

3.7 RWA in inter-domain WDM networks

In this section, we consider the problem of dynamically provisioning an end-to-end *lightpath* that spans multiple wavelength-routed optical domains. By proposing a similar aggregation technique described earlier, we present a heuristic to solve the inter-domain RWA problem. We first proceed by formulating the joint problem of routing and wavelength assignment. We then present using simulations carried out on a multi-domain network topology in order to evaluate the proposed heuristic in terms of LSP blocking probability under dynamic network state.

3.7.1 Static RWA problem formulation

Denote by F^{sd} the total traffic demand (lightpath requests) from any source s to any destination d, and by F^{sdw}_{ij} the number of lightpath requests from source s to destination d on link (i, j) and using wavelength w. Since wavelength w on any link (i, j) can only be assigned to only one path, we have $F^{sdw}_{ij} \leq 1$.

Each link in the network is assigned a cost c_{ij}. The cost of using a link can vary according to the length of the link, the amount of traffic on the link, or can be constant throughout the network. The link capacities in the network may also vary, but for simplicity, we assume all links have same capacity W. The problem can be formulated with the objective function to minimize the network cost:

$$\textbf{Minimize: } \sum_{\forall(i,j)\in E} c_{ij} F_{ij}^{sd} \tag{3.13}$$

Subject to:

- link capacity constraint:

$$\sum_{\forall(i,j)\in E} F_{ij} \leq W \tag{3.14}$$

- flow conservation constraint:

$$\sum_{\forall i} F_{ij}^{sdw} - \sum_{\forall j} F_{jk}^{sdw} = \begin{cases} -F_j^{sdw} & \text{if } s = j, \text{ lightpath head} \\ F_j^{sdw} & \text{if } t = j, \text{ lightpath tail} \end{cases} \tag{3.15}$$

- and wavelength continuity constraint:

$$F_{ij}^{sdw} = \begin{cases} 1, & \text{if path } (s,d) \text{ uses wavelength } w \text{ on link } (i,j) \\ 0, & \text{otherwise} \end{cases} \tag{3.16}$$

$$\sum_{\forall(s,t)} F_{ij}^{sdw} \leq 1 \tag{3.17}$$

where,

$$\sum_{\forall w} F^{sdw} = F^{sd}, \qquad \text{signifies the total number of lightpaths requested by (s,d) pair} \tag{3.18}$$

The static lightpath establishment problem as formulated above is NP-complete [Gare 79]. The dynamic lightpath establishment problem is even more difficult to solve.

Therefore, heuristical methods are generally employed for solving both the routing and the wavelength assignment problems. In the following section, we present heuristics using topology aggregation to solve the inter-domain dynamic RWA problem.

3.7.2 Hierarchical shortest path first heuristic

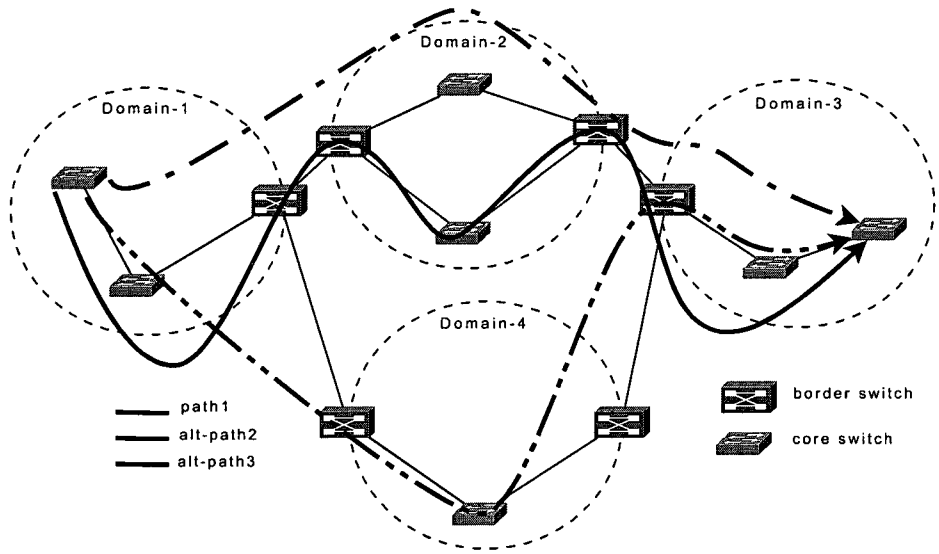


Figure 3.8: Lightpath setup across multiple domains using the HSPF heuristic

We propose a hierarchical approach to routing the end-to-end optical lightpath that we refer to as Hierarchical Shortest Path First (HSPF). The process involves calculating the entire "loose route" of the lightpath through the network. As the request crosses different domains, the call-processing entity at the ingress BN of each domain expands the loose path to compute an intra-domain path using detailed information about the domain's internal topology. This process can be repeated within each domain, with the routing entities within each domain using different algorithms to expand the sub-path portion passing through their domain. The algorithm that nodes execute to compute and signal the light path includes the below steps:

1. the destination node address is resolved to determine destination domain (*e.g.*, local *vs* global). We assume every node is identified by the tuple $(node-id, domain-id)$
2. between each pair of BNs in a domain, path discovery is performed to determine available wavelengths along the path $w^a_{bn_i,bn_j)}$.

3. an aggregate link or Forwarding Adjacency (FA) link (bn_i, bn_j) is flooded to PCEs in neighboring domains. The cost of the link computed/updated is based on w^a according to 3.20.

4. at the source-domain PCE, the shortest domain-path to is computed based on the aggregated topology which yields a set of domain hops or "loose path" (*e.g.*, $P = [d_1, d_2, ..d_m]$, where $1 < m < k$ comprising of the domain-hops to be crossed)

5. along the path to destination, each BN that receives a lightpath request for a destination (local or remote to the domain) will attempt to expand the path to nearest BN connecting to downstream domain or (destination if request reaches destination domain).

6. the lightpath is signaled from the source to BN bn_1. If successful the connection request is forwarded from bn_1 to bn_2,*etc.*.

7. at any time, if a BN cannot find a feasible path for lightpath request, the source node is notified and the request is blocked.

8. if during signaling of the inter-domain lighhtpath the request gets blocked (*e.g.*, due to unavailable wavelength), the request crancks back to the BN which can in turn attempt an alternate path that it caches– refer to Figure 3.8.

In our simulations, the link costs used are:

1. Unity fixed cost to reflect merely the hop count:

$$c_{ij} = 1, \qquad \forall (i,j) \in E \tag{3.19}$$

2. adpative cost:

$$c_{ij} = \begin{cases} -log(1 - \frac{1}{w_{ij}^a}) & \forall w_{ij}^a > 1,\ (i,j) \in E \\ 1 & \forall w_{ij}^a <= 1,\ (i,j) \in E \end{cases} \tag{3.20}$$

where w_{ij}^a is the available wavelengths on link (i,j), and the term $-log(1 - \frac{1}{w_{ij}^a})$ – or penalty function– denotes the measure of willingness a link offers to accept a call request. Dijkstra's is performed after constructing the cost matrix to determine the lightpath with the least cost. To allocate an available wavelength, we apply the First-Fit (FF) assignment algorithm [Spat 00].

Moreover, it is possible that a BN caches multiple pre-calculated link-disjoint alternatives paths (APs) to improve routing performance when signaling the lightpath along the first path fails. Multiple alternatives reduce blocking significantly when link-disjoint APs are considered as compared to the single route case [Ho 02].

We study and compare results of simulations run using our proposed schemes against that of a hypothetical flat topology where every node in the network has full network state visibility in its database. We consider the two cases where links are either assigned a fixed or adaptive cost as per Equation 3.19 and Equation 3.20.

3.7.3 Numerical results and analysis

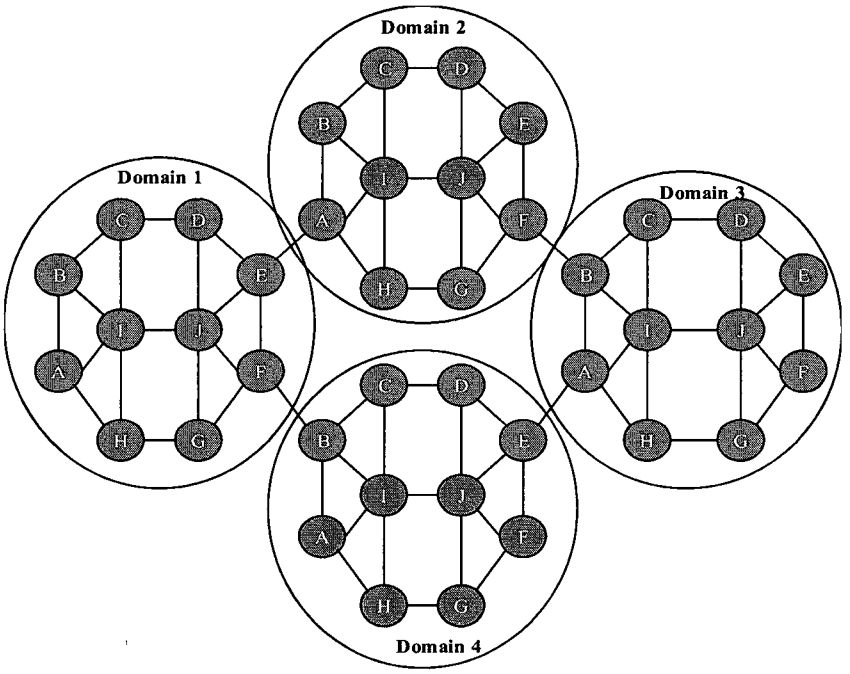


Figure 3.9: Simulated multidomain WDM network

This section presents numerical results for a 4-domain, 40-node, and 72 bi-directional link network shown in Figure 3.9. All links in the network have a 4-wavelength capacity. Calls are generated according to a Poisson process with an arrival rate. The holding times for calls are exponentially distributed with an average holding time h. The traffic load (in Erlang) is obtained by the formula $\rho = \lambda \times h$.

To study the performance of the proposed scheme, we distinguish between two types of the generated traffic: intradomain "local" calls ρ_{local}, and interdomain "global" calls as ρ_{global}. We define n as the ratio of global to local traffic:

$$n = \frac{\rho_{global}}{\rho_{local}} \tag{3.21}$$

Throughout simulations we vary $n = \{0, 0.1, 0.5, 1\}$. In each run, 10,000 lightpath call requests are generated and the blocking probability is measured. The blocking probability is computed as the ratio of blocked requests to the generated requests. For the alternate path routing approaches, we consider the first two link-disjoint paths for each source-destination pair. We vary the traffic load up to a value where considerable blocking in call requests is observed. Blocking probability is defined as the probability that connection cannot be established due to resource contention along the desired route.

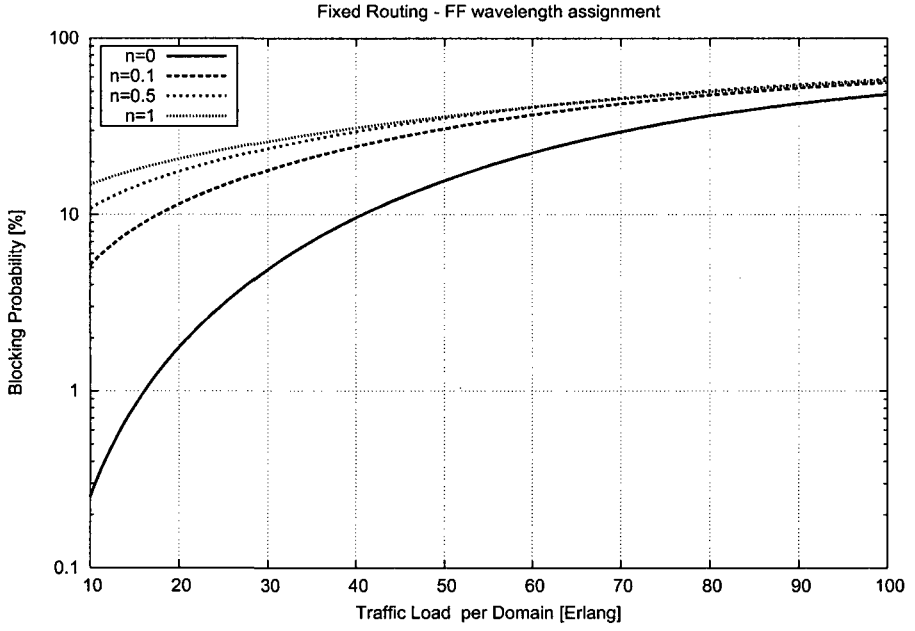


Figure 3.10: Lightpath blocking probability of multidomain network using HSPF with fixed cost

Figure 3.10 and Figure 3.11 show the blocking probability of the network when the HSPF scheme is applied under static link-costs, and with static fixed alternate routing with 1 alternate link-disjoint path respectively. In the latter case, a request is not declared

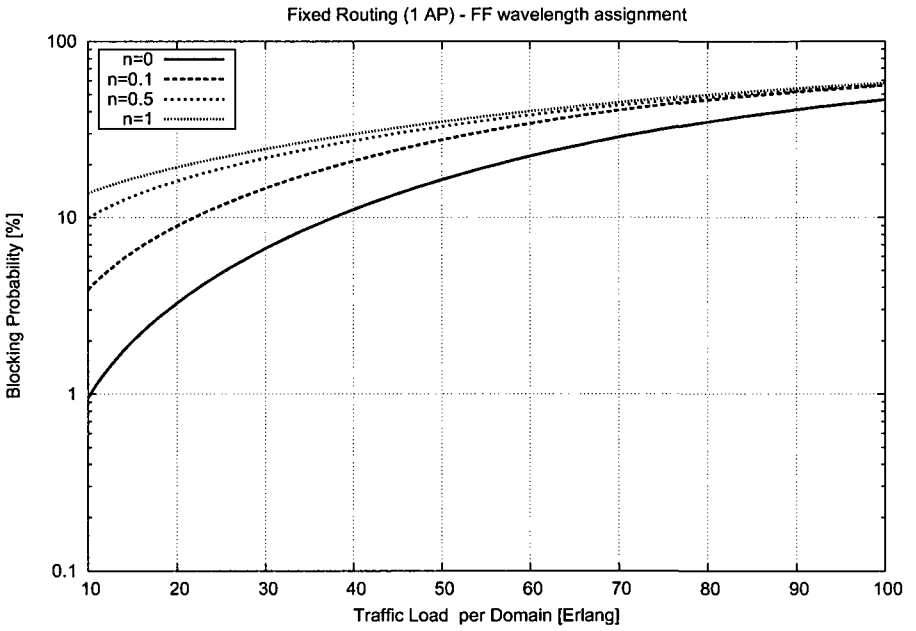


Figure 3.11: Lightpath blocking probability of multidomain network using HSPF with fixed cost and 1 alternate path

blocked if fails during signaling. Rather, another attempt to use the alternate path to signal the lightpath is performed. If after this the lightpath fails to be established the request is declared blocked. Figure 3.11 shows improvements of blocking probability when the alternate-disjoint path is considered over the single-path approach.

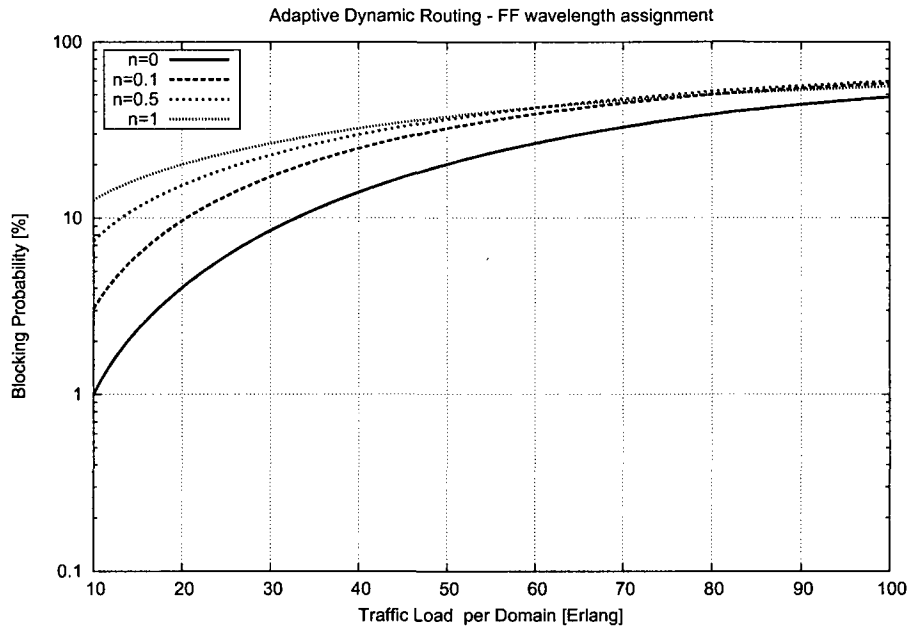


Figure 3.12: Lightpath blocking probability of multidomain network using HSPF with adaptive cost

Figure 3.12 and Figure 3.13 show the blocking probability of the network when the HSPF is applied with adaptive link-costs, and with adaptive dynamic alternate routing with 1 alternate link-disjoint path respectively. In all cases, the First-Fit wavelength assignment algorithm was applied to choose a free wavelength. Again, results show that the there is improvement in the blocking probability when choosing alternate routing over mere shortest path routing. This would be more pronounced when higher numbers of alternate routes are pre-computed. Also, the adaptive link-cost offered a lower blocking percentage since it spreads lightpaths over multiple paths over the network by dynamically reacting to the state of the network.

Figure 3.14 shows the blocking probability of the system when implemented as a flat topology. Here, every node has complete knowledge of the whole network topology and is able to compute the strict end-to-end path. As expected, results show that the performance of this scheme outperforms the hierarchical approach. However, it does suffer from scalability problems as well as from ability to compute paths across heterogeneous nodes (for example, different wavelength density across domains).

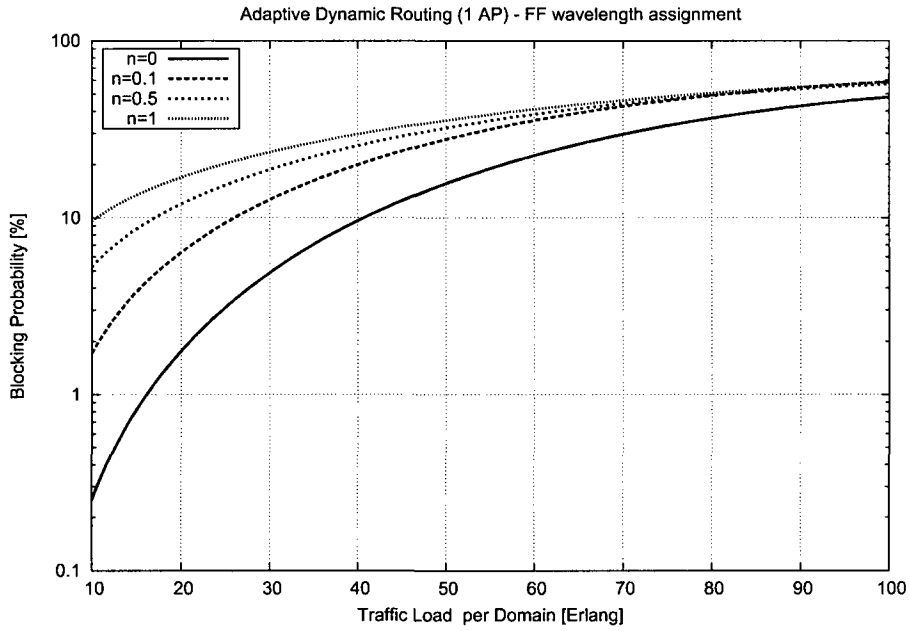


Figure 3.13: Lightpath blocking probability of multidomain network using HSPF with adaptive cost and 1 alternate path

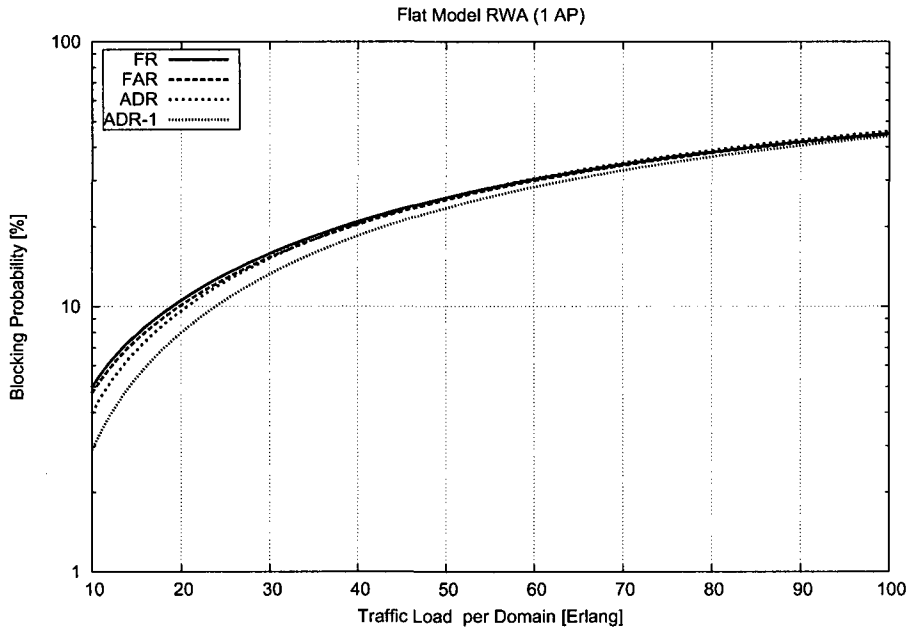


Figure 3.14: Lightpath blocking probability for flat model network

3.8 Hierarchical multi-constrained path computation

State-dependent QoS-enabled path computation schemes also necessitate the provisioning of scalable path computation solutions that take into account the QoS requirements of prospective LSP or lightpath requests as well as the available network resources. The QoS requirement of a connection is given as a set of constraints, which can be link, or path. A link constraint specifies the restriction on the use of links. For example, the wavelength continuity constraint of a lightpath connection requires that all links composing the path have the same wavelength available. A path constraint specifies the end-to-end QoS requirement on a single path– for example, a request can specify the upper limit on the end-to-end delay, or availability on the path that a optical lightpath may traverse.

In this section, we consider the problem of dynamically provisioning the end-to-end inter-domain optical lightpath that is also subject to a number of QoS constraints, namely cost and delay. We assume the user specified an upper limit on these constraints on per-request. To solve this, we propose a hierarchical lightpath provisioning algorithm for the computation and setup of the end-to-end constraint-based path. We assume requests will specify a threshold for

Using this approach, the QoS path selection process is based on a mixture of detailed and aggregated state information. The designated PCE, in each domain, can then compute a feasible QoS "loose path" using a heuristic for solve the Multi-Constraint Problem (MCP). BNs in subsequent domains compute feasible intra-domain sub-paths between BNs based on detailed information about the domain's internal topology. If at any time a BN can not expand a feasible sub-path within its domain, an error can crankback to source node to notify it of the blockage, or a new path at the BN can be attempted. To do this, we adopt a heuristic based on [Jaff 84] to simplify solving the MCP path calculation problem.

For example, in Figure 3.15 the intra-routing algorithm calculates an internal path or a set of paths between the BN pair (A, B) and assigns its metrics to the logical link in the topology aggregate. After constructing such an aggregate, a domain advertises it to all other BNs or PCEs interested in knowing this aggregated information. PCEs within each domain, will have detailed information about their own domain's state and aggregated information about other domain states.

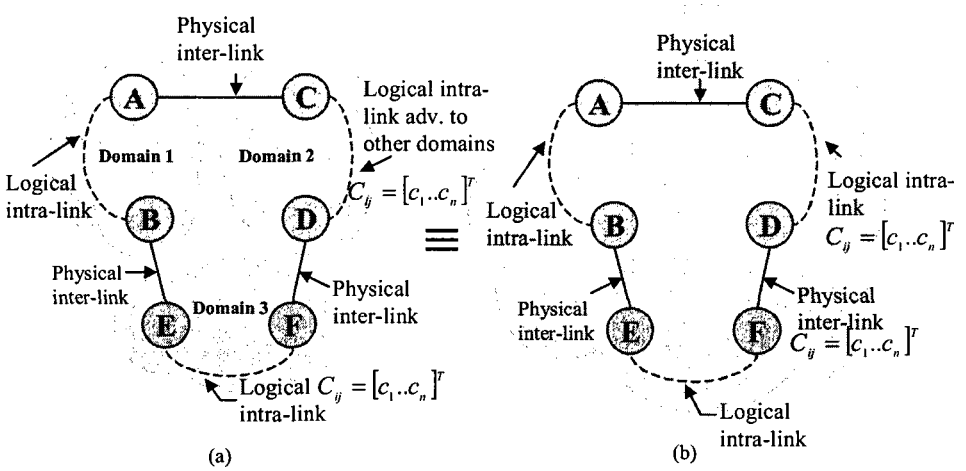


Figure 3.15: Hierarchical multi-domain network: (a) hierarchical representation, (b) transformed flat topology network

The objective of a QoS path computation scheme is to select network paths with sufficient resources that satisfy a connection's QoS request. In general, the provision of certain QoS requirements between two end-points in a network depends upon the performance properties of individual network elements such as links and nodes. The problem of computing a path subject to multiple additive constraints is an NP-complete problem that cannot be exactly solved in polynomial time [Gare 79].

3.8.1 Proposed hierarchical multi-constrained path computation scheme

In WSON networks, the performance of a network depends not only on the available physical resources, but also on how it is controlled. The objective of the RWA algorithm is to maximize the number of connection-calls serviced given certain user requirements and network resource constraints.

Assuming an multi-domain network of k inter-connected WSON domains, we represent each domain network by directional graph, $G(V, E)$ where V are set of nodes, and E set of unidirectional links. Denote by V_{bn} a subset of V representing all BNs in the domain. Assuming a full mesh topology aggregation (TA) scheme is applied in each domain, the resulting aggregated graph for each domain can be modelled as $G_{agg} = (V_{bn}, E_{agg})$ where

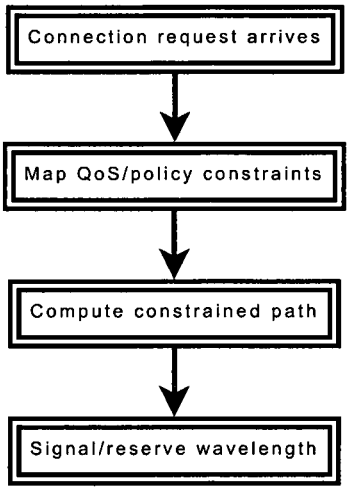


Figure 3.16: Constraint-based RWA procedures

E_{agg} is a set of aggregate links connecting all BN pairs, *i.e.*, $E_{agg} = \{(i,j) : bn_i, bn_j \in V_{bn}\}$.

BNs of a domain connect to BNs of other domains through inter-domain links. Denote by is E_{inter} the set of inter-domain links of the fully-meshed multi-domain network. The centralized PCE views the network topology as a graph of BNs from all domains connected by aggregate and inter-domain links. Denote by $E_{agg} = \cup_{\forall k} E^k_{agg}$ as the superset of all aggregate links learned by the centralized PCE from all k domains such that, $G_{cent} = (V_{cent}, E_{cent})$, where $V_{cent} = \cup_{\forall k} V^k_{bn}$, and $E_{cent} = E_{inter} \cup E_{agg}$.

Given a source node s, a destination node d, and a constraint vector $c = [c_1, c_2, \cdots, c_q]$, $q \geq 2$, we call a path p from s to d a MCP feasible path if $wt(p)_q \leq c, \forall q$. For a given QoS request and constraint vector c, QoS path computation seeks to find a feasible path p satisfying $wt(p) \leq c$ based on the current network state information– refer to Figure 3.16.

It is known that multiplicative constraints can be transformed to additive with an additional transformation step. Also, in the case of concave (or restrictive) constraints, the problem can be further simplified by first pruning out all links that do not satisfy these constraints. In this section, we mainly focus on additive metrics, namely delay and cost. Cost of the link is typically intended as an abstraction that could be mapped into a number of link metrics (*e.g.*, available wavelength or number of calls using the link). Cost of the link is typically intended as an abstraction that could, in practice, be mapped

into a number of metrics (*e.g.*, available wavelength or number of calls using the link). For our simulations, we assign an adaptive cost described in Equation 3.20 that assigns each inter- and intra-domain link in the network a cost dependent on w_{ij}^a, the number of available wavelengths on link (i, j).

In this case, the term $-log(1 - \frac{1}{w_{ij}^a})$ denotes the measure of willingness a link offers to accept a call request and usually referred to as the penalty function. The greater the number of available wavelengths, the lower is the probability that a request will be blocked. Here, the term $(1 - \frac{1}{w_{ij}^a})$ denotes the measure of willingness a link offers to accept a call request. The greater the number of available wavelengths, the lower is the probability that a request will be blocked.

We propose a linear energy function based on the convergence of multiple weights into a single metric. A similar approach was previously proposed by [Jaff 84]. By proposing energy functions, multiple QoS weights can be translated into a single metric. For example, to simplify the 2-constrained QoS routing problem, we adopt the approach by Jaffe [Jaff 84] that proposes the linear energy function $f(p) = a_1 w_1(p) + a_2 w_2(p)$, where $wt_i(p)$ is the i'th weight of path p, and a_1 and a_2 are two positive multipliers. Jaffe proposed that for a given constraint vector (c_1, c_2), when $\frac{a_2}{a_1} = \sqrt{\frac{c_1}{c_2}}$ the shortest path p computed by minimizing $wt(p)$ is feasible with maximum probability [Jaff 84].

3.8.2 Numerical results and analysis

In our simulations, we consider the cost and delay constraint path computation problem. The cost-delay constrained path computation problem is a very common requirement for many multimedia and real-time applications. We also adopt the earlier mentioned heuristic by Jaffe to solve for the multi-constraint shortest path problem. Hence, a path $p = (v_0, v_1, v_2, \cdots v_n)$ has two associated characteristics:

$$wt_c(p) = \sum_{i=0}^{n-1} c_{(i,i+1)} \tag{3.22}$$

$$wt_d(p) = \sum_{i=0}^{n-1} c_{(i,i+1)} \tag{3.23}$$

where $wt_c(p)$ and $wt_d(p)$ are the path end-to-end cost and delay respectively. In Figure 3.15, we show a sample hierarchical network of interconnected domains, and the

equivalent transformed graph of interconnected border nodes.

The blocking probability is examined under dynamic traffic for the same 40-node, 72 bi-directional link and 4-domain meshed network shown in Figure 3.9. All links in the network are assumed to have a 4-wavelength capacity. Call requests in each domain are generated according to a *Poisson* process with average arrival rate λ with average call holding times exponentially distributed with average hold time h. The traffic load per domain ρ is obtained by $\rho = \lambda h$. The source-destination pair (s, t) for each call is selected randomly with a uniform probability for local and global traffic.

We assign link delays to uniformly distributed random numbers to represent variations between lengths of fibers connecting pair of nodes in the network. For each QoS request, we randomly generate the two-constraints delay, and cost with uniform distribution. In our simulations, we also assume that call requests are equally probable to be local and global. For the allocation of an available wavelength, we apply the First-Fit wavelength assignment algorithm. In each experiment we generate 100,000 call requests and measure the blocking probability of the network. Blocking probability is defined by the ratio of number of blocked calls to the total number of calls generated.

The blocking probability is examined under dynamic traffic while varying the delay constraint for different traffic loads $\rho = \{20, 50, 100\}$. Figure 3.17 shows the blocking probability of the QoS call requests while varying the delay constraint and fixing the cost constraint. Simulations show that blocking of call requests decreases as the dominating constraint of QoS calls is relaxed (*i.e.*, delay in our case). Also, as traffic load per domain increases, more QoS call requests are likely to be blocked resulting in higher blocking probability.

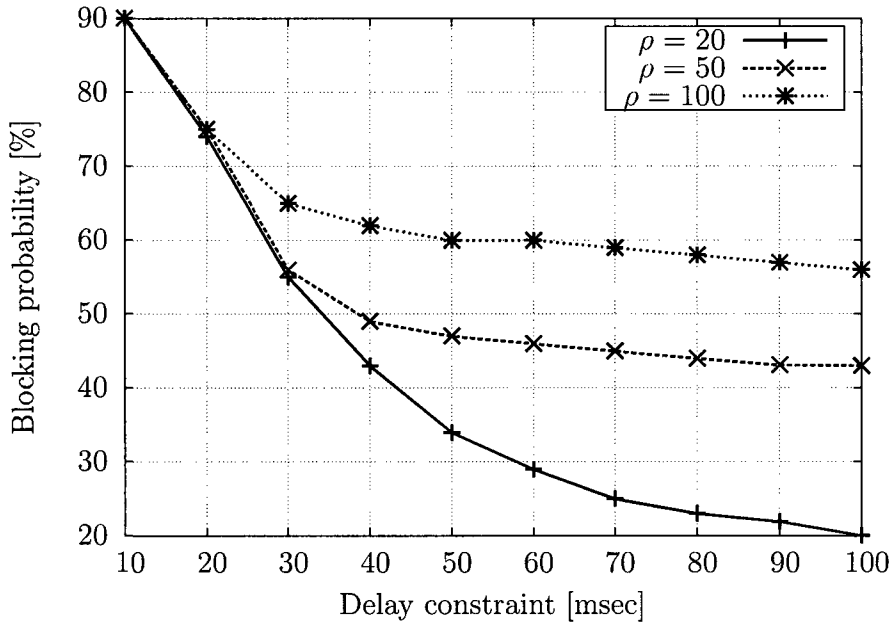


Figure 3.17: LSP Blocking probability versus delay constraint for different aggregate traffic load per domain

3.9 Conclusions

In this chapter, we proposed a novel link aggregation scheme suitable for multi-domain and multi-layered within a single carrier network, and capable of achieving control layer scalability as well as reducing the complexity of the path selection algorithms. The scheme encompasses SRLG information and utilizes the concept of trees across vertical and horizontal network-layer partitions and capable of determining SRLG diverse paths across domains and layers.

We also considered the problem of provisioning an end-to-end optical lightpath spanning multiple WSON domains. For this, we studied the case where requests specify certain QoS constraints for computing the optical paths, and proposed a heuristic based on aggregating multiple link costs to simplify the MCP problem. Using simulations run on two inter-domain topologies, we evaluated the effectiveness of the proposed scheme by comparing results for lightpath LSP blocking against a hypothetical single domain network.

Chapter 4

Availability-bounded Cooperative PCE Path Computation

4.1 Introduction

The emergence of mission-critical and other multimedia applications such as Voice over IP (VoIP), Videoconferencing, E-commerce, and Virtual Private Networks (VPNs) has translated into stringent QoS and availability requirements for carrier networks. Availability, which is the probability that a system (component, channel, connection, *etc.*) will be found in the normal operating state at a random time in the future, is a metric to measure the connection or service reliability. For enhancing the availability of services, several failure handling methods that vary in their spare resource requirements and re-routing time exist.

Large Internet Service Provider (ISP) networks have responded to those requirements by offering Service Level Agreements (SLAs) to their customers. SLAs are contracts between an ISP and its customer obligating the ISP to maintain a certain grade or level of service. ISPs are keen on offering SLAs because SLAs permit differential treatment of the customer traffic. This differential treatment can yield economic benefit to the service providers by, for example, not having to allocate a protection path to all connections. In this way the providers can make the best use of their resources given their SLA constraints. On the other hand, the customer wants SLA guarantee because they can ensure the rigid level of performance they pay for, and be compensated for the lack thereof. In addition, SLAs provide an incentive for the customer to subscribe for services

since they can choose the level that suits their need, and not have to pay premium for unnecessary features.

While many carriers, today, offer SLA contracts [ATT 09, Veri 09], these guarantees, however, are usually limited to the boundary of a single ISP only. Customers have started to express interest in similar requirements for communications with at least neighboring ISPs. Providing such inter-domain SLAs is difficult as it requires a cooperation among several ISPs. By offering end-to-end SLAs, a carrier could collectively profit from higher revenue traffic. At the same time, a customer whose sites belong to different carrier networks could benefit from the same kind of performance assurance a single-carrier customer receives. Hence, the ability to guarantee such SLAs across multiple carrier networks in a truly end-to-end manner becomes highly desirable.

MPLS is currently used by several ISPs to carry the high-value traffic for which service availability is a critical QoS dimension. Compared to pure IP, MPLS has several advantages in the inter-domain environment. First, it is possible to establish explicitly routed MPLS LSPs with QoS constraints such as availability, delay, or bandwidth. Second, several techniques exist to quickly reroute such LSPs in case of failures— *e.g.*, TE FRR. Third, LSPs with the same source and the same destination may follow different paths through the Internet. In order to preserve these QoS requirements, it is desirable to route the high-value traffic across paths that meet the availability constraints with the least monetary cost. A key factor in achieving this is the ability to perform constraint-based path computation across multiple ISP networks.

Path computation in large, multi-domain, and multi-layered networks has become more complex and demanding in terms of CPU resources. With the wide deployment of GMPLS and the variety of switching technologies, path computation constraints have increased and have become more stringent than simple bandwidth or administrative constraints. Therefore, supporting these complex computation problems— when even heuristic algorithms are computationally intensive– at every network node becomes an expensive option.

Contrary to the intra-domain case, in the inter-domain case QoS properties of the path outside a domain are not known at the ingress node of the LSP. Moreover, such information is not distributed by exterior gateway protocols (such as BGP) that run between domain BNs. This problem is especially complicated when the LSP path traverses domain boundaries of different competing carrier networks. In general, path computation can be performed either at the time of, or ahead of, service provisioning. When the

end-to-end inter-domain path is computed on a single node, such a computation is called centralized path computation. In distributed path computation there are multiple nodes that perform the path computation and may act independently or cooperate.

In this chapter, we examine methods for extending the SLA offering across ISP boundaries. These methods require each ISP to be able to offer SLAs within its boundary, but not necessarily require any knowledge about the internal structure of other ISPs. With both performance and cost considerations, we perform case studies to determine a method most beneficial to the ISP community as a whole. We propose a distributed cooperative heuristic to provision an availability-bounded least cost service for connections spanning multiple carrier networks. To achieve this, we propose extensions to the inter-domain collaborative PCE Backward Recursive Path Computation (BRPC) procedure.

4.2 Problem definition

Some inter-domain paths carry traffic that must be assured of high quality and high reliability transfer. In a multi-carrier environment there may be several alternate paths between customer end-sites. The preferred path among providers is usually determined by domain-specific policies, or by choosing the shortest AS-hop path (something also referred to as "hot-potato routing"). However, especially when multiple classes of service are offered within a carrier network, it may be desired to compute the path that satisfies the requirement with the least end-to-end cost. The optimality of a computed constrained path and how loosely that path is specified depends on the amount of network information available at the path computation node.

In the context of PCE-based mechanisms, a major issue is that the PCE in the source domain has to compute inter-domain paths based on a very limited visibility of the topology and network state, yielding solutions that are far from optimal. To cope with this, enriched topological and path state information needs to be aggregated and made available to the PCE in the source domain [Yann 06b]. Generally, there are two approaches to publishing such information about paths in remote domains, namely using

- "push" techniques that advertise summarized information about domain internals to nodes in remote domains, or
- "pull" techniques that, triggered by arrival of a request, can query a remote entities

(*e.g.*, collaborating PCEs) for an end-to-end path.

Due to scalability issues associated with the amount of summarized information exchanged between domains, "push" techniques are only suitable in multi-domain networks that fall under the same administrative entity— otherwise known as intra-carrier communication among a limited number of domains. The connection provisioning algorithms discussed in Chapter 3 rely on network-wide knowledge of aggregated topology and resource information. Carriers commonly regard such information confidential and never disclose it with other carriers. This is evident from the use of exterior gateway protocols for routing across carrier networks. The BGP protocol masks the internal topology details and only advertises reachability of network addresses and the dynamic behavior (addition and withdrawal) of such addresses to adjacent networks.

Within the PCE architecture, the distributed the Backward Recursive Path Computation (BRPC) [Vass 08a] technique was proposed by the IETF as a potential solution to the complex inter-domain TE LSP path computation problems with partial domain visibility. Several of the proposed implementations, however, make the assumption that the list of domains to be crossed— that in turn dictate the set of collaborating PCEs– is known a priori. Hence, given the set of domains or ASes to be traversed, the BRPC guarantees the computation of an optimal path that crosses the predefined AS-chain. In the case of an arbitrary set of meshed domains, the standardization does not however indicate how the input AS-chain is to be selected. In most cases, such AS-chain selection is determined based on information extracted from inter-domain BGP RIB tables which do not always warrant a feasible constrained path, nor they guarantee the selection of the optimal inter-domain path when it exists.

As we saw in the previous chapter, in general, when protection for LSPs is desirable, algorithms that guarantee connection survivability will provision pre-planned path-protection for all connections, regardless of each connection's requirement. However, this can lead to inefficient use of resources (if dedicated protection is assigned to a connection that does not require superior availability), or unnecessary demand rejection (if a connection demands superior availability, but only shared protection is offered). Therefore, to facilitate differential treatment of customerŠs traffic, connection survivability, as well as resources, should be guaranteed on a per-connection basis.

Not only should the LSP survivability be guaranteed on a per-connection basis, but the guarantee should also be measurable in a quantitative manner. The pre-defined

Table 4.1: Typical availability objectives for various IP service classes [Vogt 03]

Service classes	**Availability**	**Outage Time**
Basic Internet service	99.90%	8.76 hours/year
Australian DSL service	99.93%	6.13 hours/year
QWest wireless service	99.98%	1.75 hours/year
Broadband access	99.995%	0.438 hours/year

approach to handling LSP protection, such as letting the high-priority LSPs use optical-layer protection and the lower-priority LSPs use IP-layer protection, as described in previous chapter is sometimes insufficient. This is because customer usually demands a more concrete threshold for guaranteed services. The LSP survivability guarantee can be specified in terms of minimum connection availability, defined as the fraction of time the service is available between both ends of a connection. Table 4.1 shows typical availability objectives for different service classes reported by Telcordia [Vogt 03].

The multi-constrained shortest path problem is known to be NP-complete, and directional metrics such as diversity constraints expand the complexity. This problem is even more complex when LSPs have to cross carrier boundaries. In this case, it is crucial such information is not disclosed to neighboring competing carriers. Moreover, the inaccuracy and/or inadequacy of disseminated information can have negative commercial implications due to selection of paths crossing cost non-optimal carrier networks. The difficulty lies in the QoS inter-connection arrangement. For example, suppose two ISPs offer relative classes of service, such as gold, silver, and bronze. The gold class in one ISP may not necessarily guarantee the same parameters as the gold class in another ISP. If the service-level parameters in each traffic class are not disclosed, it would be difficult to match a class of service in one ISP with that in an adjacent one. Even if the parameter specification for each class of service is known, maintaining the constraints all the way to the end is still problematic. Coordination among all transiting ISPs would be necessary.

Moreover, the inter-connection path computation scheme should also designate how to apportion the SLA constraints to each local provisioning entity in each domain. This is not a problem for a concave constraints such as bandwidth since the threshold will remain the same for all transit networks. However, for additive constraints such as delay, or multiplicative constraints such as reliability or availability, a clear distribution policy has to be determined beforehand so an entity in each domain the way these thresholds

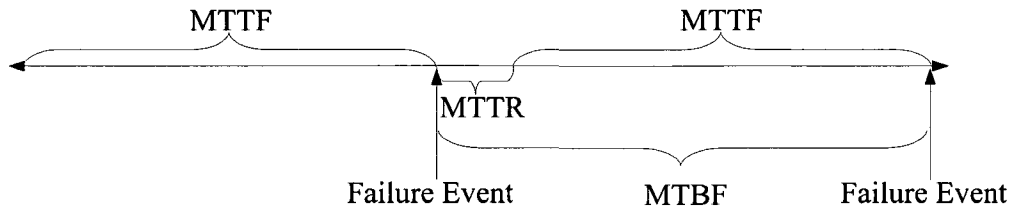


Figure 4.1: An illustration of MTTF and MTTR

are distributed among collaborating domains could greatly affect the feasibility and/or cost of the end-to-end service.

4.3 Service availability as a path constraint

Service availability between two nodes in a network refers to the ability of the network to perform its primary function (*e.g.*, packet forwarding in routers). This involves physical connectivity, link-layer protocol connectivity, and network-layer protocol connectivity. Today's ISPs offer their customers a guarantee of port availability as part of their SLAs. This simply represents the uptime of a single network element, *i.e.*, , the hardware by which the customer attaches to the ISP's network. However, this does not, in any way, provide guarantees about the destinations that a customer can reach at a given point in time. A key requirement for service availability between two nodes is the existence of connectivity between the two nodes at any of the traversed switching layers. A link failure at any of the switching layers will impact the path availability and consequently impact service availability.

It is desirable to guarantee survivability on per connection basis, since this provides a means to measure its survivability in a quantifiable manner. Availability for devices can be expressed as a function of the Mean Time Between Failure (MTBF) and Mean Time To Repair (MTTR):

$$Availability = \frac{MTBF}{(MTBF + MTTR)} \tag{4.1}$$

The time between failures in MTBF refers to the time between the occurrence of failures, whereas time to failure in MTTF is the time between the repair of a failure and the

occurrence of the next failure as shown in Figure 4.1. Therefore,

$$MTBF = MTTF + MTTR \tag{4.2}$$

The definition of service availability in a multilayered network typically considers the following factors:

- network topology,
- mapping of higher-layer links onto the underlying infrastructure,
- inter-dependence of higher-layer network elements,
- failure characteristics of links/routers, and
- routing protocol convergence time.

Network component: Availability of a network component is usually calculated based on the component's failure rate and average time to fix a failure. The steady state availability can be expressed as the function of time of the system is up over a long period of time T:

$$Availability = \lim_{T \to \infty} \{\frac{uptime}{T}\} \tag{4.3}$$

Path: The availability of path can be computed as multiplication of individual availabilities of network components along path since path p is available only when all network components— links and nodes– along its path are available.

4.3.1 Link availability across layers

In a multilayered transport network, the availability of a higher layer link (*e.g.*, IP link) depends on the underlying link layer protection (*e.g.*, optical layer). For example, assuming (1+1) protection scheme at optical layer (server layer), denote by L^f_{wp} the set of fibers that are crossed by the primary lightpath at the optical layer, and L^f_{bp} the set of fibers crossed along the backup lightpath. The IP link (i, j) will fail only when both the primary and the backup lightpaths fail simultaneously. The availability A_{ij} of link (i, j) in this case can be written as:

$$A_{ij}(1+1) = \prod_{l \in L^f_{wp}} A_l + \prod_{l \in L^f_{bp}} A_l - \prod_{l \in L^f_{wp} \cup L^f_{bp}} A_l \tag{4.4}$$

On the other hand, for shared light protection (*e.g.*, 1:N, single backup for N lightpaths), the link will fail if:

- both the primary and the backup lightpaths fail, or
- the backup does not fail but the primary and at least one other lightpath that share the same backup fail at the same time.

Let L^f_s be a set of fibers in other lightpaths that share the same backup lightpath. The availability of a 1:N protected link in this case can be written as:

$$A_{ij}(1:N) = \prod_{l \in L^f_{wp}} A_l + \prod_{l \in L^f_{bp} \cup L_s} A_l - \prod_{l \in L^f_{wp} \cup L^f_{bp} \cup L^f_s} A_l \tag{4.5}$$

Note that Equation 4.4 and Equation 4.5 already take into account reduced availability in case L^f_{wp}, L^f_{bp}, and L^f_s are non-disjoint. At the IP layer, the LSP's availability, in turn, depends on the underlying lightpath protection.

4.3.2 LSP availability

The availability of an LSP when carried over a single path (*i.e.*, unprotected) is equal to its path's availability. If the LSP, however, is path protected, the availability can be expressed as a function of A_{wp}, the primary path availability, and A_{bp}, the backup path availability since LSP fails only when both the working and backup fail simultaneously. We can write:

$$\begin{aligned} A_{lsp}(pp) &= 1 - (1 - A_{wp})(1 - A_{bp}) && (4.6) \\ &= A_{wp} + (1 - A_{wp})A_{bp} && (4.7) \end{aligned}$$

Multilayered LSP availability. Given an LSP subsets L_{unprot} of unprotected links, L_{1+1} of $1+1$ protected links, and $L_{1:N}$ of $1:N$ protected links. The availability of that LSP can be written as:

$$A_{lsp} = \prod_{l \in L_{unprot}} A_l \bullet \prod_{l \in L_{1+1}} A_l \bullet \prod_{l \in L_{1:N}} A_l \tag{4.8}$$

Shared protected LSP availability. Assuming the LSP is using shared-path protection along a path wp, denote by Ω_{wp} the set that contains all primary paths (except wp) whose backup paths are sharing some resources with bp. Let $|\Omega_{wp}| = \sigma$, and $\hat{P}^i$ be the probability that exactly i primary paths in Ω_{wp} fail. Let δ_i be the probability that wp can get backup resource when both wp and other i primary paths in Ω_{wp} fail.

The availability of the LSP can be computed by Equation 4.9. In this case, the LSP is available if wp is available, or if wp is unavailable and bp is available, and the failure on wp happens before failures to other primary paths in Ω_{wp} sharing the backup resource bp. We assume wp and all other primary paths in Ω_{wp} fail independently. Hence, $\delta_i = 1/(i+1)$. Note, it is possible that the two primary paths in Ω_{wp} traverse the same fiber link so their path failures may not be independent. In this case, $1/(i+1) \leq \delta_i \leq 1/2$ when $i \geq 1$. $\delta_i = 1/(i+1)$ will give the lower bound of the LSP availability in this case.

$$A_{lsp}(shared) = A_{wp} + \sum_{i=0}^{\sigma} \delta_i (1 - A_{wp}) A_{bp} \hat{P}^i \tag{4.9}$$

Inter-domain LSP availability. The availability of an inter-domain LSP that traverses multiple domains offering different protection types for each portion of the LSP crossing the domain:

$$A_{lsp}(inter) = \prod_{d=1}^{k} A_{lsp}^d - \prod_{d=1}^{k} A_{lsp}^d \tag{4.10}$$

Given a path p that traverses k domain-hops $D_1, D_2, \cdots, D_k$ to the tail-end, we say p is a reliable path for constrained LSP if and only if it satisfies the inequality:

$$A_p = \prod_{d=1}^{k} (a_d) \geq \Delta \tag{4.11}$$

where a_k is the availability of the sub-path in domain k, and Δ is the requested availability. Computing the logarithm and multiplying both sides by -1, while noting that service availabilities are between 0 and 1, we get:

$$-log(A_p) = -log(a_1) - log(a_2) - \ldots - log(a_k) \leq -log(\Delta) \tag{4.12}$$

If the cost of link (i, j) is defined as a function of its availability (*e.g.*, $c_{ij} = -log(a)$), the cost is additive and the path with minimum cost will be the path with maximum availability (such a path is called the most reliable path (MRP)). This multiplication-to-summation technique can then be used to compute the MRP. If the availability of a MRP is lower than Δ, then this path is not reliable enough. Note here that the availability of a link each, and consequently the availability of domain sub-segment paths in each domain can be increased adding more protection as required as described in Equation 4.11.

4.4 TE constraint problem formulation

The Traffic Engineering problem entails minimizing the overall path cost (*e.g.*, IGP, TE, and/or aggregate transmission delay) over all data flows, each of which is subject to certain constraints such as link bandwidth capacity. The TE problem is similar to the restricted-shortest path problem whose goal is to find a shortest path that does not violate a resource constraint. In label switched networks, TE LSPs can be established along paths that could be longer than the shortest hop ones— for example, in order to not violate the constraint– potentially incurring higher costs (*e.g.*, higher end-to-end transmission delays). Thus the objective, in this case, is to minimize the total cost whereby shorter paths are always favored as long as they do not violate the bandwidth constraints.

In MPLS terminology, LSPs are set-up between an ingress-egress pair of LSRs. Each connection request arrives at an ingress router which determines the explicit path for the LSP according to the current TE topology and to the available capacities at the IP

layer. It is assumed that every LSR runs a link state routing protocol with extensions for flooding link residual bandwidth advertisements.

Consider a single domain IP/MPLS network which consists of LSR nodes and TE links connecting them. The network topology can be modeled with a directed graph $G = (V, E)$, where V denotes the set of LSR nodes and E the set of TE links, and $|V| = n$ and $|E| = m$. Denote by (i, j) a link from node i to node j.

Some of the information pertinent to the network topology and links states are characterized by:

W_{ij}: the capacity of the TE link (i, j) – *e.g.*, the total reservable bandwidth

c_{ij}: the cost for transmission on TE link (i, j)– *e.g.*, the communication delay, TE, IGP, or adaptive cost

β_{ij}: the maximum allocation multiplier or over-subscription factor on TE link (i, j)

The parameters describing the LSP data flows are:

O: set of all LSPs, where s^o and d^o are the ingress and egress nodes of LSP o.

b^o: bandwidth requested by LSP o.

hp^o: maximum path hops for a LSP o

The TE constraint-based routing (CBR) problem is characterized by a set of demands (LSPs)– each described by a specific set of attributes– that are to be routed through the network. The objective is to select the optimal placement of LSPs through the network while adhering to the constraints imposed. The unknown variables that need to be determined based on optimizing a certain objective function and satisfying a set of constraints are as follows:

The traffic allocated by demand o on link (i,j) is denoted by x_{ij}^o:

$$x_{ij}^o = \begin{cases} 1, & \text{if LSP } o \in O \text{ is routed over link } (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{4.13}$$

The induced load ι_{ij} on a link (i,j) is:

$$\iota_{ij} = \sum_{o \in O} x_{ij}^o, \quad \text{for any } (i,j) \in E. \tag{4.14}$$

Resource based optimization would lead to an objective function that minimizes the sum over all links of the product of the cost and the total flow on each link. The link cost can typically be assumed any of the applicable additive link attributes (*e.g.*, delay, IGP cost, or unity for hop-count *etc.*). It is also possible that the link cost be defined as a function of resource availability. In this case, the link cost function can serve as a penalty function that is monotonically increasing and convex over $[0, W_{ij}]$ with $\lim_{x \to W_{ij}} c_{ij} = +\infty$. Hence, the objective can be written as:

$$Minimize \quad Z = \sum_{\forall (i,j) \in E} c_{ij}(\iota_{ij}) \tag{4.15}$$

Or,

$$Minimize \quad Z = \sum_{\forall (i,j) \in E} \sum_{\forall o \in O} c_{ij} b^o x_{ij}^o \tag{4.16}$$

The basic set of constraints are:

$$\sum_{o \in O} b^o x_{ij}^o \leq W_{ij} \beta_{ij} \tag{4.17}$$

$$x_{ij}^o \geq 0, \qquad \forall \text{ link } (i,j) \in E \text{ and LSP } o \in O \tag{4.18}$$

$$\sum_{(i,j) \in E} x_{ij}^o \leq hp^o, \qquad \text{for all } o \in O \tag{4.19}$$

$$\sum_{\forall l|i=s^o} x_{il}^o - \sum_{\forall l|i=d^o} x_{il}^o = b_i^o, \qquad \text{for all } i \in V \tag{4.20}$$

where:

$$e_i^o = \begin{cases} 1, & \text{if } i = s^o, \text{ ingress LSR for } o \\ -1, & \text{if } i = d^o, \text{ egress LSR for } o \\ 0, & \text{otherwise} \end{cases}$$

Variable x_{ij}^o takes on value 1 if, and only if, the TE path of LSP o uses TE link (i, j).

From the above set of equations, Equation 4.17 imposes bounds on the data traffic on each link. Equation 4.18 guarantees that the nodes selected for each LSP o form a simple path from its source s^o to its destination d^o. Equation 4.19 imposes a bound on the number of hops for an acceptable path for LSP o. Equation 4.20 is the flow conservation constraint that states that traffic for each ingress-egress pair incoming to a node has to be equal to the outgoing traffic from that node.

The TE problem has been shown to be an NP-complete problem. However, service providers still require efficient, scalable and "not necessarily" an optimal solution for this problem. It is desirable, therefore, that algorithms converge quickly to a feasible solution which has bounded deviation from optimality. Various algorithms may be compared and refined based on empirical and simulation studies and experience. It is often necessary to have online TE tools to solve, in real-time, problems such as the connection admission, CBR and rerouting. In this case, efficient methods to interface with the network routing protocols and network management system are necessary. In other situation, for example network capacity planning and reoptimization, off-line tools are useful for solving non-real-time problems. Direct interfacing to routing protocols and network management system are optional in this case.

A feasible path is one that has sufficient residual (unused) resources to satisfy the QoS constraints of a connection. The basic function of QoS routing is to find such a feasible path. In addition, most QoS routing algorithms consider the optimization of resource utilization, measured by an abstract metric cost. As such, the cost of a link can be defined in dollars or as a function of the buffer or bandwidth utilization. The cost of a path is the total cost of all links on the path. The optimization problem is to find the least-cost path among all feasible paths.

4.4.1 Availability bounded TE problem formulation

In this section, we assume each link $(i,j) \in E$ is characterized by an availability a_{ij} that describes the degree of the protection that the link acquires, and a cost c_{ij} that is a monotonically increasing function (*e.g.*, based on a monetary cost or some measure of the link's capacity).

Given an LSP o, and a positive availability constraint Δ. The *Least Cost Availability Bounded* (LCAB) path computation problem can be to find the least cost path p whose availability does not violate the user-set threshold, *i.e.*, :

$$\min wt(p) : p \in P(s,d) \text{ and } a(p) \geq \Delta \tag{4.21}$$

where $P(s,d)$ is the set of paths from the source node s to the destination node d. Denote by $P'(s,d)$ the sub-set of paths in $P(s,d)$ such that their end-to-end path availability is bounded by availability Δ. Recall Equations 4.11 and 4.12 a path $p \in P(s,t)$ is in $P'(s,t)$ if and only if:

$$\sum_{\forall\ (i,j)\in p} -log(a_{ij}) \leq -log(\Delta) \tag{4.22}$$

Denote by U the set of all ingress-egress pairs where one or more LSPs originate or terminate, and O the set of LSP, and Θ set of all offered protection service classes, the objective function is to minimize the cost of the LSPs paths in O:

$$Minimize \sum_{o\in O}\sum_{\theta\in\Theta}\sum_{(i,j)\in E} b^o x_{ij}^o c_{ij}^o \psi_{ij}^o c_{ij}^\theta, \qquad \forall p \in P(s^o, d^o) \tag{4.23}$$

where,

$$x_{ij}^o = \begin{cases} 1, & \text{if LSP } o \text{ uses link } (i,j) \\ 0, & \text{otherwise} \end{cases} \tag{4.24}$$

$$\psi_{ij}^o = \begin{cases} 1, & \text{if LSP } o \text{ uses protection service class } \theta \in \Theta \text{ on link } (i,j) \\ 0, & \text{otherwise} \end{cases} \tag{4.25}$$

subject to the constraints:

$$\sum_{\forall o \in O} x_{ij}^{o} < W_{ij}, \forall (i,j) \in E \tag{4.26}$$

$$\sum_{\forall\ (i,j) \in E} -log(x_{ij}^{o} a_{ij}^{o}) < -log(\Delta), \forall o \in O \tag{4.27}$$

The LCAB problem is among a set of problems of finding a feasible path with two or more independent path constraints which has been found to be NP-hard [Gare 79] – that is finding the theoretical optimum solution for the problem can not deterministically be found in polynomial time.

There has been several heuristics proposed in the literature to solve similar problems. In Chapter 2 we present a survey of techniques proposed that tackle this specific problem. In this chapter, we base our inter-domain path-computation heuristic on the Lagrange Relaxation Aggregated Cost proposed (LARAC) algorithm proposed in [Jutt 01] which is a heuristic based on minimizing modified mixed cost function $c_\alpha := c + \alpha \dot{a}$.

4.4.2 Inter-domain problem formulation.

An inter-domain connection request is characterized by a source AS-node, a destination AS-node, certain requsted bandwidth b and an end-to-end constraint or bound (*e.g.*, availability). Let $G(V, E)$ be a graph where V is the set of AS-nodes and E the set of inter-AS logical connections.

We distinguish between four types of traffic:

Intra-domain traffic. Traffic that is generated at and destined to an LSR within a domain, say I.

Outgoing inter-domain traffic. Traffic that is generated in a domain I, and destined to an LSR in a domain outside I.

Transit inter-domain traffic. Traffic that is generated from and destined to an outside domain. It uses the domain as a passage. This type of traffic can be similar to

outgoing inter-domain traffic but generated at the incoming BN LSR. The outgoing requirements for the BN LSRs are modified to incorporate this amount of traffic incoming from their incoming inter-domain links.

Incoming inter-domain traffic. Traffic that is generated in a source LSR outside of the domain and destined to an LSR that is within the domain.

As discussed earlier, the inter-domain path computation involves computing path subsegments along the selected inter-domain path. We assume an inter-domain path can be computed using one of the earlier presented hierarchical abstraction models in Chapter 2. Assuming a complex nodes abstraction method, we start by modelling the global network of domains by a logical network (directed graph) constituted by nodes representing domains, and logical links representing inter-domain links.

The next step would be to apply the previously mentioned intra-domain TE techniques to determine an inter-domain path computation scheme defined in Section to compute the domain path.

Once the inter-domain LSP path is determined, the inter-domain LSPs requested resources are used as constraints for the intra-domain TE problem. For this reason, we assume that the domain being considered, domain k consists of its set of V nodes and set E intra-domain links. The topology of domain k can be modelled as described earlier by graph $G_k(V, E)$. Furthermore, we assume each domain k knows the following:

- the intra-domain traffic demand between nodes i and j, $\mathcal{F}_{ij}$
- the outgoing inter-domain traffic demand, $\mathcal{F}_{iJ}$, where $I \neq J$
- the total incoming traffic demand destined to an egress node in domain I, F_{iI}
- $\mathcal{F}_{gJ}$, traffic demand from BN g entering domain I and destined to domain J Denote by $g \rightarrow i$ and $g \leftarrow i$ inter-domain links g *in* and *out* of border node i, respectively.

The intra-domain TE problem can be formulated as shown in section. The objective for the intra-domain TE problem can be as stated in Section 4.4.

The total inter-domain load $\mathcal{T}_{iJ}$ that a node i in domain I sends to domain J (including both locally and externally generated/relayed) appears like local traffic generated in i is computed as:

$$\mathcal{T}_i^J = \begin{cases} \mathcal{F}_{iJ}, & \forall i \ : \ i \text{ internal LSR, and } J \neq I \\ \mathcal{F}_{iJ} + \sum_{g \rightarrow i} f_{gJ}, & \forall i \ : \ i \text{ is BN LSR, and } J \neq I \\ \sum_{g \rightarrow i} \mathcal{F}_{gJ}, & \forall i \ : \ i \text{ is BN LSR, and } J = I \\ 0, & \forall i \ : \ i \text{ internal LSR, and } J = I \end{cases} \tag{4.28}$$

The total traffic $\mathcal{R}_i^J$ that an LSR node i in domain I receives from another domain and relays to domain J appears like local traffic destined to i. It is computed as:

$$\mathcal{R}_i^J = \begin{cases} \mathcal{R}_i^I, & \forall i \text{ and } J = I \\ 0, & \forall i \ : \ i \text{ internal LSR and } J \neq I \\ \sum_{g \leftarrow i} \mathcal{F}_{gJ}, & \forall i \ : \ i \text{ BN LSR and } J \neq I \end{cases} \tag{4.29}$$

Define:

$P(i,j)$: the set of paths between LSR routers i and j

π_{ij}^p: fraction of the internal traffic demand between LSR routers i and j on path p

$\phi_{ij}^{p(J)}$: fraction of the inter-domain traffic between LSR routers i and j on path p destined to domain J

$\mathcal{F}_{ij}^J$: traffic rate from LSR i to LSR j and LSR j relays to domain J

The global traffic on each link can be determined as:

$$tf_{ij} = \sum_{\forall(i,j)} \mathcal{F}_{ij} \sum_{p:l \in p} \pi_{ij}^p + \sum_J \sum_{(\forall(i,j)} \mathcal{F}_{ij}^J \sum_{p:l \in p} \phi_{ij}^{p(J)} \tag{4.30}$$

The linear constraints are: no double booking, that is for same LSP, do not reserve more than single unit of BW at any link along path p.

$$\sum_{\forall p \in P_{ij}} \pi_{ij}^{p} = 1, \qquad \forall (i,j) \tag{4.31}$$

$$\sum_{\forall p \in P_{ij}} \phi_{ij}^{p(J)} = 1, \qquad \forall (i,j),\, J \tag{4.32}$$

and

$$\sum_{j} \mathcal{F}_{ij} = \mathcal{T}_{i}^{J}, \qquad \forall i,\, J \tag{4.33}$$

$$\sum_{i} \mathcal{F}_{ij} = \mathcal{R}_{j}^{J}, \qquad \forall j,\, J \tag{4.34}$$

4.5 Proposed PCE cooperative heuristic

In Chapter 2, we gave a detailed overview of the major approaches to path computation for inter-domain LSPs, namely, the IP shortest path, the per-domain path computation, and the PCE centralized and cooperative based path computation.

There are several motivations for the deployment of the PCE-based architecture, among those are off-loading the highly CPU-intensive path computation functions from the control plane of core routers to some other specialized network nodes, *e.g.*, PCE(s). Cooperative path computation is one of several possible ways that PCEs in different domains can collaborate to compute the end-to-end path. For example, consider an end-to-end path computation problem for a LSP connection, and assume that the detailed path information from the ingress of a certain domain ("current domain") is not known a priori. We say a PCE in the current domain and a group of PCEs in other domains have cooperated or collaborated when they exchange path computation information that would determine the path beyond the ingress to the next domain.

The BRPC is a is a multiple-PCE cooperative path computation technique that assumes PCEs in each of the traversed domain recursively collaborate— starting from the destination domain– in order to compute the constrained TE LSP path. The BRPC, however, makes the assumption that the list of domains to be crossed by the LSP is known a priori. In this case, the computed path is merely the "shortest" path that can be obtained

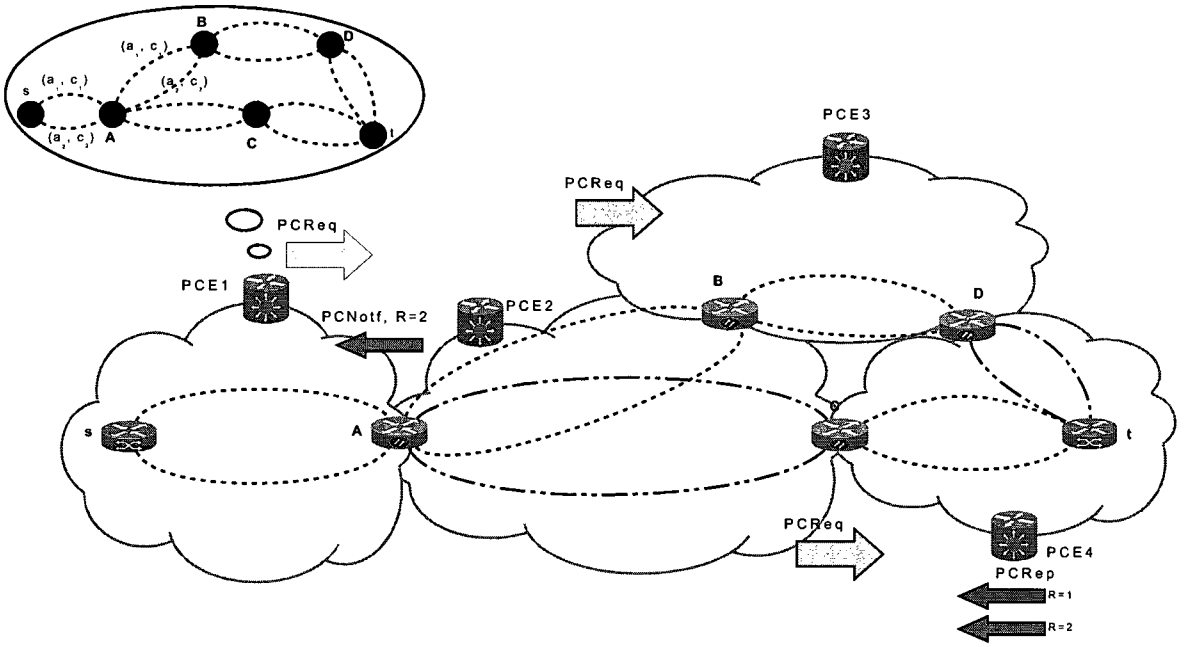


Figure 4.2: Extended-BRPC example

along this specified inter-domain path. Moreover, when the optimization of two or more constraints is sought, the BRPC does not always yield a feasible path. Moreover, in practice, the assumption that the AS or domain path is known a priori is not always true, especially in a heavily meshed domains were multiple domain paths exist to the destination domain.

Moreover, given that more and more providers are implementing Differentiated Services (DS)-TE within their networks, it is desirable that the correct TE class is identified and utilized within each of the transitted domains to achieve the least cost availability-bounded end-to-end path. ISPs today will typically associate a certain price c_v with using a class-type v.

For example, assuming each domain defines Θ classes within its boundaries, where *class-type 0* resembles the protection service with the highest service-availability (and naturally the highest price per kbps), and *class-type* θ the lowest service-availability and price-per kbps.

As described earlier, when considering the availability of a link as a cost, an SPF within a domain yields the MRP within the domain. The BRPC algorithm, in this case, yields the MRP sub-path in each domain along the inter-domain path, and as a result the

MRP inter-domain path. However, while this yields a feasible path when it exists, the computed path needs not necessarily be the lowest price feasible one as we shall show. Even, when several candidate PCE(s) along the multiple available domain paths are consulted, this still returns the MRP path among the available ones with no regards to the different services advertised in each domain along the inter-domain path.

Algorithm 4.1 LCAB heuristic executed by SPCE

LCAB(s, t, c, Δ)

$p_c := Dijkstra(s, t, c)$
if $(a(p_c) \geq \Delta)$ **then**
  return p_c
end if
$p_a := Dijkstra(s, t, a')$, where p_a is the MRP, and $a' = -log(a)$
if $(a(p_a) < \Delta)$ **then**
  return "There is no solution"
end if
loop
  $\lambda := \frac{c(p_c)-c(p_a)}{a(p_a)-a(p_c)}$
  $temp := Dijkstra(s, t, c_\lambda)$
  if $c_\lambda(r) = c_\lambda(p_c)$ **then**
    return p_a
  else if $(a(r) \leq \Delta)$ **then**
    $p_a := temp$
  else
    $p_c := temp$
  end if
end loop
where $a(p)$ is the availability of path p, $c(p)$ is the cost of path p, and $Dijkstra(s, t, \cdot)$ returns $\cdot$-minimal path between the nodes s and t.

To solve such problems, it is typical to define heuristics that find close to optimal solution. In this section, we describe a novel distributed inter-domain multi-constraint path computation heuristic that we refer to as Extended BRPC (X-BRPC) that takes into consideration the feasibility of the path due to a certain constraint (*e.g.*, availability) as well as the total service cost (*i.e.*, the TE classes advertised by each domain) in the computation of the inter-domain path.

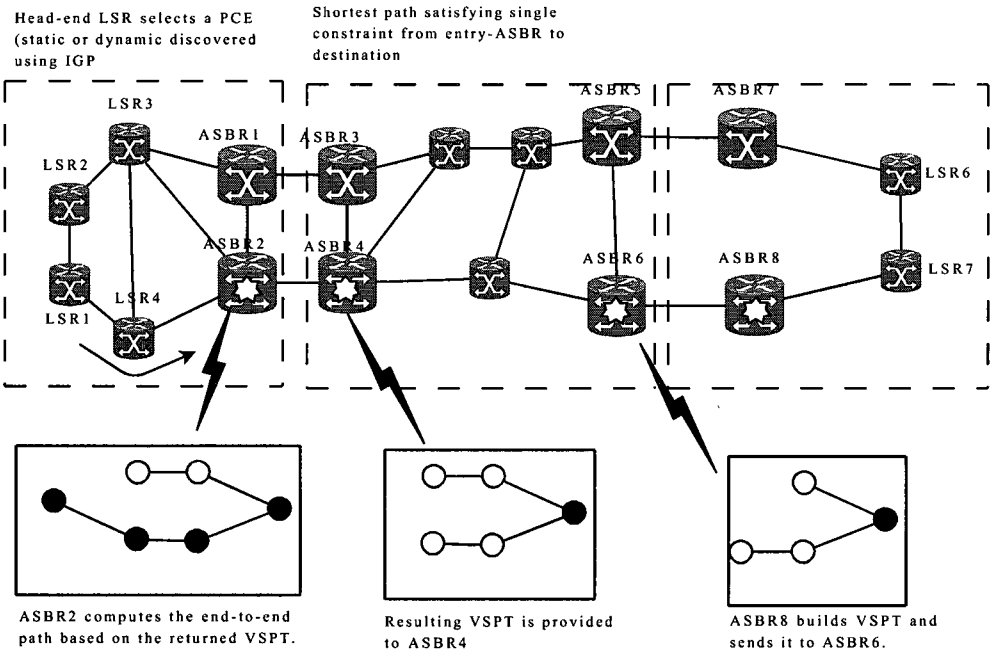


Figure 4.3: BRPC PCE distributed computation model

Using our proposed scheme, the PCC (headend LSR) sends a PCReq for a constraint path computation of an inter-domain LSP to a PCE in its domain. The source domain PCE (SPCE) then determines one or more PCReqs to candidate downstream PCE(s)that are capable of further computing the inter-domain path. When multiple downstream PCE(s) are consulted, we propose to maintain a counter within each PCReq that gets incremented when multiple downstream PCE(s) are contacted in-parallel. The request(s) are forwarded between PCEs, domain-by-domain, until the PCE responsible for the domain containing the LSP destination is reached. The PCE in the destination domain creates a tree of paths to the destination– the Virtual Shortest Path Tree (VSPT)– and passes this back to the previous PCE in a PCReq. As opposed to the BRPC, the extended VSPT is multi-graph produced by having one or more paths from each entry-BN to exit-BNs of a transit domain, or to the destination in the case of the destination domain. Each virtual link in the extended VSPT, in this case, represents a specific service offered by the domain with a certain availability and cost. In this case, $VSPT(k)$ of upstream domain is appended to $VSPT(k+1)$ by adding service-paths from each entry-BNs to exit-BNs of domain k.

The SPCE keeps track of the received/expected number of PCReqs. Once all PCReqs

are received, or depending on a certain threshold, the SPCE runs the LCAB heuristic to determine a feasible path. As an optimization, the SPCE can prune links that do not satisfy the requested availability and bandwidth before running the LCAB algorithm. The path, if found, is returned to the ingress LSR so it can initiate the signaling of the LSP.

Figure 4.2 illustrates the use of the X-BRPC in the computation of a constrained LSP path by means of cooperative PCEs. An LSP with minimum availability has to be established between s and t. There are two domain paths to reach the destination domain D4, namely $D1, D2, D4$ and $D1, D2, D3, D4$. After querying for constraint paths along the 2 domain-paths, the SPCE, PCE1, composes the virtual topology and runs the multi-constraint heuristic to find a feasible path.

4.5.1 Comparison of available techniques

Table 4.2 shows the comparison of the messaging complexity that the proposed X-BRPC scheme generates against previously mentioned schemes. The complexity of signaling of X-BRPC— due to exchange of PCE PCReq/PCReq PCEP messaging– that the X-BRPC entails is $O(kK)$ where k is the number traversed downstream domains and K is the maximum number of contacted downstream PCE(s) at any PCE.

At any of the traversed domains along the path computation domain-path, assuming BN_{en} domain entry-BNs and BN_{ex} domain exit-BNs, the cardinality of the extended-VSPT that the X-BRPC entails that the SPCE has the use to compute a feasible path is $O(kBN_{en}BN_{ex})$ as opposed to $O(BN_{ex})$ for the traditional BRPC.

To improve the efficiency of the used resources due to PCEP control messaging, and size of the exchanged VSPT, one way would be to cache the polled topology at the SPCE and to only trigger new polling requests when a threshold is crossed. This threshold can be either timer-driven (*e.g.*, expiry of a stale information timer), or triggered based by no feasible path found when using the cached virtual topology.

4.6 Numerical results and analysis

In this section, we evaluate the PCE-based path computation heuristic that was presented earlier in terms of path computation for inter-domain LSPs with a maximum end-to-end

Table 4.2: Comparison of inter-domain path computation techniques

Technique	**PCEP per-computation signaling**	**Virtual topology size**	**Number of constraints**
Per-domain ERO	-	$O(E)$	2 or more per-domain, and hop-count globally
Centralized PCE	k^2	$O(kE)$	2 or more globally
BRPC	k	$O(kBN_{ex})$	1 (IGP or TE cost)
X-BRPC	kK	$O(kBN_{en}BN_{ex})$	2 or more globally

availability constraint. The objective is to maximize the number of established TE LSPs with certain availability/SLA requirement while minimizing the total network resource usage and overall service cost.

We present results of simulations runs that are carried out using the discrete-event network simulator that we developed to replicate closely the behavior and timing of the overall inter-domain TE LSP path computation and signaling processes. The simulator is described in more details in Appendix A.

To obtain meaningful results applicable to realistic provider topologies, simulations are carried out on a multi-domain network with multiple BN LSRs sitting at the edges of adjacent domains. Figure 5.6 shows the topology used for running the simulations. It consists of four fully-meshed domains with a total of 26 TE-LSRs, and 48 bi-directional TE-links. In this topology, domain-2 connects to all other domains and acts as a backbone domain for the other connected domains. TE Links are all assumed to be of the same capacity of OC-192 (9.6 Gbps). The propagation delays on links is assumed proportional to distance of the laid fibers between neighboring LSRs, and is generated from a randomly with normal distribution to reflect variations in distance of laid fibers between each router pairs. TE link availabilities are assumed to be evenly distributed over (0.9, 0.99, 0.999, 0.9999, 0.99999). The availability requirements for each LSP requests are assumed to be uniformly distributed over (0.96, 0.97, 0.98, 0.99, 0.999). In our simulations, we assign each TE link a dynamic cost function that integrates the resource state usage as well as the availability on each link as:

$$c_{ij} = \begin{cases} -log(1 - \frac{1}{m_{ij}} \times a_{ij}) & \text{where } m_{ij} \text{ is the residual available bandwidth at link } (i,j) \\ \infty & \text{if } m_{ij} = 0 \end{cases} \tag{4.35}$$

The inter-domain TE-LSP requests are generated according to a *Poisson* process with a mean aggregate arrival rate of requests λ_T. For each request, the headend and tailend LSRs s and t are equally likely chosen from the source/destination domain's set of TE LSRs. Holding time of each LSP request follows a negative exponential distribution with a mean holding time of $h = 10$ seconds. The bandwidth demanded by each of the TE-LSPs are assumed to be evenly distributed over $[1/10 - 5/10]$ of the total link capacities. Any blocked request at any stage in signaling is dropped and no retrying is considered (*i.e.*, no crankbacks).

All LSRs in the simulated topology are assumed TE capable and able of processing RSVP-TE extensions. Intra-domain LSRs are all assumed capable of either signaling per-domain loosed-path RSVP-TE LSPs, or acting as PCCs to generate PCE path computation requests for inter-domain LSPs. BNs are all assumed capable of handling either: per-domain TE loose-path expansion, or performing end-to-end PCE path computation.

To evaluate the effectiveness of the proposed schemes, we use as performance metrics as the average LSP signaling time, and total LSP request blocking probability, the average LSP path hop length. TE LSP requests are blocked due to either: 1) no feasible TE path is found that satisfies the LSP requested bandwidth and/or end-to-end availability constraint (*i.e.*, blocked at path computation time prior to signaling the LSP), or 2) the resources on any of the traversed links/domains become unavailable (*e.g.*, get reserved by other competing LSP requests) after the LSP signaling has initiated. For each simulation run, 100,000 requests are generated and processed and the results for performance metrics are collected.

We compare the performance of the proposed heuristic with two other techniques that were discussed in details in Chapter 2, namely, the per-domain loose-path expansion and the traditional BRPC PCE heuristic. In this case, the traditional BRPC PCE heuristic is equivalent to the proposed heuristic when only one domain-hop path is explored (*i.e.*, $K = 1$). As well, we present results for simulations run over a hypothetical flat network topology replicating the same multi-domain topology.

Figure 4.4 compares the average inter-domain LSP request blocking probability prob-

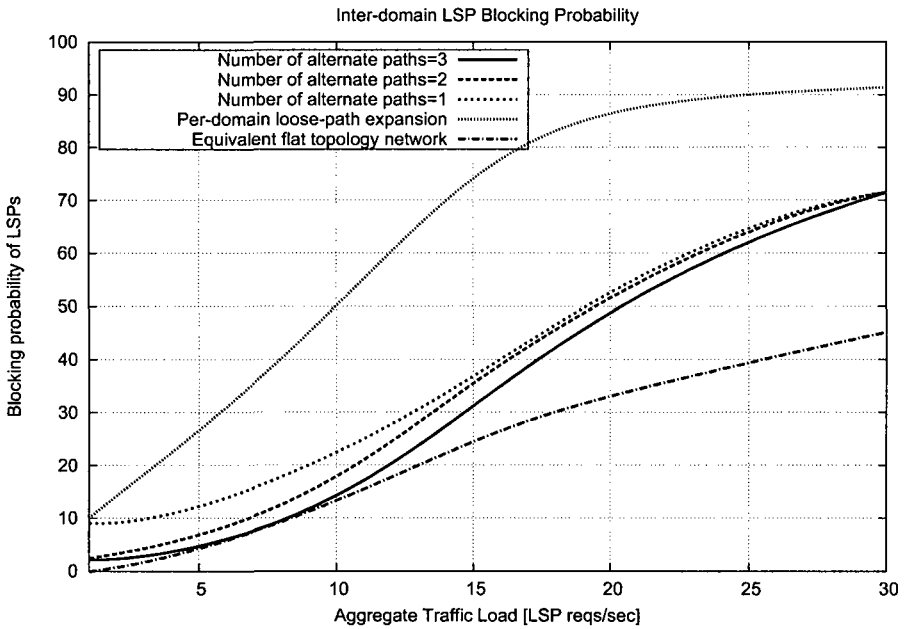


Figure 4.4: Average inter-domain LSP blocking probability versus traffic load for different path computation techniques

ability as a function of aggregated inter-domain traffic load when running simulations using 1) the extended-BRPC heuristic for $K = 1, 2, 3$ alternate inter-domain paths, 2) the per-domain loose-path expansion technique, and 3) and a hypothetical flat topology network that resembles the multi-domain network. Firstly, we notice that the LSP request blocking for simulation run using the flat network model experienced lower blockage than any of the other models/techniques. This is expected since the headend LSR, in this case, was able to autonomously compute the end-to-end path— without any additional delays due to contacting other PCE nodes, *etc.*. This minimizes the time needed to initiate the LSP signaling process, improving the chances of the request from being blocked along its signaling path. Secondly, we notice that the per-domain path computation technique experienced higher blocking rate that the other PCE-based technique. We can explain this since the per-domain loose path expansion technique relies solely on the inter-domain RIB to select the next domain hop without any consideration to the LSP request's availability requirement. When network loads gets higher, this shortest domain-hop path eventually get overloaded causing requests to get blocked when BNs fail to expand loose sub-paths within their domains. We can also notice that the LSP request blocking improved when more alternate inter-domain paths were attempted (*i.e.*, as K

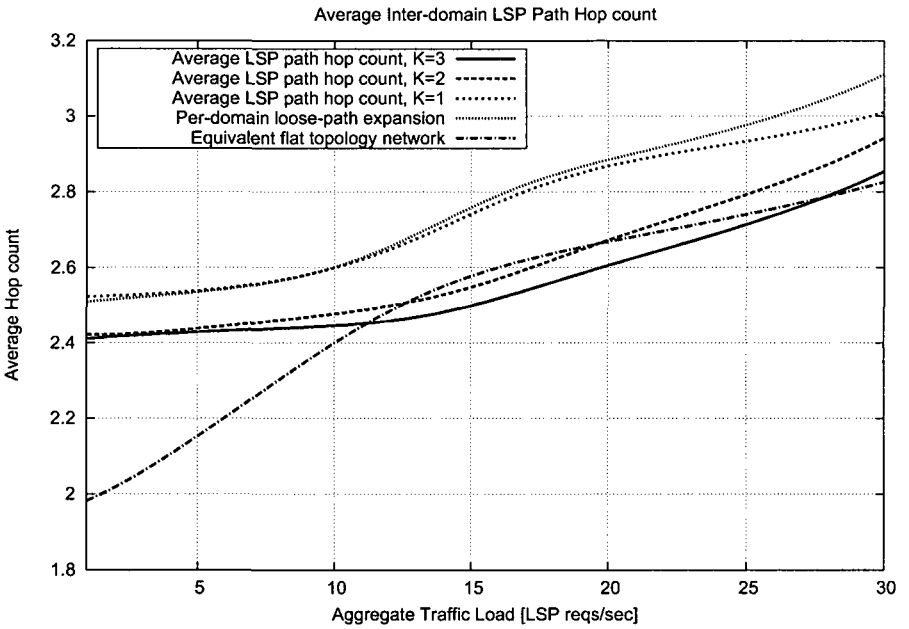


Figure 4.5: Average inter-domain LSP path hop count versus traffic load for different path computation techniques

increased). This is since the possibility finding feasible path satisfying the availability requirement improves as more potential domains are queried. Moreover, the SPCE will always select the least cost path among the multiple available feasible paths resulting in lower potential of the request being blocked later at signaling time.

Figure 4.5 compares average inter-domain LSP path hop count for each of the previously mentioned models. We notice at lower traffic rates, lower hop paths are initially favored. At higher traffic loads, links along the shortest path start to get overloaded, causing new requests to start favoring longer hop paths; hence, resulting in an increase in the end-to-end LSP path hop count.

Figure 4.6 shows the average LSP blocking probability per availability requirement when fixing the LSP request end-to-end availability requirement. As expected, requests with higher availability requirements experienced higher blockage than those with lower requirements. Again, the LSP blocking per availability requirement decreased more alternate domain-paths K were queried for feasible paths.

Figure 4.7 shows the inter-domain LSP request blocking per availability requirement class $0.96, 0.97, 0.98, 0.99, 0.999$ when varying the traffic load. Here we distinguish between two

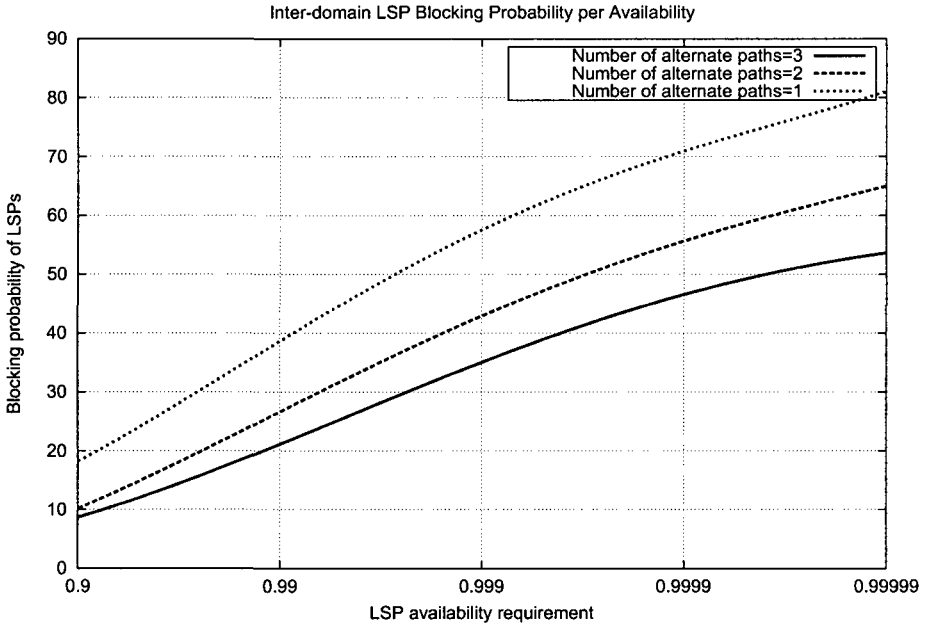


Figure 4.6: Average inter-domain LSP blocking type for different availability requirements at different traffic loads

types of LSP blockage: that 1) which happens prior to the LSP signaling (*e.g.*, due to no feasible path found), and 2) which happens while LSP signaling process is in progress– *i.e.*, after a feasible path is found. We notice that at lower traffic rates, the number of LSP requests blocked due to no feasible path found constitute most of the LSP blockage. At higher traffic rates, however, we notice that blockage due to unavailable resources, after feasible path has been identified, constitute the bulk of the LSP request blockage.

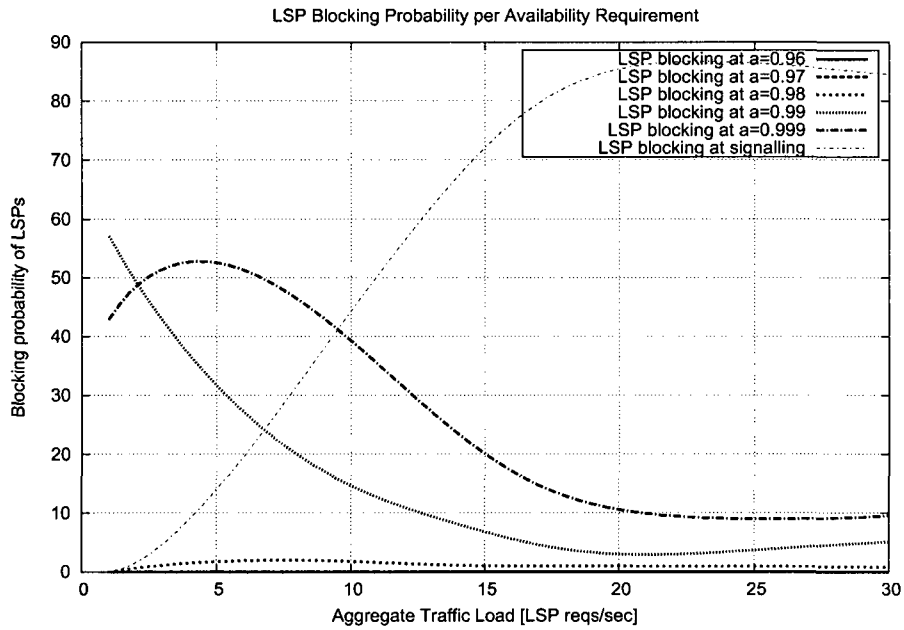


Figure 4.7: Inter-domain LSP blocking probability for different path computation techniques

4.7 Conclusions

This chapter has addressed the problem of finding a availability-bounded inter-domain feasible paths for TE LSPs that span multi-carrier networks. Specifically, we have proposed an algorithm takes into account the user requested bound on the LSP path availability as well attempts to minimize the overall cost of the LSP path across multiple alternate domain paths. We have shown that this problem is NP-hard to solve, even for the case of a single domain when the full TE topology is known a priori at the path computing node. The problem is further amplified in the multi-domain case due to the lack of exchange of network state information across competing meshed carrier networks.

Based on the PCE architecture, we have proposed a solution that extends the existing BRPC to allow polling for multi-service information along the domain-paths. The proposed scheme also involves contacting/querying multiple PCE(s) for virtual topologies along several alternate domain paths. The SPCE can, later, run a multi-constraint path heuristic on collected virtual topology in order to find a feasible path.

We have evaluated our proposal against two other schemes, the BRPC and the per-domain loose path expansion, by running simulations run on a multi-domain network using a self-developed inter-domain discrete-event simulation tool. We have shown superiority of our proposed scheme against the traditional BRPC technique as well as the per-domain loose path expansion. The per-domain expansion techniques does not take the user requirements beforehand when signaling the LSP. Hence, it is possible that the LSP be blocked due to no feasible sub-path can be expanded at any of the traversed domains. Our scheme also out performed the traditional BRPC, since ours will query multiple alternate PCEs in different domain for feasible paths.

Chapter 5

Dynamic PCE Selection for Inter-domain Path Computation

5.1 Introduction

The introduction of the PCE architecture is a key enabler to path computation in a multi-domain environment. The PCE itself is a generic term that encompasses several building blocks of the architecture including, the PCC and PCE routers, the PCE signaling Protocol (PCEP), and the PCE Discovery (PCED) procedures. There are various ways that PCEs can collaborate to compute inter-domain end-to-end paths; of particular interest to us in this chapter, is cooperative way that PCEs exchange path requests in the context of a specific end-to-end path computation prior to signaling the LSP.

Specifically, the selection of a specific PCE-chain from among a set of multiple available candidate PCEs that can service the path computation request, in each domain, can have a direct impact the total computation time of an end-to-end path. We show that depending on the information available to each PCE, it is possible to enhance the total path computation time for the inter-domain LSP path, and consequently improve the blocking probability of the LSP. The aim is to quantify the impact of different PCE selection schemes on the total path computation time and the amount of established LSPs.

We demonstrate through probabilistic analysis that by distributing information about the PCE congestion state, the end-to-end path computation can be made more efficient

in terms of the overall path computation time. To achieve this, we propose a novel heuristic for distributing the path computation requests among the available candidate PCEs and prove superiority of our heuristic over other techniques used.

In the remaining of the chapter, we study the effect of the PCE selection scheme on the path computation time and the LSP blocking probability. We also present heuristics that minimize the total path computation time and increase the probability of success for the LSP signalling. Finally, we evaluate the proposed PCE selection heuristics by running simulations using the inter-domain discrete event simulation tool over a generic multi-domain network.

5.2 Problem definition

The PCE architecture introduces the concept of policy in the context of path computation. When PCE-based computation procedures are used to compute LSP paths spanning multiple domains managed by different ISPs, policy becomes a key component of the architecture responsible for ensuring that requests are directed to a specific PCE according to pre-established contract agreements (*e.g.*, rate at which requests are sent, total amount of requested bandwidth, total number of TE LSPs, accounting, *etc.*). Furthermore, in some case, crossing domain boundaries requires constraint mapping (bandwidth pool, preemption, affinity, *etc.*) should the two neighboring domains make use of different conventions.

Requests for path computation of an inter-domain TE LSP can be performed by either using a single centralized PCE instance that has TE topology visibility over all of the other areas/domains, or can be distributed among multiple PCE instances— each responsible for each domain and usually referred to as the PCE chain— that cooperate to find the full path. PCEs may cooperate to find all domains along the path to destination, with full path information in each domain, or partial path information that will have to be expanded in each domain. What is common to all these scenarios is that a group of PCEs exchange information in response to a specific path computation instance, and generate path information that goes beyond the next domain ingress.

It is necessary to know the sequence of domains and PCE(s) when BRPC is used to compute the path of TE-LSP across multiple domains. One or more PCE routers, usually residing at the domain or area borders, are capable of performing optimal and/or diverse

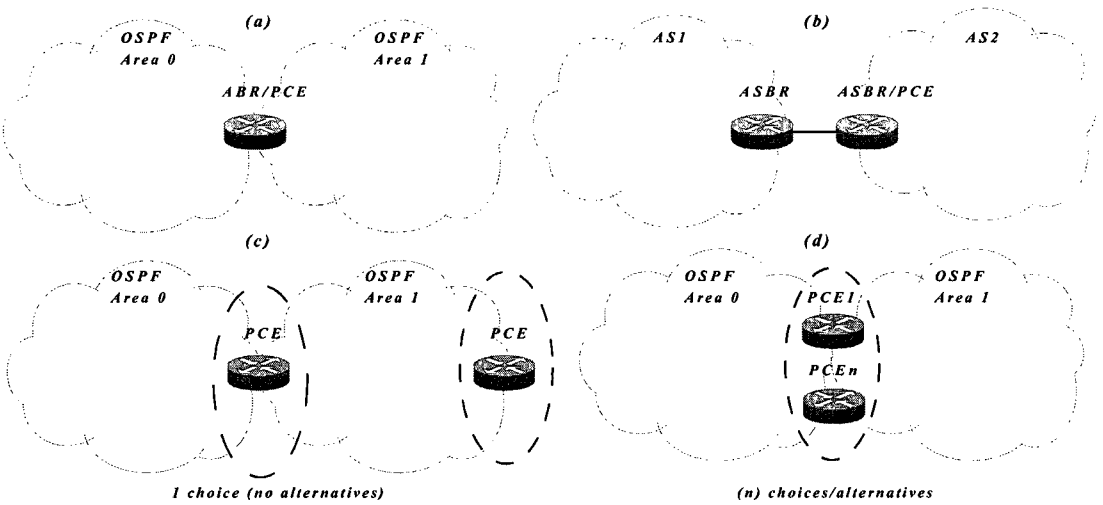


Figure 5.1: Examples of inter-domain PCE deployment. Above: (a) inter-area case, (b) inter-AS with inter-links. Below: (c) single PCE next hop, and (d) multiple PCE next hop alternatives

path computation for TE LSPs on behalf of other domain nodes referred to as PCCs. The head-end acting as a PCC determines and forwards a path request with the desired path computation constraints to one of the eligible PCE(s) in its domain (see Figure 5.2(a)). According the BRPC, once a PCE determines that it can not solely compute a end-to-end path, it can consult other PCE(s) present neighboring domain(s) to do so. If no next PCE can be found or if the next hop PCE of choice is unavailable, the procedure stops, and a path computation error retracts its steps back to the PCC.

When a PCE receives a request for which it has direct visibility to the TE LSP destination node in its TED, it performs a local path computation for the shortest path from all BNs to the destination node, forms a Virtual Shortest Path Tree (VSPT) of paths from all entry-BNs to destination and replies with it back to the upstream PCE. The upstream PCEs in their turn append their own paths from entry-BNs to exit-BNs until the VSPT reaches the Source PCE— the initiator of the request.

In cases where multiple PCEs are capable of serving the path computation request, the upstream PCE may select a subset of these PCEs based on some local policies or heuristics (see Figure 5.1). This downstream PCE node selection process that in turn defines the full PCE chain is crucial in the amount of overall time taken to compute the full end-to-end path.

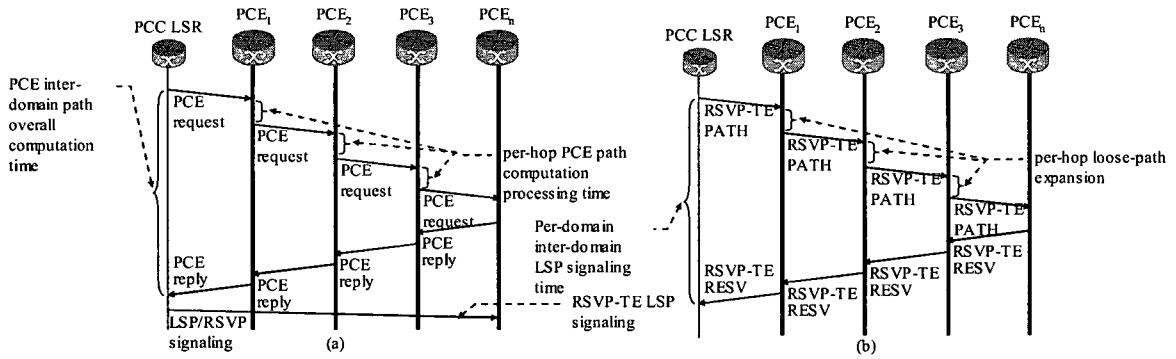


Figure 5.2: (a) PCE-based path computation and signaling time, and (b) Per-domain based loosely routed explicit path LSP overall path signaling time.

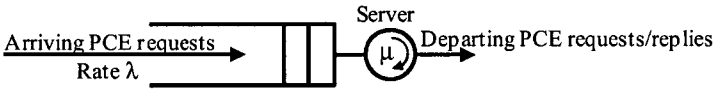


Figure 5.3: $M/M/1$ queue analytical model of a PCE server

Typically, routing information– *e.g.*, reachability announcement to the TE LSP destination (*e.g.*, by an area border router or autonomous system border router), or another user set routing policy is used to define a set of eligible PCEs that are capable of processing further the path computation request. However, among the set of candidate PCE(s), the decision to elect a certain PCE and forward the path computation request can significantly affect the overall end-to-end path computation response time depending on the degree of congestion on each of the selected PCEs along the PCE chain, and the available resources for the computation process.

Consequently, as the path computation time increases considerably, the signaled LSP can be potentially blocked at transit nodes due to reasons such as: 1) the PCC timing out while waiting for the path reply (*e.g.*, after a certain acceptable time-out elapses), or 2) a change in network state or use of stale information while computing the— *e.g.*, due to other competing intra/inter-area LSPs acquiring resources that were previously available, especially after a considerable time had elapsed since when the time of path computation in that domain was performed.

5.3 PCE selection for path computation

There are a number of schemes that can be considered to elect a preferred PCE from a set of a candidate PCEs that can collaboratively compute the overall end-to-end inter-domain TE LSP's path. Based on a eligibility criteria, a PCE determines a set of downstream PCEs. A cooperative PCE can then use a heuristic to rank and/or distribute PCReqs among the eligible PCE(s). In this section, we present two approaches that attempt to forward the PCReqs requests to PCEs that are least loaded.

In the first approach, the selection decision is done on per PCE-hop along the PCE chain to determine the preferred next hop PCE to further process the path computation request. Each PCE keeps a record for the response performance to its peering PCE(s) (*e.g.*, based on the path computation response time) and utilizes this in its decision making. We refer to this approach as per-hop PCE selection.

The second approach relies on keeping an average of the response times for every downstream PCE and sharing this state with a tier-1 PCE instance that overlooks all other domains. We refer to this as source-specified PCE approach (SSPCE).

We describe in more details the two approaches and present some results for simulations run over on the topology described in Figure 5.6.

5.4 Problem formulation

Assume that a PCE r can be modeled as an $M/M/1$ queue where path computation requests arrive at rate λ^r. Also assume that a request takes on average time τ^r to process– *i.e.*, with average path computation service rate $\mu^r = \frac{1}{\tau^r}$. The expected response time can be written as:

$$Rt^r_{exp} = \frac{1}{\mu^r - \lambda^r_T} \tag{5.1}$$

and λ^r_{Tot} being the total rate of arrival of path computation requests from all possible upstream PCE(s) to node r.

The expected number of requests q queued in PCE r can be written as:

$$Rq^r_{exp} = \frac{\lambda^r_{Tot}}{\mu^r - \lambda^r_{Tot}} \tag{5.2}$$

Considering the case where arriving requests at a PCE node can be forwarded to $\mathcal{M}$ downstream PCEs. The probability that any particular PCE request is directed to a particular PCE is $\frac{1}{\mathcal{M}}$, assuming eligible PCEs are equally probable. It follows the probability that exactly $\mathcal{X}$ out of $\mathcal{R}$ requests are directed to that PCE is:

$$\hat{P}(\mathcal{X}) = \binom{\mathcal{R}}{\mathcal{X}} \mathcal{M}^{-\mathcal{R}}(\mathcal{M}-1)^{\mathcal{R}-\mathcal{X}} \tag{5.3}$$

where,

$$\binom{\mathcal{R}}{\mathcal{X}} = \frac{\mathcal{R}!}{\mathcal{X}!(\mathcal{R}-\mathcal{X})!} \tag{5.4}$$

The following properties then apply:

1. The stability condition is given by $\lambda_{Tot} < \mathcal{M}\mu$.
2. The quantity $\rho = \frac{\lambda_{Tot}}{\mathcal{M}\mu}$ gives the utilization or load of a PCE.
3. The steady-state probabilities for $\rho < 1$ are given by:

$$p_0 = \left[\sum_{k=0}^{\mathcal{M}-1} \frac{(\mathcal{M}\rho)^k}{k!} + \left(\frac{(\mathcal{M}\rho)^{\mathcal{M}}}{\mathcal{M}!}\right)\left(\frac{1}{1-\rho}\right)\right]^{-1} \tag{5.5}$$

$$p_k = \begin{cases} p_0 \dfrac{(\mathcal{M}\rho)^k}{k!}, & k \leq \mathcal{M} \\ p_0 \dfrac{\mathcal{M}^{\mathcal{M}}\rho^k}{\mathcal{M}!}, & k \geq \mathcal{M} \end{cases} \tag{5.6}$$

For example, consider the case where a PCE takes on average $\tau = 2.5\,milliseconds$ to complete a TE path computation transaction, (*i.e.*, requests are serviced at rate $\mu = 400\,req/sec$), Figure 5.4 shows the expected number of PCE requests queued as a function of load ρ. As the total rate arrival of PCE requests approaches the rate of service the

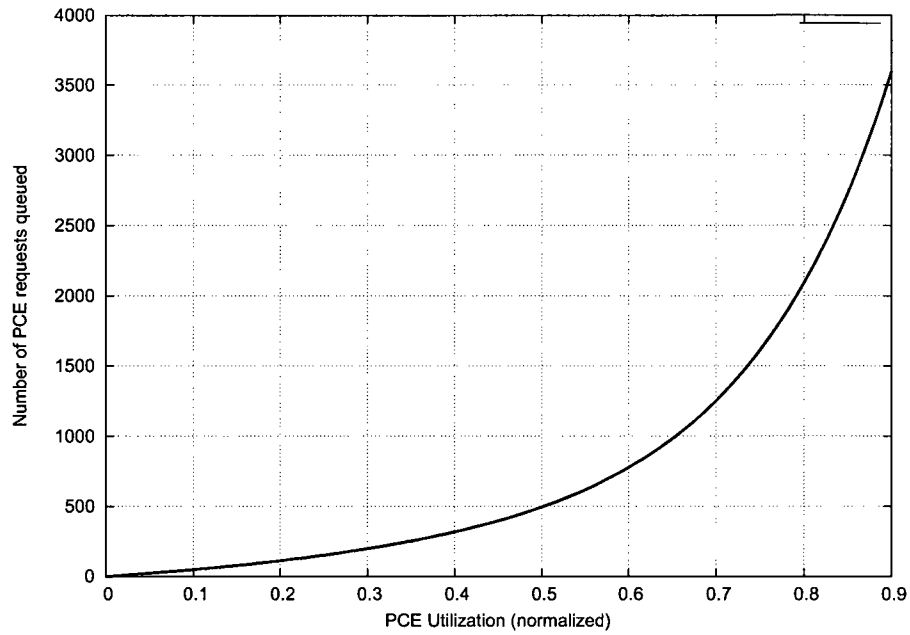


Figure 5.4: PCE path computation requests queued as a function of PCE utilization (1 downstream PCE)

path computation requests μ, the number of queued PCE requests and subsequently the expected response time increases significantly.

Within a time interval $Ti^{\mathcal{X}}$, assuming $\mathcal{X}$ requests arrive at a certain PCE r, there will be no queuing of requests at PCE r as long as $\mathcal{X}\tau < Ti^{\mathcal{X}}$. The latency experienced in this case for each request will be τ. When $\mathcal{X}\tau > Ti^{\mathcal{X}}$, the request processing rate is limited by the server, each request averaging one transaction in each period $\mathcal{X}\tau$. The extra delay per each request transaction is $(\mathcal{X}\tau - Ti^{\mathcal{X}})$.

We can observe from previous discussion that the number of redundant/eligible PCEs $\mathcal{K}$ that can process a certain path computation request plays a key role in bounding the maximum response time experienced by the PCE request. In general, the base (minimum adequate) number of PCEs must provide acceptable response time, so must not be very far into the overload region.

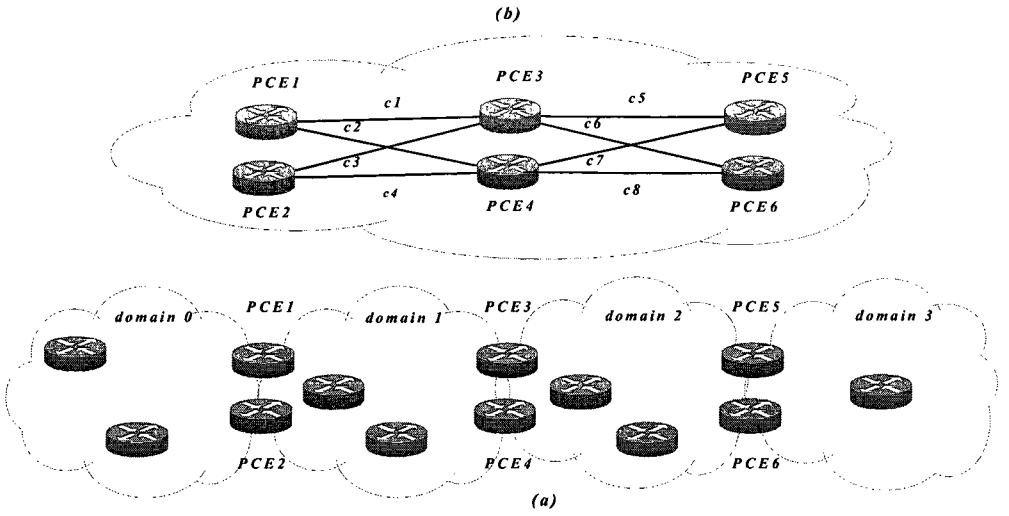


Figure 5.5: PCE traffic engineering virtual topology formation

5.5 Proposed heuristics for PCE selection

5.5.1 Source specified PCE selection

This scheme assumes that each PCE keeps track of mean aggregate arrival rate of path computation requests and uses this information to periodically flood a metric associated with the cost of using that PCE for a path computation. Each PCE creates a PCE virtual topology in the form of a graph $G(V, E)$ where V corresponds to the set of PCEs in all areas, and E the set of virtual links/edges between peering PCEs and whose link-weights reflect the amount of system utilization for that downstream PCE (see Figure 5.5). In summary, PCEs that are highly utilized will flood a large metric to be used for the virtual links and consequently will be avoided in any future path computations.

The originator of the PCE path computation request (*e.g.*, PCC) directs the PCReq to the PCE in its domain. In turn, the PCE computes the least PCE utilized path based on the topology of eligible PCEs. The PCE-hops sequence can then be carried in the PCReq message and used at each PCE hop to determine the downstream PCE.

In order to define a meaningful link cost associated with expected response time when using that PCE, we attempt to model each PCE as an $M/M/1$ queue. Equation 5.7 defines the PCE utilization ρ^r which represents the amount of load at any PCE router r. It is defined as the ratio of the total request arrival rate λ^r_{Tot} to the request service

rate μ^r at PCE r, *i.e.*, :

$$\rho^r = \frac{\lambda^r_{Tot}}{\mu^r} \tag{5.7}$$

The load intensity on each PCE which serves as an indication of the total number of requests queued awaiting processing at PCE r is shown in Equation 5.7. Note as $\rho \longrightarrow 1$, the cost the virtual link to use PCE r increases dramatically. This means that as the PCE server queue grows in size the PCE is penalized by increasing this cost so that incoming requests can use other PCEs to compute their path.

$$c_r = \frac{\rho^r r}{1 - \rho^r} \tag{5.8}$$

5.5.2 Per-hop PCE selection

This approach assumes the PCE selection is performed at each traversed PCE hop to elect a preferred next hop PCE to further process the request. Using this scheme, we present three heuristics to partition the computation requests among the candidate NHOP PCEs: 1) by equally partitioning requests (*e.g.*, round-robin distribution), 2) by forwarding requests to the least response time PCE, and 3) by adaptively partitioning the requests using token quotas heuristic based on an average response time to each candidate NHOP PCE.

5.5.3 Round-robin PCE selection

The Round-robin PCE selection (RRS) method assumes that requests are distributed equally in a round-robin fashion among a number of eligible PCEs that are capable of processing further the path computation request. In this scheme, requests from a certain source can be assumed to be locally distributed evenly among the available candidate PCEs. However, this does not guarantee global request balancing among the all candidate PCEs, and hence, can lead to some PCEs being overloaded with large queue of requests leading to increased delays in the overall path computation response. Algorithm 5.1 describes the steps included in executing this scheme.

Algorithm 5.1 Round-Robin PCE selection

pce_req_round_rob_sel(PCReq req)

Input:

req: PCE path computation request.

$S_{req} \leftarrow$ *eligible_pce_list(req)* {Extract eligible set of PCE(s) for request r}

$r \leftarrow$ *get_pce_longest_idle_time*(S_{req}) {Get PCE with the longest idle time}

Forward *req* to PCE r

set_last_req_sent_at(r); {Timestamp last time a request was sent to PCE r}

5.5.4 Least response PCE selection

The Least Response PCE (LRS) selection heuristic assumes that each PCE preserves locally an average path computation response time to each peering PCE. The downstream PCE would be always picked based on the least PCE response time. This scheme will achieve relatively better load request load balancing among PCEs. However, depending on the accuracy of the average response time stored locally, might not always yield the best PCE to forward to. Algorithm 5.2 describes the steps included in executing this scheme.

Algorithm 5.2 Least-response time PCE selection

pce_req_least_resp_sel(PCReq req)

Input:

req: PCE path computation request.

$S_{req} \leftarrow$ *eligible_pce_list(req)* {Extract eligible set of PCE(s) for request *req*}

$r \leftarrow$ *get_pce_least_resp_time*(S_{req}) {Get PCE with the longest idle time}

Forward *req* to PCE r

5.5.5 Token-based PCE selection

The Token-based PCE (TBS) selection heuristic also assumes an average response time is kept to each of the peering PCE(s). Arriving requests are partitioned among the eligible PCEs based on a token-based quota policy that is define based on the recorded average response times for each peering PCE. The main idea behind this heuristic is to penalize the PCE with the higher path computation response time by sending to it less path computation requests at any one time. Note, we assume in this case that

the partitioning quotas (tokens) can be updated periodically or once every time a path computation reply is received from downstream PCE. Algorithm 5.3 shows the steps included in executing this algorithm.

Algorithm 5.3 Token-based PCE selection heuristic

pce_req_token_sel($PCReq$ req)

Input:

req: PCE path computation request.

/* On arrival of PCE path computation request on PCE r */

while More PCE path compute request in queue **do**

$S_{req} \leftarrow$ *eligible_pce_list*(req) {Extract eligible set of PCE(s) for request req}

$r \leftarrow$ *get_pce_highest_tokens*($\mathcal{M}$) {Get PCE with the highest number of tokens}

Forward r to PCE r

$Tk_{req}[r] \leftarrow Tk_{req}[r] - 1$ {Decrement PCE request tokens for PCE r}

end while

/* On arrival of PCE path computation response on PCE r */

while More PCE path compute responses in queue **do**

Update path response time for PCE r

Update the new token array quotas $Tk^S \rightarrow f(R^S)$, where

$f(\cdot) \leftarrow\downarrow \frac{min(Rt^S)}{req_r}\mathcal{K}$

end while

Given a set of eligible PCE(s) S that can process a path computation request r, and the vector of path computation response times for S Rt^S:

$$Rt^S = \begin{bmatrix} rt_1 \\ \vdots \\ rt_n \end{bmatrix}, \quad \text{where } rt_{min} = Min(Rt^s) \tag{5.9}$$

The vector of tokens Tk^S for the set of PCE(s) S can be generated as:

$$Tk^S = \begin{bmatrix} tk_1 \\ \vdots \\ tk_n \end{bmatrix}, \quad \text{where } tk_i =\downarrow \frac{rt_{min}}{rt_i}\mathcal{K} \tag{5.10}$$

Table 5.1: Token quotas for $\mathcal{K} = 5$

Peer PCE	Average Response Time (ms)	Num. of Tokens
PCE1	110	5
PCE2	130	4
PCE3	140	4
PCE4	170	3

Table 5.2: Update Token quotas vector for $\mathcal{K} = 5$

Peer PCE	Tokens After					Avg. Resp. Time (ms)
	Req. #1	Req.#2	Req.#3	Req.#4	Req.#5	
PCE1	4	3	3	3	2	110
PCE2	4	4	3	3	3	130
PCE3	4	4	4	3	3	140
PCE4	3	3	3	3	3	170

where $\mathcal{K}$ is a constant that controls the size of the burst of path computation request that can be sent to any one preferred PCE r before a new update is applied.

Table 5.1 shows the token quotas computed for the set of eligible PCEs based on their respective average path computation response for $\mathcal{K} = 5$. The token quotas are computed according to Equation 5.10. $\mathcal{K}$ is chosen in order to limit large path computation request bursts to any one PCE at one time. High values of $\mathcal{K}$ would overload the selected PCE with many requests (big burst) before selecting another eligible PCE. Table 5.2 shows the updated token vector quotas after PCE receives a burst of 5 path computation requests. The table shows the updated token quotas after each request is assigned and forwarded to a PCE. After processing the full burst of requests, PCE1 receives 3 requests from this burst while PCE2 and PCE3 receive a single request each.

5.6 Evaluation of Proposed Heuristics

In this section, we evaluate the four interdomain PCE selection heuristics presented earlier in terms of path computation time for the inter-domain LSPs. The objective is to determine the quality of the paths computed by the different techniques, in terms

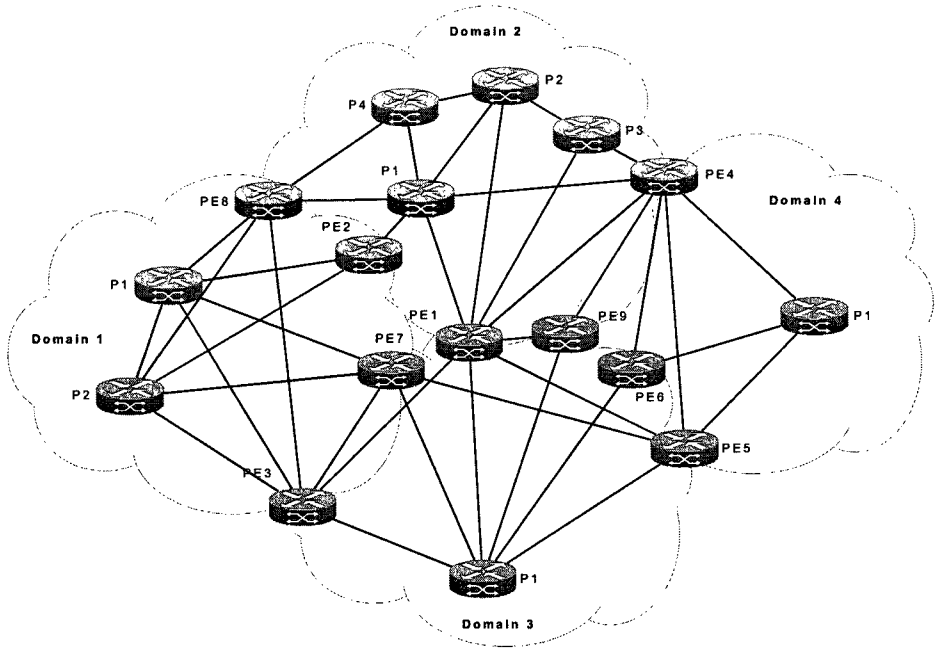


Figure 5.6: Multi-domain network topology used in simulation runs

of LSP blocking probability. We interest ourselves at the amount of traffic that can be carried inside the topology with each path computation technique. In addition, we aim at evaluating the amount of signalling required by the techniques.

We present results of simulations runs that are carried out using the discrete-event network simulator that we developed to replicate closely the behavior and timing of the overall inter-domain TE LSP path computation and signaling processes. The simulator will be described in more details in Appendix A.

To obtain meaningful results applicable to realistic provider topologies, simulations are carried out on a multi-domain network with multiple BN LSRs sitting at the edges of adjacent domains. Figure 5.6 shows the topology used for running the simulations. It consists of four domains and a total of 26 TE-LSRs, and 41 bi-directional TE-links. In this topology, domain-2 connects to all other domains and acts as a backbone domain for the network. TE Links are all assumed to be of the same capacity of OC-192 (9.6 Gbps). The propagation delays on links is assumed proportional to distance of the laid fibers between neighboring LSRs, and is generated from a randomly with normal distribution to reflect variations in distance of laid fibers between each router pairs.

The inter-domain TE-LSP requests are generated based on a *Poisson* process with mean

aggregate arrival rate of requests at each node λ_{Tot}. For each request, the ingress and egress LSRs s and t are equally likely chosen from the source/destination domain's set of LSRs. The hold-times for LSPs are exponentially distributed with mean holding time of $h = 10$ seconds. The bandwidth demanded by each of the TE-LSPs requests is set to one-tenth of the link's capacity. Any blocked request at any stage in signaling is dropped and no retrying is considered (*i.e.*, no crankbacks).

All LSRs in the simulated topology are assumed TE capable and able of processing RSVP-TE extensions. Intra-area LSRs are all assumed capable of either signaling per-domain loosed-path RSVP-TE LSPs, or acting as PCCs to generate PCE path computation requests for inter-domain LSPs. BNs are all assumed capable of handling either: per-domain TE loose-path expansion, or performing end-to-end PCE path computation.

In order to simulate the PCE processing delay due to the path computation process as per the discussion in Section 5.4, we simulated a PCE server as an $M/M/1$ queue. Path computation requests are serviced at rate $\mu^r = \frac{1}{\tau^r}$ – *i.e.*, each PCE takes on average a time τ^r to complete processing a path computation request. For our simulations, we assume $\tau^r = 2.5$ milliseconds (*i.e.*, requests serviced at rate = 400 req/sec). For the TBS scheme, the token request burst constant $\mathcal{K} = 5$ is chosen to limit large PCE path computation request bursts and avoid overloading the selected PCE.

To evaluate the effectiveness of the proposed schemes, we use as performance metrics the average PCE path computation time, the average LSP signaling time, and total LSP request blocking probability. The PCE path computation time is computed from the time a PCC initiates and forwards the PCE path computation request until it receives the corresponding path computation reply (see Figure 5.4(a)). The average signaling time is computed from the time an LSR dispatches an RSVP PATH discovery message to the time the ingress LSR receives a RESV reservation message declaring a full LSP establishment (see Figure 5.4(b)). TE LSPs are blocked due to: 1) no available resources on any of the traversed domains before initiating the LSP signaling (*i.e.*, no valid TE path that satisfies the LSP requested bandwidth), or 2) the resources on any of the traversed links/domains become unavailable after the LSP signaling has initiated. For the per-domain path computation approach, the LSP-signaling time includes the LSP signaling (RSVP-TE signaling time) as well as time to perform loose path expansion process on each of the BN LSRs. For each simulation run, 100,000 requests are generated and processed and the results for performance metrics are collected.

Figure 5.7 compares the performance of each of the proposed PCE selection schemes

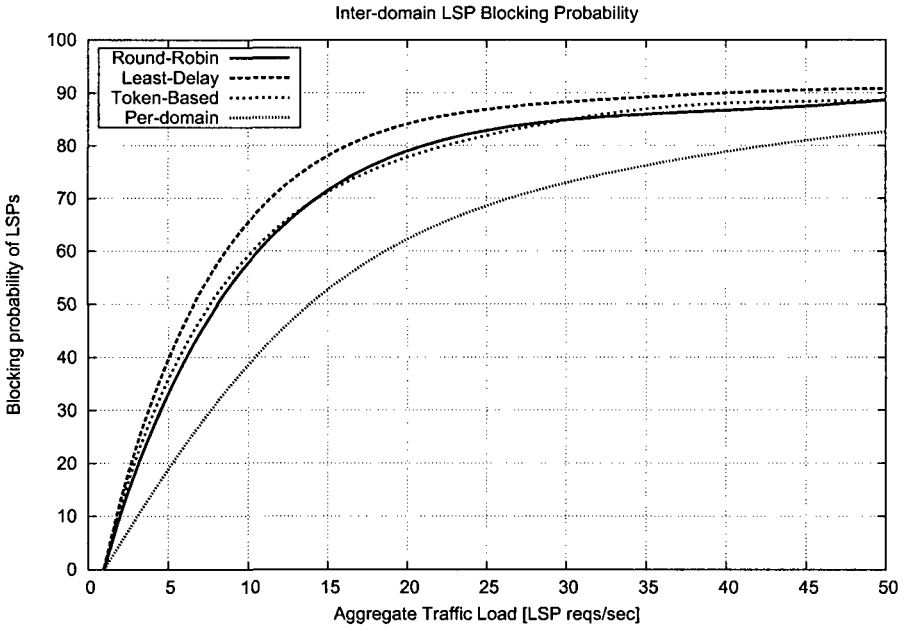


Figure 5.7: Comparison of the proposed PCE selection schemes

for the PCE-based approach: Round-Robin, Least-delay, and the Token-based, as well as results for the per-domain based approach (ERO expansion) in terms of blocking probability as a function of the aggregate rate of generated traffic. All PCEs in this run are assumed to have the same path request service rate. The results show that the TBS scheme performed closely to the RRS scheme and both better than the Least-Response delay scheme. Since service rates for all PCE LSRs are the same, the PCE request queues in each are likely to grow comparably. In this case, the both schemes will partition the arriving PCE requests almost evenly among the eligible PCEs. However, when having PCEs with mismatching request service rates, path request will be serviced at different rates on different PCEs. This yields to PCE queues growing unevenly among eligible PCEs, and response times varying from each PCE.

Figure 5.8 shows the results for the inter-domain LSP blocking probability when having PCEs with mismatching service rate- in this case, we set the rate on PCEs PCE1, PCE2, PCE3, and PCE5 to twice the rate of the other PCEs (see Figure 5.6). As results show, the Token-based scheme in this case out performed the RR scheme since it always prefers and forwards the computation requests to the PCE with higher number of tokens (less response time), while the RRS scheme partitions requests evenly among the eligible PCEs

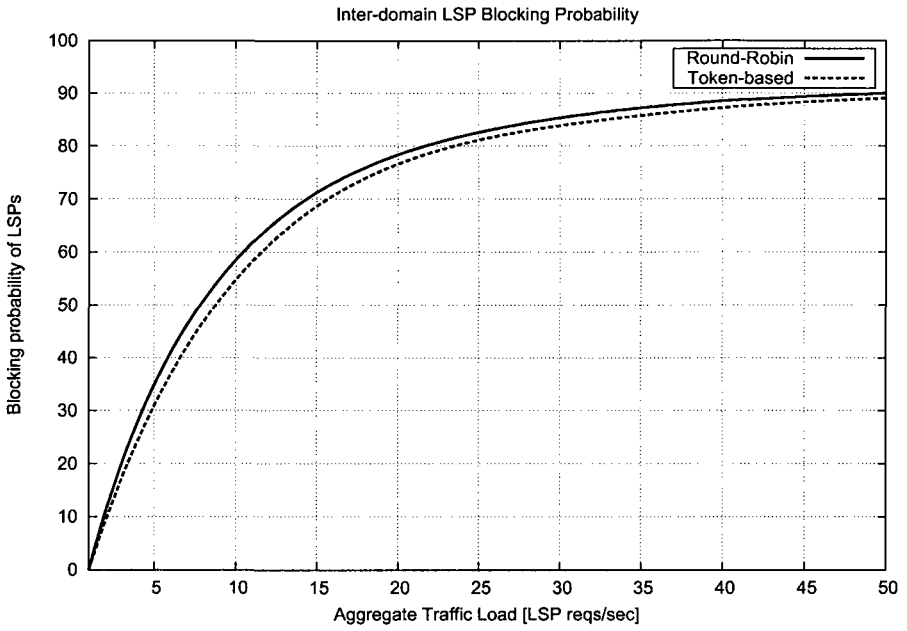


Figure 5.8: Comparison of the proposed PCE selection schemes

with no regard to downstream PCE load or expected response time.

Figure 5.9 compares the performance of previous mentioned schemes in terms of average PCE path computation time. From results, we observe that as the average path computation response time increases the probability of blocked LSPs also increases. This increase in blocking probability is attributed to link resources that were available at the time of path computation in transit domains being no more available at signaling time due to being claimed by other competing LSPs.

Figure 5.10 shows the average inter-domain LSP signaling time computed for the PCE based approach and the per-domain based approach. As shown, the average LSP signaling time in the per-domain approach is considerably larger than in the PCE-based approach. This is due to the fact that in the per-domain based approach the path computation time, performed for the loose path expansion at each border LSR, is accounted for in the LSP signaling process, while in the PCE-based approach the computation of the complete end-to-end path is done prior to signaling the LSP in this case. This is also demonstrated in the lower LSP blocking probability in the per-domain based approach shown Figure 5.7 as opposed to other PCE-based LSP establishments.

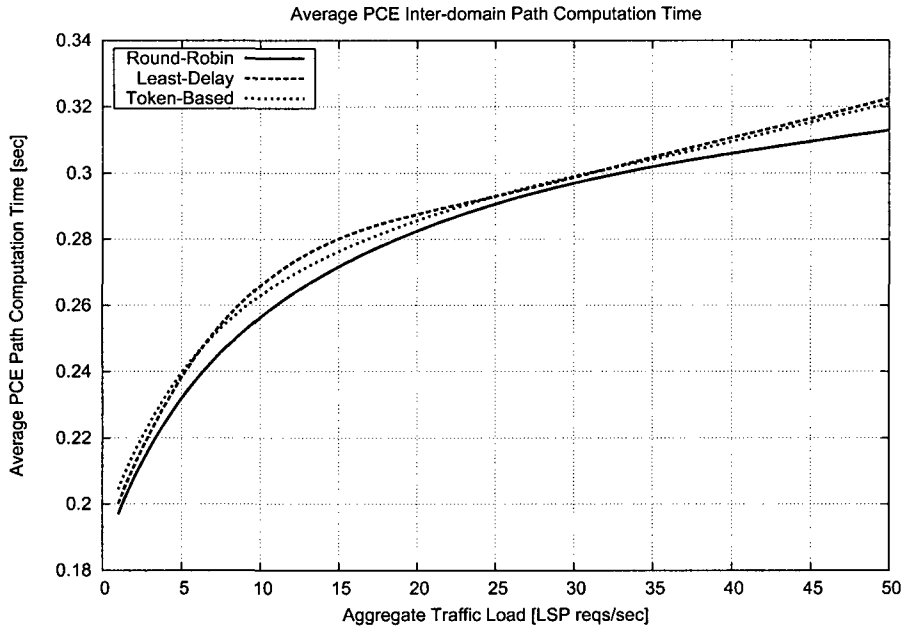


Figure 5.9: Comparison of the proposed PCE selection schemes

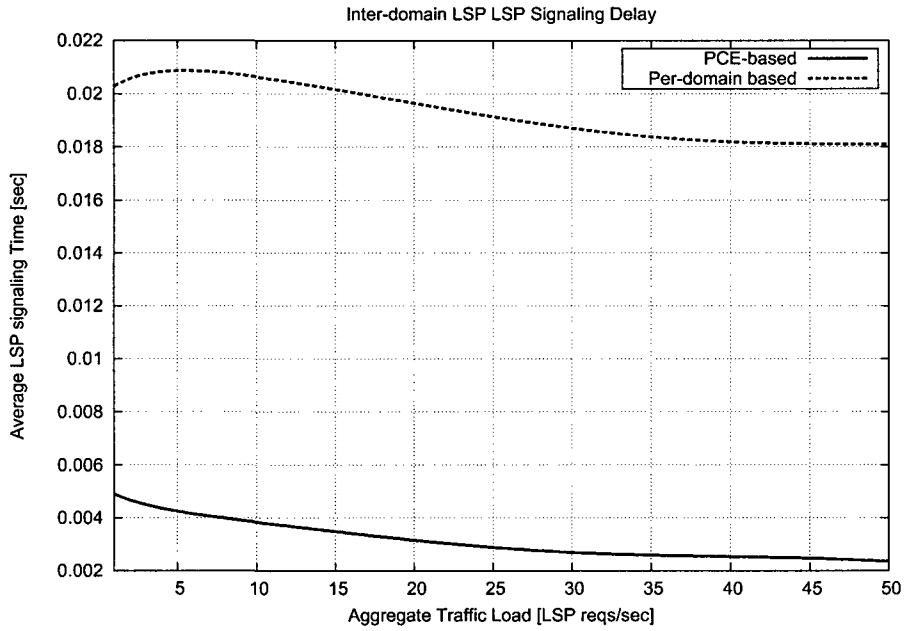


Figure 5.10: Comparison of the proposed PCE selection schemes

5.7 Conclusions

In this chapter, we examined the effect of per-hop PCE path computation delay on the overall path computation response time and its effect on the probability of inter-domain LSP signalling blocking.

We presented a number of schemes to distribute the PCE path computation requests among multiple candidate downstream PCE(s) so that the overall path computation response time can be minimized. Two basic approaches were highlighted: the source-specified PCE hops, and the per-hop PCE selection. The source-specified PCE hop approach relies on PCEs to flood a metric associated with its state load intensity. Other PCEs can use this information to compose a virtual PCE topology and pre-select a set least loaded PCEs that can be traversed to compute the overall end-to-end path.

In the per-hop PCE selection approach, we presented 3 different schemes to partition the requests locally among a set of eligible PCEs using: 1) round-robin PCE selection, 2) least response time selection, and 3) token-based PCE selection. The first scheme assumes that path computation requests are distributed evenly among eligible PCEs. The second scheme assumes that path computation requests are always forwarded to the PCE with the least response time. Finally, in the third scheme, partitions the PCE path computation requests adaptively among the set of eligible downstream PCEs based on tokens computed based the average PCE path computation response time for each. In our simulations, we showed that the latter scheme results in an improved global load balancing of the path computation requests among the candidate set of PCEs, and consequently, minimizes the overall path computation time for the end-to-end path and the LSP blocking probability for inter-area or domain LSPs.

Chapter 6

Extensions for P2MP Inter-domain TE Tree Computation

6.1 Introduction

The success of IPTV and, more generally, of multimedia transmissions over the Internet has recently increased the interest in Point-to-Multipoint (P2MP) transmission services. These are applications that require data delivery from a single source to multiple destinations. Hence there is an opportunity for transport service providers to sell very complex and valuable services that could deliver significant revenues.

Today, service providers who offer multicast services are facing the challenging task of efficiently delivering multicast traffic both within and across their networks. Traditionally, multicast routing optimization has been formulated as the well-known Steiner Tree (ST) problem [Hwan 92], with the major objective to deliver multicast traffic at least cost (*e.g.*, by consuming minimum bandwidth resources). This is achieved by calculating the path of the tree at the root node. The latter executes an algorithm that aims to minimize a cost function defined by the sum of links of the tree.

IP multicast provides P2MP communication. However, there are no TE capabilities or QoS guarantees with existing IP multicast protocols. TE and CSPF capabilities are crucial to enable and scale the intelligent management of network resources, prevent congestion, and enable sub-second fault restoration around network failures.

A P2MP LSP is a service that delivers data traffic with specified characteristics from a

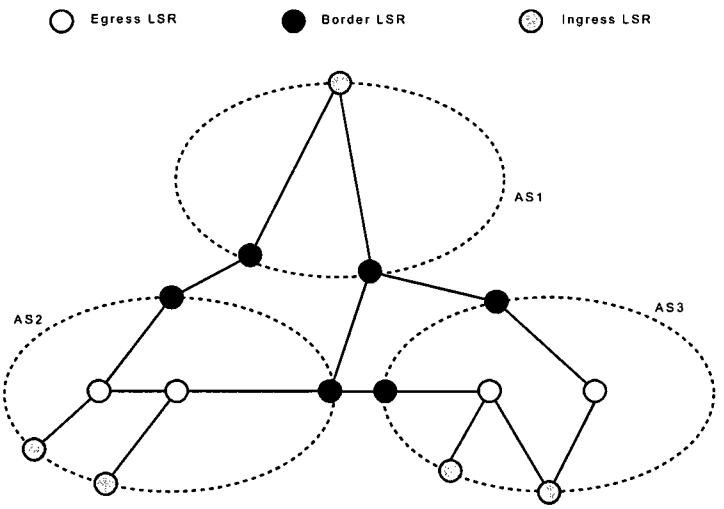


Figure 6.1: Inter-domain multicast tree construction

single source (P2MP root) to one or more destinations (or simply leaves) with an agreed-upon QoS, blocking probability or resilience against network failures. When realizing multicast applications using Point-to-Point (P2P) TE LSPs, the multicast groups are still duplicated at the source PE across each of the LSPs, even when one or more of P2P LSPs share links along their paths. Such scenario decreases the usefulness of multicast as a bandwidth-saving technology and increases the bandwidth used in the MPLS backbone linearly with the number of PEs.

A P2MP tree (see Figure 6.1) is a graphical representation of all TE links that are committed for a particular P2MP LSP. In other words, a P2MP tree is a representation of the corresponding tunnel on the TE network graph. A sub-tree is a part of the P2MP tree describing how the root or an intermediate P2MP LSPs solve the main problem of packet duplication where P2P TE LSPs traverse common links. The computation of a P2MP tree requires three major pieces of information. The first is the path from the ingress LSR of a P2MP path to each of the egress LSRs, the second is the traffic engineering related parameters, and the third is the branch capability information.

An explicitly routed P2MP LSP consists of a number of paths, each from the ingress LSR to an egress LSR, defined as Source-to-Leaf (S2L) sub-LSPs. These paths are set a priori (source routing), without requiring a multicast routing protocol in the backbone. The IETF's RFC 4875 [Agga 07] describes a solution to allow a non-ingress LSR to be a replication/branch LSR, able to replicate the incoming data on one or more outgoing

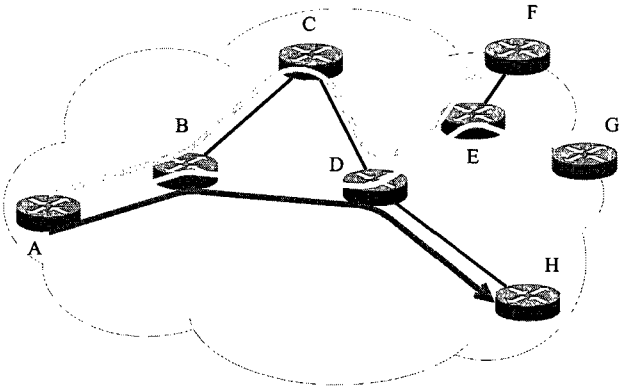


Figure 6.2: Sub-LSP remerge pair at (D): {A,B,C,D,E,F} and {A,B,D,E,G}. Sub-LSP cross-over pair at node (D): {A,B,C,D,E,F} and {A,B,D,H}

interfaces. Thus, an explicitly routed P2MP LSP is set-up by grouping multiple source-to-leaf sub-LSPs and relying on data replication at branch nodes; such a solution uses RSVP-TE as signalling protocol, without requiring a multicast routing protocol in the core network.

As with all other RSVP controlled LSPs, P2MP LSP states are managed using RSVP-TE extensions— similar to the P2P TE LSP case. One of the requirements of signaling a P2MP LSP is that the resulting LSP state must be "remerge-free" [Agga 07]. The term "re-merge" refers to the case of an ingress or transit node that creates a branch of a P2MP LSP that intersects with the P2MP LSP at another node farther down the tree. This may occur due to such events as an error in path calculation, an error in manual configuration, or network topology changes during the establishment of the P2MP LSP. Figure 6.2 shows an example of two sub-LSPs {A,B,C,D,E,F} and {A,B,D,E,G} that form a remerge pair since share the same output resources at (D), but do not share any input resources at same node. On the other hand, the two sub-LSPs {A,B,C,D,E,F} and {A,B,D,H} form a cross-over pair since they do not share both input and output resources at node (D). While cross-over sub-LSPs still indicates sub-optimal network resource usage, this however does not cause packet duplication at the receiver leaves which is undesirable.

As MPLS TE becomes the preferred choice for network operators, the computation of "remerge-free" P2MP tree paths to a high number of destinations is also becoming a challenge. This requirement becomes more acute to address when P2MP LSP spans multiple domains. This is because in an inter-domain environment, the ingress node may

not have topological visibility into other domains to be able to compute and signal a re-merge free P2MP LSP. This is because in an inter-domain environment, the ingress node may not have the full TE topological visibility into other domains to be able to compute and signal a re-merge free P2MP LSP.

Recent works carried within IETF have extended the MPLS for inter-domain tunnel set-up; for example in [Agga 07]. In an inter-domain environment, signaling for a given S2L or a set of S2Ls may contain loosely routed explicit LSPs paths. Many issues related to inter-domain P2MP TE LSP path computation and signaling of such loosely routed P2MP TE LSP such that it is remerge-free are still an open discussion.

In this chapter, we focus on multicast routing across multiple domains. We consider the problem of computing "remerge-free" inter-domain P2MP trees for the setup of P2MP TE LSPs. To do this, we propose novel heuristics for generating P2MP inter-domain trees that not necessarily yield the optimal ST, nevertheless insure "freeness" of remerges for its between S2L pairs. We also propose some extensions to the PCE architecture to allow such tree computation. Specifically, we propose the notion of passing the in-progress sub-tree within the PCReq when computing paths to new destinations within the tree.

This chapter is organized as follows. First we state the problem of that we attempt to address. We present a formal mathematical formulation to the problem. We proceed to describe the heuristics that we are proposing to solve the problem. We present results of simulations that we run using the several proposed heuristics. Finally, we conclude with our conclusions that we draw from this chapter.

6.2 Problem definition

Generally, inter-domain P2MP tree (*i.e.*, one source and a plurality of destinations across a plurality of domains, also referred to as inter-domain P2MP trees) is particularly difficult to compute, even for a distributed PCE architecture. For instance, while the BRPC recursive path computation may be well-suited for P2P paths, P2MP path computation involves multiple branching path segments from the source to the plurality of destinations. As such, inter-domain P2MP path computation may result in a plurality of per-domain path options that may be difficult to coordinate efficiently and effectively between domains. That is, when one or more domains have a multiple ingress and/or

egress border nodes, there is currently no known technique for one domain to determine which border routers another domain will utilize for the inter-domain P2MP tree, and to limit the computation of the P2MP tree to those utilized border nodes.

Also, such per-domain path computation may result in a merging of segments of the P2MP tree, *e.g.*, where two or more diverse segments from one domain are merged in a downstream domain, thus rendering the diverse segments in the upstream domain substantially unnecessary or inefficient.

A trivial solution to computation of inter-domain P2MP tree would be to compute shortest inter-domain P2P paths from the source to each destination and then combine them to generate an inter-domain Steiner P2MP tree. This, however, may require replication of incoming packets to all the P2P LSPs at the ingress PE (*e.g.*, when the branch point is chosen to be the head-end node) to accommodate multipoint communication. Obviously, this solution is very inefficient for a couple of reasons. First, it places more replication burden on the ingress PE and hence has poor scaling characteristics, and second it does not make use of bandwidth sharing when one or more S2L LSPs share links along their paths, hence wasting bandwidth resources, memory and MPLS label space in the network.

Apart from path computation difficulties faced due to the inter-domain topology visibility issues mentioned earlier, the computation of the minimum Steiner tree — *i.e.*, one which guarantees the least cost resulting tree– is an NP-complete problem [Wint 87]. Moreover, adding and/or removing a single leaf to/from the destination set may result in an entirely different tree. In this case, the frequent Steiner tree computation process may prove computationally intensive, and result in frequent tunnel re-configuration which may even cause network destability. There are several heuristic algorithms presented in the literature that approximate the result within polynomial time that are applicable within the context of a single-domain.

6.3 P2MP TE problem formulation

The P2MP tree can be determined in a variety of ways. For example, it could be fully computed or fully specified via configuration. It is also possible that a network operator specifies a P2MP tree as a root, a set of leaves, and an ordered lists of intermediate nodes and/or links that the tree should traverse from the root to all or some of the leaves. In

the latter case the entity that computes the P2MP tree is expected to expand the full strict path to produce the P2MP tree that satisfies the specified constraints.

Assuming the TE network is modelled as a directed graph $G(V, E)$ where V is the set of TE-enabled LSR nodes (vertices), E is the set of TE-enabled links connecting the LSR nodes, B the bandwidth (capacity) of links, and c_{ij} the cost function for TE link (i, j), such that $c : E \rightarrow \mathrm{R}^{+}$. Denote by $|V|$ and $|E|$ the number of TE nodes and links respectively.

Let Γ be a subset of V. A Steiner Tree St is defined as a connected a-cyclic sub-graph of G which spans all nodes of Γ. The cost of tree St^{Γ} spanning Γ destinations is defined as the sum of the costs of all edges of the tree, *i.e.*, $COST(St^{\Gamma}) = \sum_{\forall(i,j)\in St^{\Gamma}} c_{ij}$. Let $\overline{St^{\Gamma}}$ be the set of all possible trees in G that span Γ. The Minimum Steiner Tree (MST) problem consists of finding a connected sub-graph $St^{\Gamma}_{min} = (V_{St^{\Gamma}_{MST}}, E_{St^{\Gamma}_{MST}})$ where $V_{St^{\Gamma}_{MST}} \subseteq G$ and $E_{St^{\Gamma}_{MST}} \subseteq E$ such that it minimizes the objective function $COST(St) = \sum_{(i,j)\in E_{St}} c_{ij}$, for all $St^{\Gamma} \in \overline{St}^{\Gamma}$.

When a P2MP LSP request arrives at an ingress LSR, the P2MP-TE path computation algorithm is responsible for finding paths for each of its destinations; each path should have enough free bandwidth to support the requested bandwidth. If multiple paths exist for a given multicast tree, the computation algorithm needs to choose one so as to optimize a certain objective function.

Formally, the path computation problem for a P2MP TE tunnel can be stated in similar way to that of the P2P TE tunnel case; that is, compute a P2MP tree subject to a set of constraints for the tunnel that is designed to carry traffic with a set of parameters from the tunnel root to a set of leaves. The problem of designing optimal trees for P2MP-TE LSPs can be formulated as:

Let set V consists of three disjunct subsets: V_{in} set of the ingress (PE) LSR nodes, V_{eg} the set of the egress (PE) LSR nodes, and V_{cr} the remaining set of P core LSR nodes.

Let O be the set of all P2MP LSP. Each LSP $o \in O$ is characterized by an ingress node $s^{o} \in V_{in}$ a set of egress nodes $d^{o} \in V_{eg}$, and the requested bandwidth b^{o} for demand o.

The set of trees $\overline{St}$ in the network defines how to route traffic, where, $St^{o} \in St$ defines the tree for P2MP LSP $o \in O$. Consider the variables:

$$x_{ij}^{o} = \begin{cases} 1, & \text{if P2MP LSP } o \text{ uses link } (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

$$y_{ij}^{St} = \begin{cases} 1, & \text{if tree } St \in \overline{St} \text{ uses link } (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{6.2}$$

The objective: The objective for our optimization problem is set to minimize the weighted sum of the used resources over all the links in the network. The weight parameter α, where $0 < \alpha < 1$ is used to minimize the number of links used by a tree (*i.e.*, for larger α) or to minimize the total capacity used (for smaller α):

$$min \quad \sum_{\forall (i,j) \in E} \left[\alpha \sum_{\forall St \in \overline{St}^{o}} y_{ij}^{St} + \frac{1-\alpha}{b_{ij}} \sum_{\forall o \in O,\ \forall St \in \overline{St}^{o}} x_{ij}^{o} b^{o} \right] \tag{6.3}$$

Capacity constraints. Capacity constraints guarantee flows will not use no more than the link capacity.

$$\sum_{\forall o \in O,\ \forall St \in \overline{St}^{o}} (x_{ij}^{o}) b^{o} \leq b_{ij} \tag{6.4}$$

Flow conservation constraints. Flow conservation constraints guarantee that a demand goes from its source to its destination, and if it enters any other node it has to leave it as well.

$$\sum_{\forall j \in V^{i\rightarrow}} x_{ij}^{o} - \sum_{\forall k \in V^{\rightarrow i}} x_{ki}^{o} = \begin{cases} 1, & \text{if } i = s^{o} \\ -1, & \text{if } i = d^{o} \\ 0, & \text{otherwise, } \forall i \in V \text{ and } o \in O \end{cases} \tag{6.5}$$

where $V^{i\rightarrow}$ is the set of nodes $j \in V$ for which $(i,j) \in E$, and $V^{\rightarrow i}$ is the set of nodes $j \in V$ for which $(j,i) \in E$.

Demand-tree mapping constraints. A demand can be carried over a link only if the tree assigned to the demand is also set up over the link.

$$x_{ij}^{o} \leq y_{ij}^{St} \qquad \forall (i,j) \in E, o \in O, St \in \overline{St}^{o} \tag{6.6}$$

Tree conservation constraints. The tree conservation constraints can be stated as: 1) the root of a tree may be only at the ingress node, 2) the leaves of a tree may be only at egress nodes, and 3) no remerges should between any two paths in the tree.

$$\sum_{\forall j \in V^{i\rightarrow}} y_{ij}^{St} = \begin{cases} 0 & i \in V_{eg} \\ <= 1 & \text{otherwise} \end{cases} \tag{6.7}$$

$$y_{ki}^{St} \leq \sum_{\forall j \in V^{i\rightarrow}} y_{ij}^{St} \qquad \forall i \in V_{eg}, \forall k \in V^{\rightarrow i}, \forall St \in \overline{St}^{o} \tag{6.8}$$

$$y_{ij}^{St} \leq \sum_{j \in V^{\rightarrow i}} y_{ki}^{St} \qquad \forall i \in V_{in}, \forall k \in V^{i\rightarrow}, \forall St \in \overline{St}^{o} \tag{6.9}$$

According to the taxonomy of the multicast routing problems proposed in [Wang 00], our reference problem is "link-constrained", as the constraints depend on the bandwidth availability at links. It is worth noting that the QoS requirements— for example in terms of end-to-end delay, availability, jitter, and losses– would imply, in general, additional constraints at tree level, and thus an additional complexity to the mathematical model.

The optimization problem formulated above is NP-hard, mainly due to the NP-hard nature of computation of the Steiner trees [Chen 95, Gare 79]. The problem is further complicated in the interdomain case due to the limited topology visibility at the source to destination in remote domains. This makes such solutions inapplicable within the context of real-time tree computation within a dynamic network. Hence, other solutions that approximate the solution are generally preferred. In the next sections, we will propose heuristical approach to achieve this.

6.4 Heuristics for inter-domain tree computation

As mentioned, due to the NP-hardness of the ST computation [Wint 87], it is practical to propose heuristics to approximate it. In this section we present three such heuristics to solving this problem.

6.4.1 Simple generalized P2P heuristic

One way to solve the P2MP inter-domain tree computation problem is to break it into individual multiple P2P path computations and combine the resultant P2P paths to generate the P2MP tree. The resulting algorithm can be as follows:

1. compute the shortest inter-domain path subject to all constraints for each destination independently, and
2. join the sub-path(s) so that a P2MP remerge-free tree is generated.

We refer to this algorithm as the Generalized P2P (GP2P) heuristic. This heuristic, however, suffers from a number of problems. First, since the computed paths are oblivious of paths selected to other destination(s), 1) inefficiencies can arise from not encouraging selected paths to share common links, and 2) the resultant tree may have remerges.

To elaborate on (1), consider the problem where a tree is sought within graph G from a source s to a set of destinations Γ. Assume that there is a path of cost $wt_c(s,\nu)$ from the source to a node $\nu \in G$, and that links exist (each of equal cost c) that connect ν to each of destinations in Γ destinations. Denote by γ the number of destinations in Γ, *i.e.*, $\gamma = |\Gamma|$. In addition, assume point-to-point paths from the source to each destination $d \in \Gamma$ exists such that: $wt_c(s,d) = wt_c(s,\nu) - m^d$, and $m^d \gg c, \forall d \in \Gamma$.

In this case, the tree of shortest paths (*i.e.*, from source to each destination) does not use the path through ν and incurs a tree cost of $T_{GP2P} = \gamma \times (wt_c^{optimal}(s,\nu) - m^d)$. While a better tree cost would have been achieved if the path through ν was taken where $St_{\text{better}} = wt_c^{optimal}(s,\nu) + \gamma \times c$.

Note, using this scheme the additional step (2) is needed so that a "remerge-free" sub-path is guaranteed. In this case, one of the P2P paths causing the "remerge" has to be updated to correct this. It is important to note that resources (*e.g.*, bandwidth) are

shared on common links; Hence it is always better to adopt algorithms that maximize link sharing, and in turn reduce the cost of the resultant tree.

Moreover, in cases where Equal Cost Multiple Paths (ECMP) are across the path to the destination, the GP2P heuristic does not guarantee that two sub-LSPs traversing a transit domain, for example, will always use the same TE links. In fact, it is desirable sometimes to loadbalance sub-LSPs in the network which again will result in these types of inefficiencies. Secondly, this scheme does not guarantee "remerges" sub-LSP paths— *e.g.*, due some paths to destinations traversing different domain-paths– which is considered undesirable for end applications.

6.4.2 Incremental in-progress sub-tree heuristic

In this section, we propose a another solution to computing the inter-domain P2MP tree that we refer to as the Incremental in-progress Sub-tree (IS) heuristic. According to the IS heuristic, when a P2MP path computation request reaches the PCE in the source domain, the PCE_{src} computes the P2MP tree incrementally finding P2P paths for each of the destinations of the P2MP tree. Each time a path to a new destination from the destination set is determined, the in-progress P2MP sub-tree maintained at the PCE_{src} is updated. When the PCE in the destination domain, PCE_{dst} receives a P2MP path computation request that includes the in-progress sub-tree, it marks all links belonging to this sub-tree and present in its TE database with zero cost. It then attempts to compute a path from the entry-BN to the new destination using the updated TE database. The steps involved in executing the IS algorithm are summarized below:

1. When a P2MP tree path computation request reaches the source domain PCE, PCE_{src}, PCE_{src} chooses a destination d_1 from the destination set Γ that resides in the farthest AS– *e.g.*, has the longest AS/domain-hop path.

2. Using the inter-domain path computation scheme (*e.g.*, BRPC), a new request is formed and propagated to collaborating PCEs so the optimal end-to-end path for d_1 can is computed and returned to the requesting PCE, PCE_{src}. Note, if PCE_{src} receives a "no feasible path found" for d_1, it selects another destination from Γ and repeats above step.

3. PCE_{src} constructs/updates the in-progress sub-tree which is composed of known paths thus far for the subset of destinations in Γ.

4. PCE_{src} selects the next destination d_2 (*e.g.*, based on next longest AS-path), generates a new path computation request that includes the encoded in-progress sub-tree (*e.g.*, in an SERO object) and initiates a new end-to-end path computation to d_2.

5. When the path computation request for a P2MP destination reaches the destination domain it inspects the presence of the in-progress sub-tree, and if found, PCE_{dst} marks all links belonging within the in-progress tree that are present in its TED with zero link costs. Note that such a cost assignment is reasonable since once traffic is delivered over a link to one destination, there is no extra cost (*e.g.*, no extra resources required) to deliver the data over the same link to another destination. Consequently, by doing so, the path selection process can be biased to use those links that are already selected by the same P2MP LSP.

6. PCE_{dst} runs a CSPF to compute a feasible path from any egress node in the in-progress tree that is local to the domain to destination d_2. If a path is found, the sub-path is appended to the in-progress tree and the in-progress tree is sent back to the PCE_{src}. If a "no" feasible path can be found from any node in the in-progress tree that are local to the destination domain to the destination, the path computation request traces its step to the upstream PCE in the upstream domain.

7. Step above is repeated to find a feasible path local to the current domain.

8. If "no" feasible path can be found traces itself back to the PCE_{src} in the source domain, the PCE_{src} can choose to select a different AS-chain for computation of the path to the destination.

9. The above steps are repeated to compute the full P2MP tree for all destinations in Γ.

Notably, we can infer a couple of observations specific to the above heuristic.

Observation 1: The tree computation does not require the knowledge of all destinations during the computation. That is, additional destinations can be incrementally added to the in-progress sub-tree after it is originally computed. Equally important is the fact that these additions do not alter the paths to existing destinations and hence do not cause re-configuration of the entire LSP every time a new destination is added.

Observation 2: The IS resultant P2MP sub-tree is always remerge free.

To prove the above, consider the problem where a P2MP tree is sought from source s to Γ destinations within a multi-domain graph $G(V, E)$, where each domain is modelled as a sub-graph $G_k(V_k, E_k)$, where c_{ij} is the link cost and $c_{ij} \geq 0$. Assume that the path to destination d_1 is first computed through the domain-chain $P(domain)^{d_1} = D_1^{d_1}, \cdots, D_n^{d_1}$, where $D_i^{d_1}$ can be any of source, transit, or destination domain for the path chosen to destination d_1.

We consider the case that k is a destination domain and generalize our proof to the other two cases. Assuming the sub-path $sp_{d_1}^k$ is selected within domain k for destination d_1. According to our proposed IS heuristic, the PCE in domain k will first mark all links along the path $sp_{d_1}^k$ with zero costs and then run CSPF to attempt to compute a path to d_2. In order for a remerge to happen, the sub-paths for d_2 and d_1 have to intersect at least at two nodes, v_i and v_j. In this case, the cumulative cost for the sub-path selected by path to d_2, $sp_{d_2}^k(v_i, v_j)$, should have lower cost to be preferred over sub-path $sp_{d_1}^k(v_i, v_j)$. However, since all links of $sp_{d_1}^k$ were marked with zero costs, this necessitates that $sp_{d_2}^k(v_i, v_j) < 0$. This is in contradiction, however, with the initial assumption that all edges have positive costs. Similarly, the same applies for sub-paths in the source/transit domains with the difference that according to IS, the PCE computes a CSPF to the BN instead of d_2.

Observation 3: The IS P2MP tree computation heuristic is highly dependent on the order in which the component paths are computed and may potentially lead to a sub-optimal tree. For example, consider the case presented earlier in Section 6.4.2, if the order of selection of destinations is d_1, d_2, d_3, the resulting P2MP tree will be St_1 with a tree cost $COST(St_1) = X$. On the other hand, if the order of selection of destinations is d_2, d_1, d_3, the resulting P2MP tree will be different, St_2, with a different tree cost $COST(St_2) = Y$.

To overcome this limitation/dependency, we propose to select the next destination based on the AS-path length. In this way, the first destination/leaf to be selected will always be the farthest (in terms of domain-hops). Subsequent destination will again be selected in farther to closer proximity. The full pseudocode for the algorithm executed for inter-domain P2MP LSP path computation is depicted below:

Algorithm 6.1 P2MP interdomain algorithm executed by head-end domain PCE

compute_interdomain_p2mp_path_src_domain(s, Γ)

Input: s: source

Γ: set of destinations

```
TreeInProgressTree ⇐ φ
QueuePQ;
for all t ∈ z do
  PQ ⇐ t
end for
sort_by_as_hops(PQ)
t := pop_longest_dest(PQ)
while (t ≠ φ) do
  PCreq_tree ⇐ compose_pce_request(t, InProgressTree)
  Pathp := brpc(PCreq_tree)
  if p == φ then
    t := pop_longest_as_hop_path(PQ)
  end if
  InProgressTree := append_path_to_tree(InProgressTree, p)
end while
return t
```

Algorithm 6.2 P2MP interdomain algorithm executed by PCE in destination domain

compute_interdomain_path_dst_domain$(PCReq, t)$

Input: $PCreq$: path computation request

t: destination

```
if pcreq_type ≠ P2MP then
  return brpc(t)
end if
InProgressTree ⇐ extract_p2mp_tree(PCreq)
if InProgressTree == φ then
  return brpc(t)
end if
for all l such that l ∈ TED do
  COST(l) ⇐ 0
end for
return min_path(InProgressTree, t)
```

6.4.3 Incremental in-progress sub-tree heuristic with clustering

As discussed in previous section, the IS heuristic dictates the computation of paths to destinations sequentially. To overcome this, we propose to enhance the IS a heuristic to take into consideration all the destinations of the tree at the initial computation time. Since destinations belonging to the same destination domain will follow the same sub-path segments in transit domains towards the destination domain. We propose that the source PCE clusters destinations residing in the same destination domain k together. A single path computation request with a subset of destinations Γ^k can then be forwarded instead of sending a destination at a time.

One possible way to do so is to select the destination that is equally far from all other destinations. Return BRPC for that destination as well as paths from that destination to all other destinations.

1. When the path computation request reaches the destination domain, the PCE_{dst} elects a destination from set Γ^k and computes the VSPT from all BN to this destination. In addition, PCE_{dst} includes the sub-paths from d to all other destinations in Γ^k.

2. When BRPC completes tracing its steps to the source domain, and a feasible end-to-end path is found to d, the paths to all other destinations in Γ^k are inferrred immediately without requiring any additional PCE signaling or computation.

3. The above steps are repeated for all k and destination sub-groups Γ^k.

6.5 Numerical results and analysis

In this section, we evaluate the P2MP inter-domain tree computation heuristics that were presented earlier. The objective is to maximize the number of established inter-domain TE source-to-leaf sub-LSPs in the network while minimizing the total network resource usage. We present results of simulation runs that were carried out using **intersim**– the discrete-event simulator described in Appendix A. Figure 5.6 shows the topology of the multi-domain network that consists of four fully-meshed domains with a total of 29 TE-LSRs, and 48 bi-directional TE-links. TE Links in the network are all assumed to be of

Algorithm 6.3 P2MP (Clustered-P2MP) interdomain heuristic executed by head-end domain PCE

compute_interdomain_p2mp_path_src_domain_cluster(s, Γ)
Input: s: source
Γ: set of destinations

```
/* Group destinations residing in same domain */
for all z ∈ Γ do
  d ← domain(z)
  Γ_d ← z
end for
for all Γ_d do
  Tree = compute_interdomain_p2mp_tree_src_domain(s, Γ_d)
  t = append_tree(Tree)
end for
return t
```

equal capacity. The propagation delays on links is assumed proportional to distance of the laid fibers between neighboring LSRs, and is generated from a randomly with normal distribution to reflect variations in distance of laid fibers between each LSR router pairs. All TE links in this simulation run were assigned unity TE link costs.

For dynamic traffic, the inter-domain P2MP TE-LSP requests were generated according to a *Poisson* process with a mean aggregate arrival rate of requests λ. We define a P2MP LSP as the container LSP that is composed of one or more source-to-leaf sub-LSPs. For each P2MP LSP request, the source and each destination of the sub-LSP(s) are equally likely chosen from the source/destination domain's set of TE LSRs. Holding time of each sub-LSP request follows a negative exponential distribution with mean holding time of $h = 10$ seconds. The bandwidth demanded by each of the P2MP TE-LSPs was ranged from 1-5 of the total TE link capacity. Any blocked sub-LSP request at any stage in signaling is dropped and no retrying is considered (*i.e.*, no crankbacks).

First, we compare the two described algorithms, GP2P and the IS, in terms of blocking probability of sub-LSPs under dynamic traffic load where sub-LSPs are established and torn down dynamically. Figure 6.3 shows the average inter-domain sub-LSP request blocking when ranging the traffic load from 1-30 P2MP LSP req/sec while limiting the

number of destinations leaves five per P2MP LSP. We can infer that the IS scheme yields better success ratio than that of the GP2P. This can be explained by the fact that the IS algorithm will prefer links that are used by existing sub-LSPs when grafting subsequent sub-LSPs for remaining destination leaves. This results in more sharing on links in the network and reduces the overall network resource usage. Results for the GP2P algorithm were higher due to the fact that it does not always prefer paths/links that are used by previous sub-LSPs, rather tries to compute the shortest cost path to the destination(s).

In this case of the hypothetical flat topology, the P2MP tree is computed by first selecting the longest destination leaf path, and then assigning zero costs to all links along its path. Subsequent paths for destinations leaves for the same P2MP LSP are computed by finding the shortest path(s) with the modified graph. Each time a new destination path is determined, its links are again assigned to zero costs until the full P2MP tree is computed. Clearly, the flat topology simulation run yielded better results than the other two schemes, mainly due to ability to immediately compute the path and signal the sub-LSP without having to wait up-front for the PCE(s) to compute and return the end-to-end path— a process which does not happen instantaneously, and that increases chances for the sub-LSP to be blocked after its signalled.

Figure 6.4 shows the average inter-domain sub-LSP request blocking when ranging the traffic load while increasing the number of destinations, in this case, to seven leaves per P2MP LSP. Again, we notice the blocking of sub-LSPs for the GP2P algorithm was higher than the other two schemes. Though the number of sub-LSPs per P2MP LSP increased, with the IS scheme the sharing property for links of the P2MP tree reduced the over-all resource usage, and in turn allowed more LSPs to be admitted on the links.

Figure 6.5 shows the results of the P2MP average LSP packing ratio per link for each of the considered tests while varying the number of P2MP destination leaves per LSP from [1-5]. We define the P2MP LSP packing ratio on each TE link in the network as the ratio of the number of P2MP sub-LSPs to the number of P2MP LSPs carried on the TE link. Higher LSP packing ratio per link indicates that each P2MP LSP uses less bandwidth to reach its destination leaves, and hence is desirable. In this test, the P2MP LSP requests arrive dynamically, but once the sub-LSP(s) are established, they remain connected throughout the simulation run and their resources are never released thereafter. We can notice, that the IS had higher LSP packing ratio per TE link than the GP2P which again asserts the results for the blocking probability earlier. The simulation run over the flat topology network had the highest LSP packing rations. Since the latter

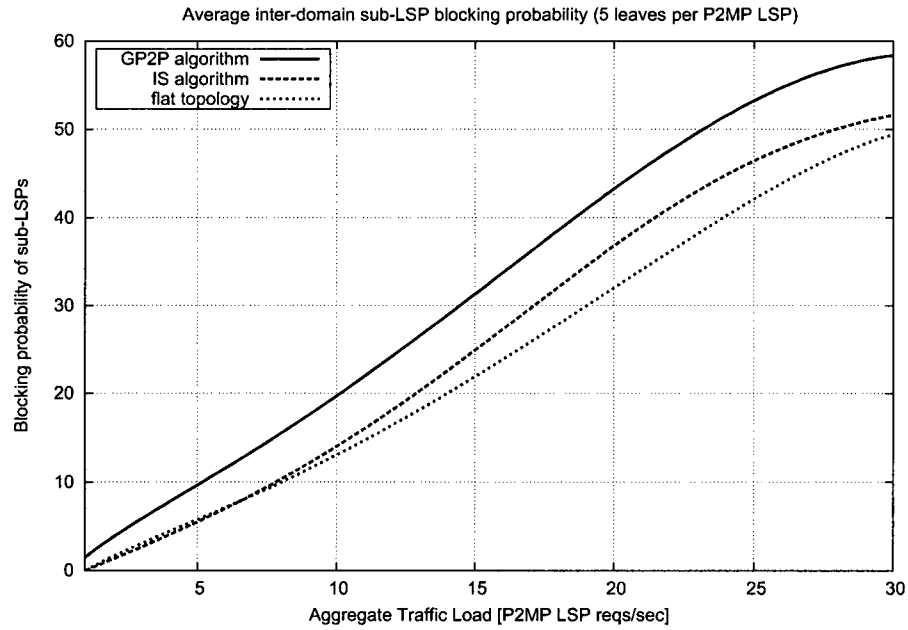


Figure 6.3: Average inter-domain sub-LSP blocking with five destination leaves per P2MP LSPs

run had the least blockage (*i.e.*, more sub-LSPs established), we deduce that more sharing happened on the links in this case.

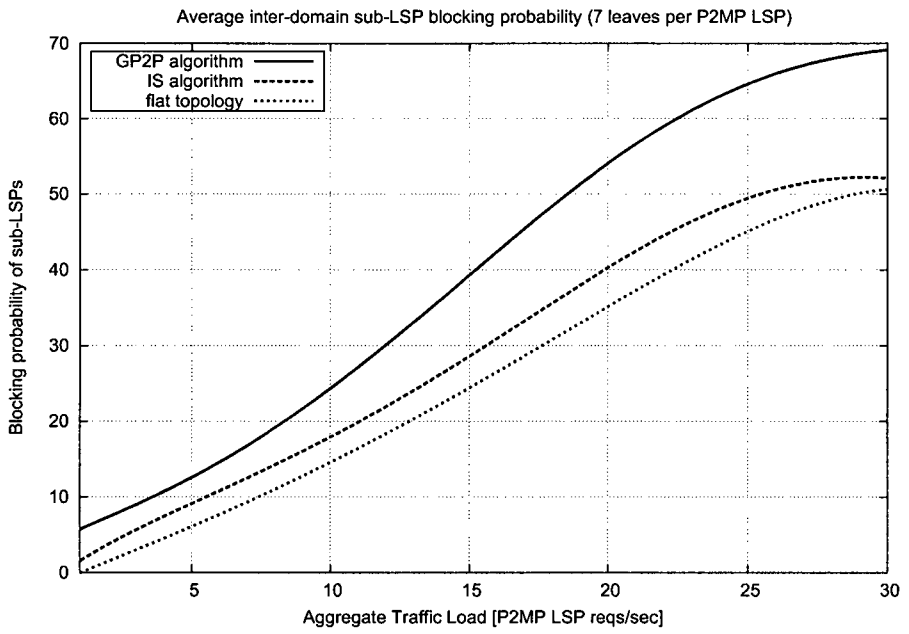


Figure 6.4: Average inter-domain sub-LSP blocking with seven destination leaves per P2MP LSPs

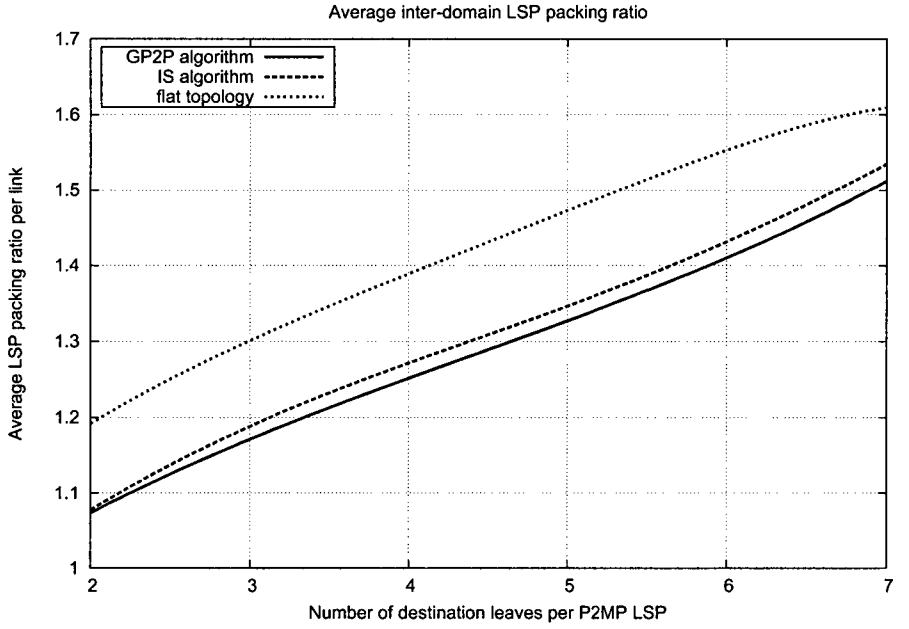


Figure 6.5: Average P2MP LSP packing ratio versus number of destinations per P2MP LSP

6.6 Conclusions

In this chapter, we studied the problem of inter-domain P2MP LSP tree computation which has gained recent interest due to the success of multimedia services over the Internet. First, we formulated the problem of TE within a multicast enabled MPLS network. We demonstrated, due to its complexity, the Steiner tree based solutions are generally not applicable within the context of realtime tree computation within a dynamic network.

We presented three novel heuristics for the P2MP tree construction for inter-domain P2MP TE LSPs under bandwidth constraints. The scalability and performance of the proposed algorithm were evaluated by comparing simulation results of the P2MP LSP blocking and P2MP LSP packing ratio per link. Simulation results have shown that the proposed IS heuristical solution is able to achieve significantly better performance than the GP2P approach. This was demonstrated by achieving higher LSP packing ratio per link which indicated that P2MP LSP overall uses less network resources to reach their destination leaves.

Chapter 7

Experimental Implementation of Multilayered VPNs

7.1 Background

Today, there is increased interest from service providers in offering diverse services and transport solutions for different traffic types over a single common infrastructure. A Virtual Private Network (VPN) is an overlay network that is built over a publicly shared network infrastructure, providing its users with a private and secure network using tunneling, encryption, as well as authentication mechanisms. VPNs are typically built over different types of provider networks, such as Ethernet, ATM, MPLS or the Internet. This solution has found popularity as an alternative to the expensive leased-lines or circuit-switched infrastructure.

VPNs are technically classified based on the underlying technology stack of the transport network that carries the VPN traffic (*e.g.*, layer-3 IP/MPLS, layer-2 SONET and ATM, or layer-1 WDM core switching networks), or the tunneling endpoint switching layer (*e.g.*, Layer-2 Ethernet, Frame Relay or ATM tunnels, or layer-3 IP tunnels). A layer-3 VPN forwards packets based on the VPN customerŠs IP information, while a layer-2 VPN forwards layer-2 frames based on information in customer's VLANs, MAC addresses, ATM Virtual Circuit Connection Identifiers (VCID), etc. Both of the previous mentioned examples, however, can be carried over a layer-1, layer-2, or layer-3 core based network. Two major flavors of VPNs exist: PE-based and the Customer Edge based (CE-based). Both offer transport services for layer-2 as well as layer-3 traffic.

In this chapter, we investigate several possibilities of inter-connecting customer layer-2 VLAN VPNs over a public network infrastructure. First, we consider transporting the inter-VLAN traffic over a Stacked VLANs (SVLAN)-based core backbone. Secondly, we consider the case of an IP/MPLS core network is used to provide the transport for layer-2 VPN traffic. In the latter case, we study two cases: 1) layer-2 VPNs using point-to-point MPLS tunnels, and 2) layer-2 VPNs using point-to-multipoint VPLS.

We also consider different implementations techniques to provision reliable services for VPN connection(s). This requires providing redundancy between end-points so that no link can be a single point of failure. Because of these requirements, redundancy solutions have to be chosen carefully. Several techniques exist to implement this, among them are: layer-2 Spanning Tree Protocol (STP), Ethernet Automatic Protection Switching, and MPLS Fast Reroute (FRR) and path-protection. Finally, we study the effect of protocol overhead for different tunneling techniques used for provider-based VPNs implementations.

7.2 Description of testbed hardware

The testbed that was used to carry out our experiments consisted of Nortel Networks, Cisco Systems, and Juniper products. The Optera Metro 1200 (OM-1200) switch-aggregator was equipped with 10/100 Mbps User to Network Interface (UNI) ports and two Network-to-Network (NNI) Gigabit Ethernet (GigE) interface-cards and were used as CE devices. The Optera Metro 8000 (OM-8000) MPLS LSRs were equipped with GigE ports and were used in the MPLS backbone. The Passport 8600 (PP-8600) router/switch was equipped with GigE ports and was used as an IP router and as a provider SVLAN-enabled layer 2 switch. The Baystack-425 and Cisco Catalyst-6500 switches were used as an 802.1q-enabled switches that also support layer 2 Class of Service (802.1p) and layer 2 multicast.

The Passport 15K (PP-15K) was used as CE device and was equipped with two OC-12 ATM over SONET interfaces, and two FastEthernet ports. The Juniper M-10, and M-160 were used as PE MPLS switches, and were each equipped with OC-48 Packet over SONET (POS), and OC-12 ATM over SONET interfaces. The OPtera Metro (OM-5200s) DWDM were used as an Optical Add Drop Multiplexer (OADM) equipped with OC-48 Packet over SONET (POS) Optical Channel Interface (OCI) tributary cards. Five

Intel-Pentium IV PC-machines were used as traffic sources/sinks in our tests.

For performance experiments and to validate inter-connectivity, we used Iperf test tool [Tiru 04] and the VideoLAN streaming server/client application to transfer unicast/multicast MPEG-2 video traffic over the network.

7.3 Case study: implementing layer-2 VPNs

7.3.1 Layer-2 VPNs over SVLANs

VLANs represent a standardized layer-2 mechanism for partitioning a single physical LAN into multiple disjoint logical LANs. VLANs are typically assigned based on the traffic port number, type of the protocol, the hardware address, or an explicit tag carried within the packet. Connectivity between distinct VLANs is achieved by routing at layer-3 usually done by a router device. The IEEE 802.1Q standard defines the packet format and required behavior for tagged-based VLANs. The Q-tag inserted in the Ethernet frames is limited to 12-bits, or 4096 unique VLAN Ids (VIDs). This limits the maximum number of VLANs accommodated within one network.

Stacked VLANs (SVLANs)— also known as Q-in-Q tagging– were introduced to address this shortcoming by adding an additional 4-byte Q-tag header to each tagged packet. SVLANs can be implemented in a provider network as a PE-based solution to provide connectivity to remote VLANs. Using this scheme, up to 4096 VLANs can be defined per customer, while the service provider uses up to 4096 SVLANs, increasing the maximum number of accommodated number of VLANs to 4096x4096.

The SVLAN technology has been widely adapted in Metro Ethernet applications due to its cost effective solution. Layer 2 Class of Service (CoS) can further be supported in the core network on per SVLAN basis.

For the testbed setup shown in Figure 7.1, the layer-2 Cisco Catalyst-6500 and Nortel Baystack-425 switches are configured as customer edge CE devices that carry 802.1q tagged traffic generated from PCs connected on each side. The two OM-1200s switches are also used as CE devices to transparently switch traffic received from PC hosts connected on each side. On the network core side, two layer-2 core switches PP-8600s, that are SVLAN-capable, receive tagged traffic from the Cat-6500 and B450, and untagged traffic from the OM-1200s and map it onto SVLAN-1 and SVLAN-2 respectively.

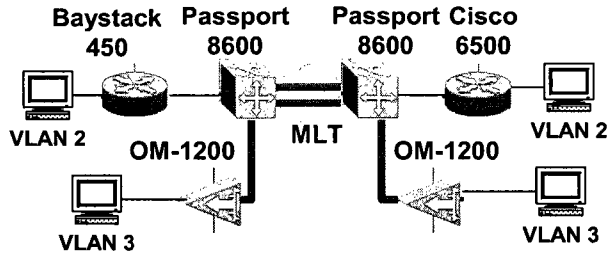


Figure 7.1: Testbed setup for MLT with SVLANs

The SVLANs are then assigned different CoS on the PP-8600s ($CoS_{SVLAN-1} = 1$, and $CoS_{SVLAN-2} = 5$) in order to test quality of service provided by the network core.

To provide link redundancy for the purpose of protection, we employ a layer-2 link bundling scheme referred to as Multi-Link Trunking (MLT) between the two PP-8600s on the GE links. MLT is a point-to-point connection that bundles multiple aggregated similar configured links into a single port group so that they actually act as a single logical port connection. By using this, higher capacity links with higher bandwidth throughput can be provisioned between end-points. A load-sharing algorithm is applied among the MLT links so as each link is activated for shared traffic transport.

The Iperf tool [Tiru 04] is run on each of the PCs in order to measure end-to-end packet loss for traffic sent across the provider core network. A fault is triggered at one of the MLT bundle member links. Figure 7.5 shows the packet loss that is incurred for each of the flows.

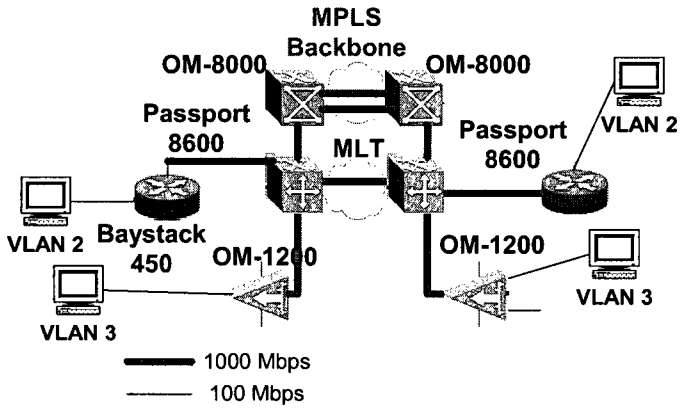


Figure 7.2: Testbed setup for MLT using SVLAN and L2VPN

Link protection with MPLS MLT In this scenario, the MLT is implemented using one physical connection between the PP-8600s and another virtual connection that is created over the underlying MPLS network described in section 5.2. The OM-8000s are added to the testbed as shown in Figure 7.2. The OM-8000s transparently map incoming packets from the PP-8600s into a VPN primary tunnel that was described in Section 5.2. Three test cases for packet loss are considered:

Testcase 1: single link failure when fault occurs on the link between the two PP-8600s. When this failure occurs, the traffic from the customer side is rerouted to the MPLS network and mapped onto the primary established tunnel.

Testcase 2: double link failure, the first fault on the link between the PP-8600s, and the second fault on the link of the primary protected MPLS LSP. In this case the customer traffic is automatically rerouted from the primary established tunnel onto the backup tunnel in the MPLS core network.

Testcase 3: double link failure, the first fault on the link between the PP-8600s, and the second fault occurring only at one of the primary tunnels in one direction in the MPLS core. In such case, the customer traffic is rerouted to the backup tunnel.

The packet loss recorded for each of the three testcases is shown in Figure 7.5.

7.3.2 Layer-2 VPNs over MPLS

MPLS also offers another way to deliver scalable, secure, and reliable VPN services. Using MPLS TE functionality and CSPF a better management of traffic and link utilization can be achieved in the provider network. MPLS LSPs can be setup to provide the generic tunneling service to connect segments of a VPN over a shared provider network, or define certain treatment (*e.g.*, class of service) for packets based on a defined filtering policy. The interior of an MPLS VPN network is made up of MPLS-aware P routers. PE routers surround the core devices and enable VPN functions.

Layer-2 traffic— such as Ethernet, Frame Relay or ATM– can be transported over an MPLS-enabled networks by realizing layer-2 VPN tunnels. Customer data can then be forwarded over the MPLS backbone based on information embedded in the layer-2

headers. Layer-2 VPNs are inherently transparent to the higher transported protocol layers, and hence support the transport of both IP as well as non-IP traffic. They also eliminate the need for service providers to participate in the customerŠs layer-3 routing with minimal changes from network.

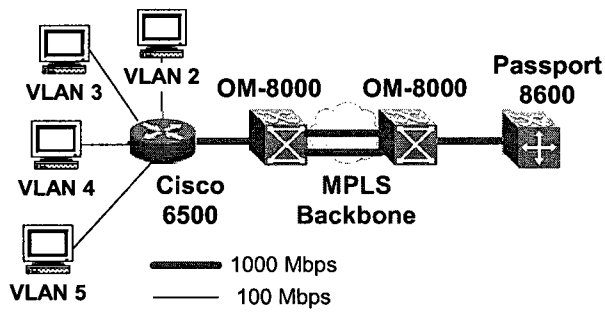


Figure 7.3: Testbed setup for point-to-point VPN over IP/MPLS

The testcase sets up MPLS P2P LSPs between PE routers. The Cisco Catalyst-6500 switch is connected to the MPLS network through the OM-8000 LSR that acts as a PE device. The other end of the MPLS network is connected to PP-8600 layer-3 router which is capable of performing routing between different VLANs (see Figure 7.3).

Four PCs are connected to the Catalyst-6500 and are also set to belong to four different VLANs (2-5). Two VPNs of with QoS service A (premium service) and QoS service C (best-effort service) are created to carry traffic from VLANs 2 and 3, and 4 and 5 respectively.

The OM-8000 LSR maps all tagged incoming packets arriving at the access port (UNI/CE side) from the Cisco-6500 to an established switched VPN based on the VCID. The traffic is then forwarded to the PP-8600 that acts as virtual router for routing between the different VLANs.

Testcase 1: We consider the effect of failure occurring at a link of the primary LSP in the MPLS network. In this case, the customer traffic is automatically rerouted from the primary established tunnel onto the backup tunnel. The packet loss recorded is shown in Figure 12.

Testcase 2: We consider the effect of a failure occurring at only one of the primary tunnels in one direction in the MPLS core. In such case, the customer traffic is auto-

matically transferred to the backup established tunnel in one direction only. The packet loss and throughput are shown in Figure 12 and 14.

7.3.3 Layer-2 VPNs using VPLS

Virtual Private LAN Service (VPLS) [Lass 07] addresses the problem layer-2 connectivity to multiple remote sites (one-to-any) in a transparent manner for CE devices. It emulates LAN across an MPLS-enabled IP network, allowing standard Ethernet devices communicate with each other as if they were connected to a common LAN segment.

In order to provide this service, client sites connected over the MPLS network expect that broadcast, multicast and unicast traffic gets forwarded to the proper site. Thus, the MPLS network must satisfy certain requirements: MAC learning, broadcast and multicast packet replication across LSP tunnels, and unknown destination unicast packet flooding.

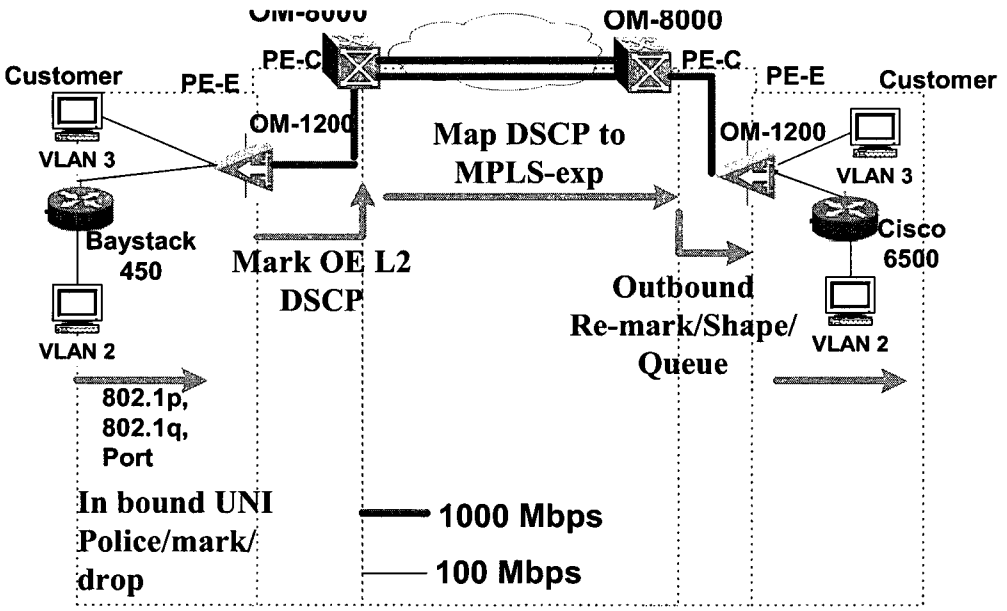


Figure 7.4: Testbed setup for VPLS VPN over IP/MPLS

In our experiment, we enable the Logical Provider Edge (LPE) aggregation architecture which is supported by OM-8000 LSRs. In this case, the PE functionality is split among two devices: PE-edge and PE-core. The PE-edge is represented by the OM-1200, and the PE-core by the OM-8000 LSR. The OM-1200 aggregates clients, maintains the for-

warding tables for each defined VPN service, signals to OM-8000 the addition of new VPN members, and applies QoS policies. The OM-8000 distributes and maintains MPLS labels, and exchanges information about VPNs with the other PE-core OM-8000. We configure the OM-1200 with two types of CoS assigned to respective ports/client as shown in Figure 10. The port connected to Catalyst 6500 is assigned a CoS equals to 4, while the one connected directly to the PC is assigned to the default CoS.

The OM-8000 marks the EXP bits in the MPLS label based on the OE L2 Diffserv Code Point (DSCP). Thus, the upper three most significant DSCP bits are mapped to the EXP bits. Similarly, on the egress, the OM-8000 sets the DSCP bits based on the EXP bits. Figure 10 illustrates the complete VPLS QoS solution on the OM-8000.

We consider the effect of the primary LSP link failure and that of a unidirectional tunnel failure. The packet loss and throughput are reported in Figure 7.7.

7.3.4 Experimental results

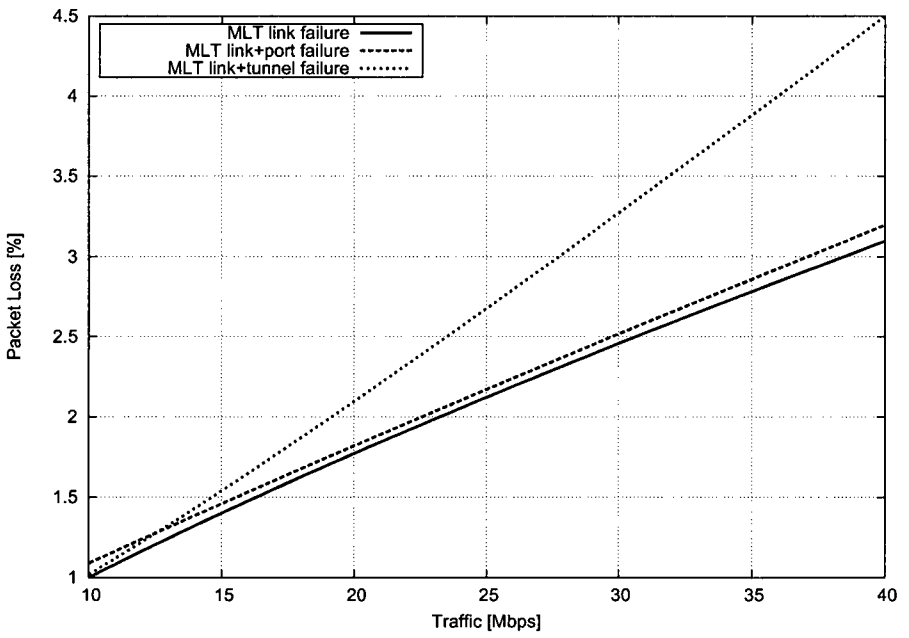


Figure 7.5: Packet loss using MLT with SVLANS and MPLS transparent VPN due to port and tunnel failures

Figure 7.5 describes the results of the experiments described in section 4.2.2. UDP traffic

was injected using the Iperf tool from one source-host to a sink-host across the testbed setup and the losses were recorded. As explained earlier, three types of failures are considered: 1) physical link failure between the PP-8600s, 2) (1) and virtual link failure (unidirectional tunnel case across MPLS backbone), and 3) (1) and virtual link failure (tunnels in both directions case across MPLS backbone). We concluded that packet losses were highest in case 3 due to two failures occurring consecutively on the 2 links. As described in Appendix B the tests were repeated over at least seven times at the of end of which the mean, upper and lower bounds were computed for the confidence interval of 95%. The confidence intervals were not shown, however, since they were negligible– example of the computations can be found in Appendix B.

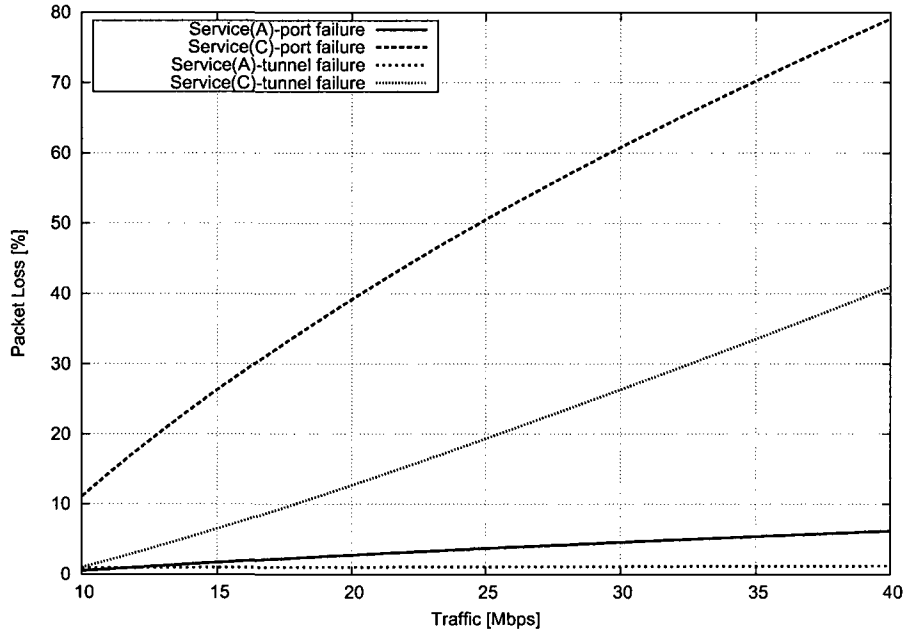


Figure 7.6: Packet loss using MPLS Switched VPN service A and service C due to port and tunnel failures

Figures 7.6, and Figure 7.7 show the results of packet loss experiments for the setup described earlier. In both cases, service A (premium service) exhibits lower packet losses than service C (best-effort service). After failure, service A is completely restored by re-routing its traffic onto the pre-configured backup tunnels, while service C is only permitted to transmit on the remaining bandwidth of the backup tunnels resulting in higher packet losses.

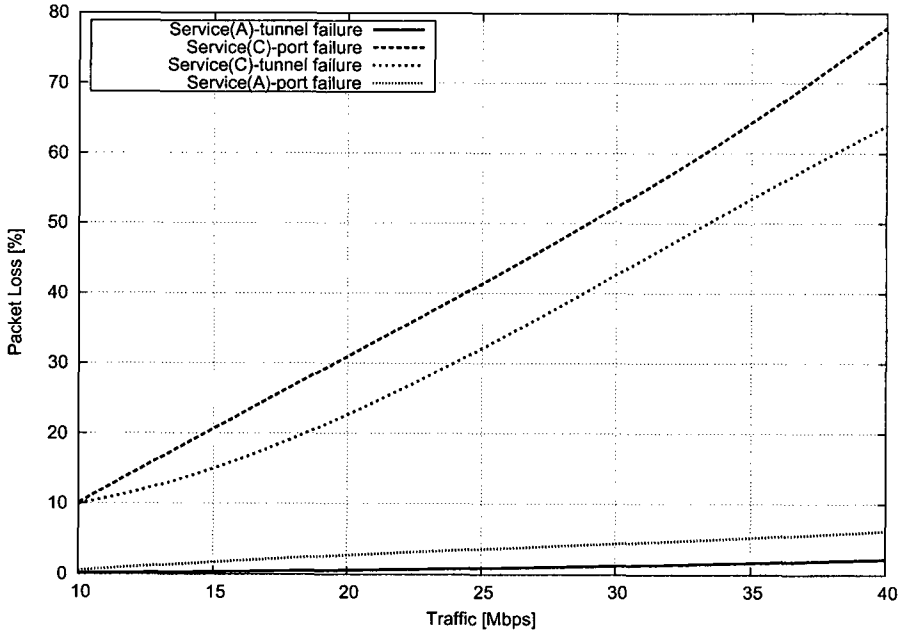


Figure 7.7: Packet loss using MPLS VPLS VPN due to port and tunnel failures

Figure 7.8 shows the measured throughput when 30 Mbps of UDP traffic is sent from source-host to a sink-host across the VPN implementation scenarios. The abrupt drop in throughput (at approximately t=6 secs) coincides with the time that the fault was injected on the primary connection. As expected, recovery from the failure was faster in the case of tunnel failure in one direction as opposed to port failure that brings down the tunnels in both directions of the primary connection. On the other hand, service C—which resembles a best-effort service in our test setup– was allowed only the available bandwidth on the backup tunnels (10 Mbps in this case) only after service A has been assigned its full backup capacity.

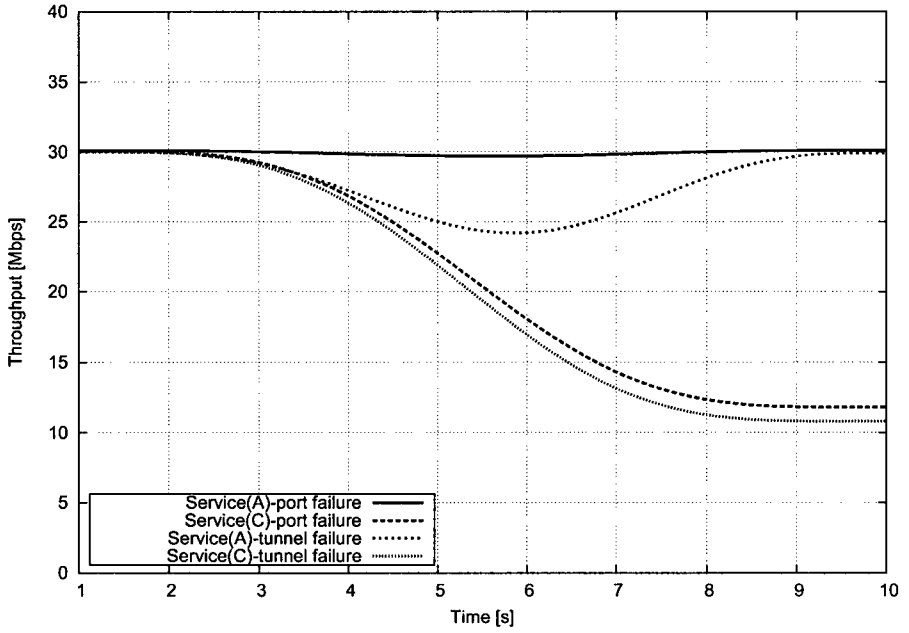


Figure 7.8: Throughput of service (A) and service (C) on MPLS Switched VPN due to port and tunnel failures

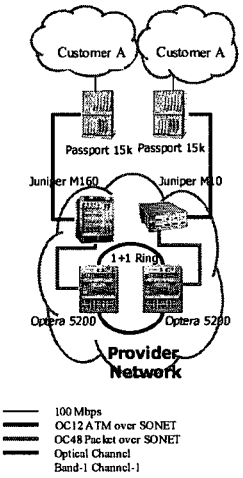


Figure 7.9: Layer-3 VPN testbed setup

7.4 Case study: evaluation of tunneling protocols overhead

As mentioned, tunneling is a technique for encapsulating a packet or frame within another packet of the same or a different network layer. One of the motivations for tunneling is bridging various heterogeneous networks that use different protocols for communication. Tunneling is also used for providing private and secure communications over a publicly shared network.

In this section, we study the effect of protocol overhead for tunneling techniques of different provider-based VPNs implementations, including: 1) Layer-2 virtual circuits using Generic Routing Encapsulation (GRE) [Hank 94] tunnels, 2) layer-2 trunks using Ethernet SVLANs, 3) layer-2 tunnels using L2TPv3 [Lau 05], and 4) layer-2 transparent bridging across heterogeneous networks using GRE and L2TPv3 tunnels. Figure 7.11 shows the several configuration setups that were used to implement the mentioned schemes for layer-2 VPNS.

We compare the performance characteristics of various tunneling implementations (see Figure 7.10) and their impact on the end-to-end data throughput. In each case, an end-to-end virtual connection is established using a tunnel or a series of concatenated tunnels spanning different heterogeneous provider domains each implementing a different tunneling technique. The concatenated tunnels constitute a virtual connection that inter-

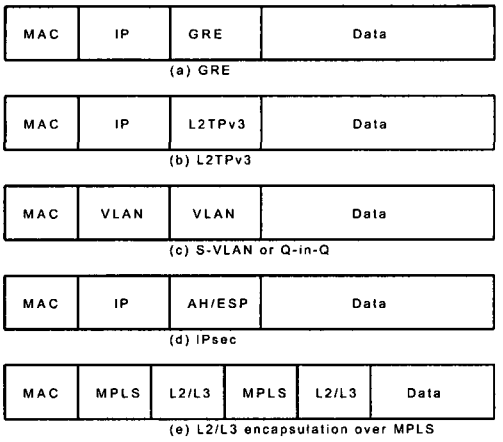


Figure 7.10: Tunneling and encapsulation protocols

connects remote customer VLAN sites.

Generally, there are two types of inter-provider relations that coexist between different network operators, namely: the peer-to-peer model, and the client-server model. In the client-server model the client domain requests service that the server domain offers. Client network domains receive transport services from their service provider to reach each other. In the peer model, a client network domain is treated as a peer of its service-provider; hence, peer domains not only receive transport services from other participating domains but also contribute new transport services to other domains. For our tests, the Iperf tool was used to generate traffic and measure throughput. Iperf is a tool capable of measuring a number of parameters including bandwidth, throughput, delay jitter, and packet loss.

Testcase 1: Layer-2 virtual circuit using GRE. For this testcase, the network is partitioned into two domains: CRC IP-domain, and the ONRL-L2 MPLS domain. The ONRL-L2 domain runs MPLS tunneling service to provide transparent transport for Ethernet VLAN-tagged traffic between the two end PC-clients in the remote VLANs as shown in Figure 7.12. The CRC domain is totally contained inside the ONRL-L2 domain and runs Ethernet over IP-based GRE tunneling service in order to stitch LSPs crossing it from side to side. Ethernet encapsulated MPLS packets arriving at one of the CRC border routers are encapsulated within IP/GRE packets, and then forwarded towards the other border router where it gets decapsulated back to an MPLS over Ethernet datagram and forwarded to the neighboring LSR. Figure 7.13 shows the packet encapsulation at

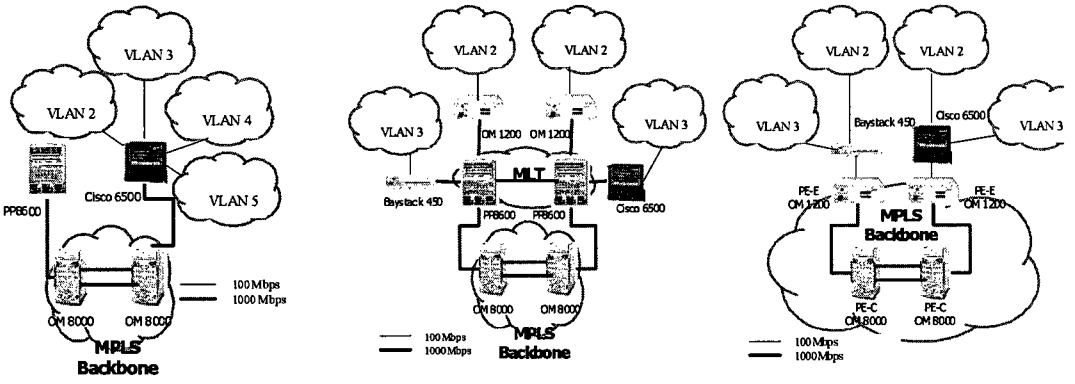


Figure 7.11: Layer-2 VPN Testbed setup: (a) Switched L2-VPN. (b) Transparent L2-VPN. (c) VPLS L2-VPN

each of the mentioned stages.

Testcase 2: layer-2 trunking using SVLANs. For this experiment, the ONRL-SVLAN and the ONRL-L2 MPLS domains are configured to provide a tunnel connection for the remote PC-clients present in the customer VLANs as shown in Figure 7.13. The ONRL-L2 provides transparent transport for Ethernet VLAN-tagged traffic over MPLS L2 tunnels. The ONRL-SVLAN domain is totally contained inside ONRL-L2 domain and runs SVLAN tunneling service in order to stitch LSPs crossing its domain boundaries. The Ethernet encapsulated MPLS packets arriving at one of the ONRL-SVLAN border-routers are wrapped within SVLAN packets and forwarded towards the next border router where they are decapsulated back to MPLS over Ethernet frames and forwarded to the neighboring LSR. Figure 7.13 shows the Packet encapsulation at each of the mentioned stages.

Testcase 3: Ethernet transparent bridging using L2TPv3. the CRC IP domain and the ONRL-L2 MPLS domains are configured to provide the same service for the PC clients in the customer VLANS as shown in Figure 3. The CRC domain uses L2TPv3 to establish tunnels that carry VLAN-tagged Ethernet frames encapsulated over MPLS. The Packet encapsulation at each of the mentioned stages is shown in Figure 7.12.

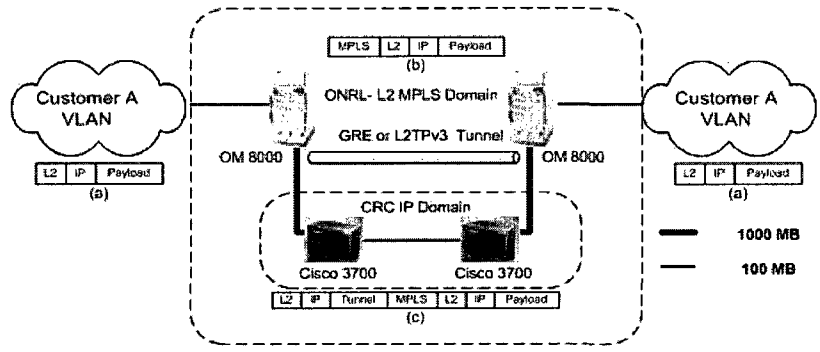


Figure 7.12: Testbed setup realizing 1) layer-2 virtual circuits using GRE tunnels, and 2) layer-2 tunnels using L2TPv3. Packet encapsulation for customer traffic: (a) in VLAN A, (b) across layer-2 MPLS network, and (c) across GRE or L2TPv3 tunnel

Testcase 4: Ethernet transparent bridging across heterogeneous networks. In this experiment, the CRC, ONRL-L2 and ONRL-L3 are arranged as peer domains that provide transit tunneling-service to end-to-end connections between the PCs in the customer VLAN networks. The ONRL-L2 MPLS and CRC IP domains are configured to provide the same services as described in the experiment in Section III.C.1 and III.C.2. The ONRL-L3 domain is configured to provide a layer-3 tunneling service over MPLS as per RFC-2547bis. In order to transfer layer-2 VLAN-tagged frames over layer-3 MPLS tunnels, an IP/GRE tunnel is first established between the ONRL-L3 edge routers. The VLAN-tagged frames are first encapsulated over IP/GRE and then encapsulated over MPLS packets. At the egress edge router, the MPLS and IP/GRE headers are extracted and the VLAN-tagged frame is restored back and forwarded to the customer network as shown in Figure 7.14.

7.4.1 Experimental results

Figure 7.15a shows the throughput of TCP flows across the established tunnels described above. The throughput was measured while varying the maximum segment size (MSS) of TCP from 350 to 1200 bytes. As expected, as the MSS increased— *i.e.*, the payload size increased– the total number of packets needed to transfer the same amount of data decreased resulting in an increase of the overall throughput of the TCP flow. However, as total packets size exceeded the MTU of the tunnel, we noticed that packets were fragmented before entering the tunnel and defragmented back at the other end point.

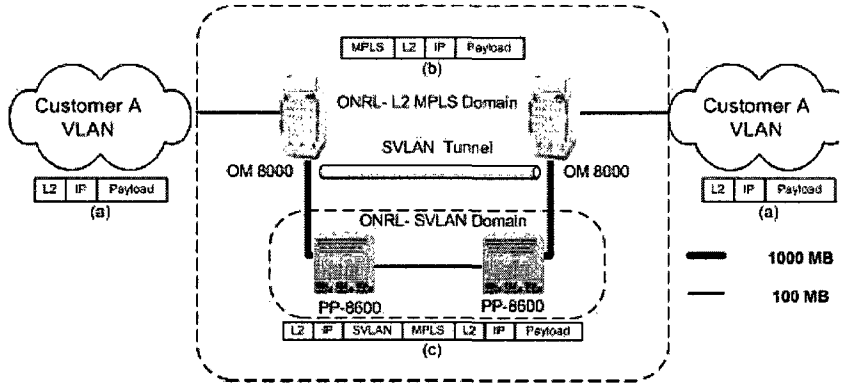


Figure 7.13: Testbed setup realizing layer-2 S-VLAN trunks in a provider network. Packet encapsulation for customer traffic: (a) in VLAN A, (b) across layer-2 MPLS network, and (c) across S-VLAN provider network

This resulted in an increase in the overhead and CPU utilization of the routers.

Using a network protocol analyzer, we compute the approximate overhead bytes per packet by sending a 1000-byte UDP datagrams and comparing it with the size of the overall captured packet (after applying tunneling encapsulation in the core of the network). Figure 7.15b shows the overhead per packet as measured when transported over each of the established tunnel in testcases mentioned earlier. Results showed that MPLS-based techniques out performed other IP-based tunneling in terms of overhead in signaling, reliability, scalability and performance. Furthermore, our tests have shown that increasing the size of the transported payload will increase the throughput as long as the packet size does not exceed the MTU size of the established tunnel. In this case, extra fragmentation and de-fragmentation becomes necessary at the end-points of the tunnel incurring extra processing power on the routers as well as a decrease in the overall throughput due to the increase in the total overhead.

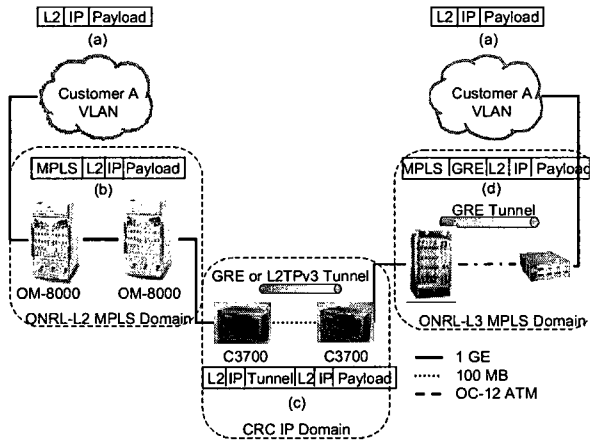
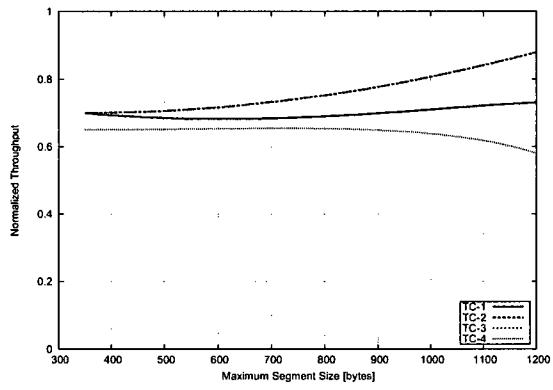
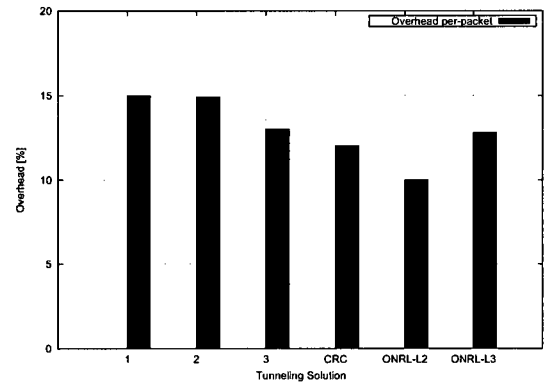


Figure 7.14: Testbed setup realizing layer-2 transparent bridging across heterogeneous networks using GRE and L2TPv3 tunnels. Packet encapsulation for customer traffic: (a) in VLAN A, (b) across layer-2 MPLS network, (c) across GRE/L2TPv3, and (d) across layer-3 MPLS based network



(a) Normalized TCP flow throughput for the implemented tunnel scenario versus maximum segment size.



(b) Percentage of maximum packet overhead per packet for established tunnels for a 1000-byte packet.

Figure 7.15: Experimental results

7.5 Conclusions

This chapter described how to design, deploy and implement provider-based VPNs over SVLANs, and IP/MPLS core networks. The layer-3 VPN approach offers transport of IP traffic only, whereas layer-2 VPN approach allows transparency of carried layer-3 traffic (*e.g.*, IPv4/v6). One of the advantages of layer-3 VPN is that it allows customers to out-source the complexity of routing management from their CE devices to the service provider PE routers. On the other hand, layer-2 solution allows the transport of native layer-2 information (*e.g.*, VLAN, CoS, etc) facilitating the configuration and management overhead on the customer end.

In addition, several protection schemes for ensuring continuity of the service were studied. The QoS schemes using layer 2 802.1p and MPLS/Diffserv were implemented and evaluated on different classes of traffic. The IP/MPLS layer-2 implementation of VPNs offers better scalability than SVLANs that are limited in number to (4096x4096). Also, IP/MPLS protection and restoration techniques proved more robust than layer-2 MLT link bundling protection. The DWDM (1+1) ring protection yielded the best recovery time of approximately 27 ms after failure occurrence.

We also experimented with different tunneling mechanisms that are capable of connecting layer-mismatched networks. We highlighted the advantage of tunneling in the ability to bridge heterogeneous networks using different protocols by implementing experiments on three network domains. Experiments included implementation of the tunneling techniques: 1) layer-2 virtual circuits using GRE tunnels, 2) layer-2 trunks using Ethernet SVLANs, 3) layer-2 tunnels using L2TPv3, and 4) layer-2 transparent bridging across heterogeneous networks using GRE and L2TPv3 tunnels. In the considered implementations, we noticed that GRE tunnel configuration was mainly static with no standard signaling protocol. On the other hand, we found that dynamic signaling of L2TPv3 layer-2 tunnels was possible. Finally, MPLS-based tunneling techniques also proved to be superior in terms of their signaling, reliability, scalability and performance. Our tests have shown that increasing the size of the transported payload will increase the throughput as long as the packet size does not exceed the MTU size of the tunnel. In this case, extra fragmentation and de-fragmentation becomes necessary at the endpoints of the tunnel incurring extra processing power on the routers as well as a decrease in the overall throughput due to the increase in the total overhead.

Chapter 8

Conclusions and Future Research

8.1 Concluding Remarks

In this chapter, we present a review of our research contributions and concluding remarks, and propose some directions for future work.

As the network relies on multiple carriers, issues of survivability across domain boundaries are becoming increasingly important. In a multi-layered and multi-domain network, finding diverse protection paths in a client layer of a domain is a challenging problem. Failure-disjointness at one layer does not necessarily infer physical path disjointness, nor domain-diversity at all lower layers. Hence, computing a pair of failure-diverse inter-domain paths involves finding paths that are SRLG-diverse at all lower layers and domains. Although the concept of SRLG was originally proposed to deal with fiber and conduit cuts, we have shown that it can be extended to include general risks associated with logical links at any layer or domain. We have proposed a novel link aggregation scheme that is suitable for the the multi-domain and multi-layered environment, and which utilizes the concept of SRLG link resource trees to compute failure-diverse paths across vertical and horizontal network-layer partitions.

Within the context of WSON networks, little effort has been invested to date for the provisioning of end-to-end optical lightpaths spanning multiple WSON domains. In general, the problem of RWA for optical lightpaths has been proven to be NP-complete within a single domain. We have proposed a heuristical solution for solving the RWA inter-domain problem that is based on assigning adaptive link costs to aggregated logical links between

BNs of the WSON domain. A PCE in the source domain learns the aggregated links states in remote domains, and can computes loose end-to-end paths for the optical light-path. We have shown that due to security concerns and the competitive nature between different carriers, such approach where domains exchange and maintain a summarized view of the network state is only applicable to the single carrier case. The performance of our proposed algorithm was evaluated by running simulations on a multi-domain WSON network under dynamic traffic load.

We have examined the problem of computing availability-constrained LSP across multiple TE-enabled carrier networks. In this case, solutions that suggest exchange of summarized network state information between domains become inapplicable. Based on the PCE architecture, we have proposed a scheme that extends the existing BRPC mechanisms to allow polling for multi-service information along several alternate domain paths. Based on the polled topology, a decision can be made as to the least cost availability feasible path among the multiple inter-domain paths. We have shown that our proposal outperforms the per-domain ERO expansion scheme with dynamic IP routing, since the latter always prefers domain IP path irrespective of the over-all path availability or cost. In addition, the traditional BRPC assumes the sequence of domains to be traversed by an inter-domain LSP is known a priori. Given this, the traditional BRPC guarantees finding the least cost path along the pre-determined domain path. However, since it only supports optimizing a single cost metric, it does not guarantee that this path is availability feasible, nor it is the least cost feasible inter-domain path.

Within the context of collaborative PCE path computation, when a PCE is not able to compute the full end-to-end path, a decision will have to be made to select and forward the TE path computation request to a downstream PCE— a process also known as PCE selection. We have shown that the downstream PCE selection process is crucial to the amount of overall time taken to compute the end-to-end inter-domain path; in turn, this affects the probability of success of the LSP call setup. Hence, it is useful for PCEs to learn the PCE load state of their peers in order to guide them in choosing the appropriate next hop PCE when multiple alternatives exist. Based on a proposal to advertise the measure of PCE load state, our proposed heuristics distribute the incoming PCE path computation requests among multiple candidate downstream PCE(s). Our scheme has shown superior overall path computation response time and LSP signaling blocking as opposed to arbitrary PCE selection process.

In MPLS networks, P2MP LSPs provide an efficient way to support network services that

require to carry traffic flows from one source to multiple destinations. RSVP-TE based P2MP LSPs allow the ingress node to control the paths taken by the traffic flows, thus enabling bandwidth guarantees to be assured. The ability to compute such paths across multiple domains has also been identified as a key requirement. The P2MP minimum Steiner tree has been proven to be NP-hard, and consequently, is not applicable in the context of realtime tree computation in a dynamic network. Our proposed incremental sub-tree solution includes extensions to the BRPC for the inter-domain P2MP tree computation under bandwidth constraints. One concern that was raised with our approach is the amount of signaling needed for the P2MP inter-domain tree computation, specifically, with large number of destination leaves. We have proposed a way to reduce this by clustering together requests for destinations leaves of the same P2MP LSP residing in the same destination domain.

Through implementation on a general-purpose experimental testbed, we have evaluated and compared several inter-testbed networking services, technologies, and interoperability mechanisms. Several protection techniques have been tested for ensuring the continuity of the VPN services, including MPLS TE FRR, layer-2 link bundling, as well as 1+1 DWDM optical protection. We have experimented with several tunneling techniques for connecting layer-mismatched networks, including layer-2 virtual circuits using GRE tunnels, trunks over SVLANs, and bridging using GRE and L2TPv3 tunnels. We have noticed that GRE tunnels were mostly statically configured with no standard signaling protocol. On the other hand, we found that dynamic signaling of L2TPv3 cross-connects was possible, and was useful in changing the attributes of the connection dynamically. We have also concluded that MPLS-based tunneling was superior in terms of its signaling, scalability and performance.

8.2 Future Research

Many extensions to the proposed schemes that we have presented in this thesis can be envisioned.

The rise of P2MP communication has revived interest in P2MP transmission services over the global Internet. We have found the area of protection and restoration for inter-domain P2MP trees to be a challenging topic that has gotten little attention so far. In particular, the problem of computing diverse inter-domain trees for TE P2MP LSPs

presents a significant challenge due to the complexity of the computations and number of constraints described earlier. Determining such disjoint inter-domain P2MP trees can add considerably to this complexity. Disjoint paths are required for end-to-end protection services and sometimes for load balancing. These may require to be fully disjoint, link disjoint, or best-effort disjoint. Another key area that is also challenging is the reoptimization of existing inter-domain P2MP trees, since small modifications to a P2MP tree (such as adding or removing a single destination leaf) can lead to completely different optimal resulting tree. We believe that the PCE architecture fits well in this context, and can present viable potential solution for these problems.

This thesis also considered the problem of availability-constrained inter-domain P2P LSP path computation. A natural extension to this would be the ability to compute availability-constrained P2MP trees where the availability for any sub-LSP path of the tree is guaranteed below a certain bound. This problem is known to be NP-complete within the single domain network. We still believe that efficient heuristics can be developed to quickly and efficiently find feasible solutions to this problem.

We have investigated several techniques to load-balance PCE path computation requests among several PCE chains. However, in this thesis, we have assumed that all path computation requests are equally computationally intensive. A future extension to this work would be to consider the case where some requests are more computationally demanding, multiple requests are bundled together, and when requests can be delegated to other PCE(s) (*e.g.*, due to lack-of-resources) before being processing.

Finally, within the context of inter-domain WSON meshed networks, computing paths with optical transport wavelength continuity constraints is still a challenge. A natural extension to our work would be to consider the problem of provisioning P2MP optical lightpaths that span multiple WSON domains. With the rise of GMPLS and layer-1 VPNs, there are several providers that have started to deploy optical VPNs and provide P2MP VPN connectivity to multiple customer sites. We believe, in this case, P2MP communication is an appealing solution for optical networks as it can reduce the amount of used network resources by encouraging wavelength channel re-use on links of P2MP tree— as opposed to signalling individual P2P lightpaths to each destination leaf. Specifically, within the inter-domain context, P2MP tree computation for WSON networks is has received little attention to date.

Bibliography

[Adam 07] D. Adami, S. Giordano, F. Mustacchio, and M. Pagano. "Design and development of a DSTE experimental testbed with Point-to-Multipoint LSP support". *In proceedings of 3rd International Conference on Next Generation Internet Networks (EuroNGI '07)*, pp. 14–20, May 2007.

[Agar 03] S. Agarwal, C.-N. Chuah, and R. Katz. "OPCA: robust interdomain policy routing and traffic control". *In proceedings of the IEEE Conference on Open Architectures and Network Programming, 2003*, pp. 55–64, April 2003.

[Agga 07] R. Aggarwal, D. Papadimitriou, and S. Yasukawa. "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)". RFC 4875 (Proposed Standard), May 2007.

[Agga 08] R. Aggarwal, Y. Kamite, L. Fang, Y. Rekhter, and I. Property. "Multicast in VPLS". Internet-Draft draft-ietf-l2vpn-vpls-mcast-04, Internet Engineering Task Force, June 2008. Work in progress.

[Ahme 96] M. Ahmed and J. H. Rus. "Technical Committee Private Network-Network Interface Specification Version 1.0". 1996.

[Alan 04] W. Alanqar and A. Jukan. "Extending end-to-end optical service provisioning and restoration in carrier networks: opportunities, issues, and challenges". *IEEE Communications Magazine*, Vol. 42, No. 1, pp. 52–60, Jan 2004.

[Ali 09] Z. Ali and T. Saad. "BRPC Extensions for Point-to-Multipoint Path Computation". Internet-Draft draft-ali-pce-brpc-p2mp-ext-00.txt, Internet Engineering Task Force, March 2009. Work in progress.

[Arte 07] A. Arteta, B. Barán, and D. Pinto. "Routing and wavelength assignment over WDM optical networks: a comparison between MOACOs and classical approaches". *In proceedings of the 4th international IFIP/ACM Latin American conference on Networking (LANC '07)*, pp. 53–63, 2007.

[Asla 07a] F. Aslam, Z. Uzmi, and A. Farrel. "Interdomain path computation: Challenges and Solutions for Label Switched Networks". *IEEE Communications Magazine*, Vol. 45, No. 10, pp. 94–101, October 2007.

[Asla 07b] F. Aslam, Z. Uzmi, A. Farrel, and M. Pioro. "Inter-Domain Path Computation using Improved Crankback Signaling in Label Switched Networks". *In proceedings of the IEEE International Conference on Communications (ICC '07)*, pp. 2023–2029, June 2007.

[ATT 09] ATT. "ATT Global IP Backbone Networks". Online, Jan. 2009. Available from http://www.corp.att.com/peering/.

[Awdu 01] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow. "RSVP-TE: Extensions to RSVP for LSP Tunnels". RFC 3209 (Proposed Standard), Dec. 2001. Updated by RFCs 3936, 4420, 4874, 5151.

[Awer 98] B. Awerbuch, Y. Du, and B. Khan. "Routing through networks with hierarchical topology aggregation". *IOS Journal of High Speed Networks*, Vol. 7, No. 1, pp. 57–73, 1998.

[Ayya 08] A. Ayyangar, K. Kompella, J. Vasseur, and A. Farrel. "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)". RFC 5150, Internet Engineering Task Force, Feb. 2008.

[Ball 97] A. Ballardie. "Core Based Trees (CBT version 2) Multicast Routing – Protocol Specification –". RFC 2189 (Historic), Sep. 1997.

[Balo 08] S. Balon and G. Leduc. "Combined intra- and inter-domain traffic engineering using hot-potato aware link weights optimization". *In proceedings of the 2008 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '08)*, pp. 441–442, 2008.

[Bane 01] A. Banerjee, J. Drake, J. Lang, B. Turner, K. Kompella, and Y. Rekhter. "Generalized multiprotocol label switching: an overview of routing and man-

agement enhancements". *IEEE Communications Magazine*, Vol. 39, No. 1, pp. 144–150, Jan 2001.

[Barr 97] R. Barry and S. Subramaniam. "The MAX SUM wavelength assignment algorithm for WDM ring networks". *In proceedings of the Conference on Optical Fiber Communication (OFC '97)*, pp. 121–122, Feb 1997.

[Bern 08] G. Bernstein, Y. Lee, and W. Imajuku. "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks (WSON)". Internet-Draft draft-ietf-ccamp-wavelength-switched-framework-00, Internet Engineering Task Force, May 2008. Work in progress.

[Bita 08] N. Bitar, R. Zhang, and K. Kumaki. "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)". RFC 5376 (Informational), Nov. 2008.

[Bouc 05] M. Boucadair. "QoS-Enhanced Border Gateway Protocol". Internet-Draft draft-boucadair-qos-bgp-spec-01, Internet Engineering Task Force, July 2005. Work in progress.

[BS A 01] M. W. B.S. Arnaud and J. Coulter. "BGP Optical Switches and Lightpath Route Arbiter". *Optical Networks Magazine*, Vol. 2, No. 2, pp. 73–81, mar 2001.

[Capo 09] A. Capone, J. Elias, and F. Martignon. "Routing and resource optimization in service overlay networks". *Elsevier Journal of Computer Networks*, Vol. 53, No. 2, pp. 180–190, 2009.

[Chai 08a] M. Chaitou and J.-L. Le Roux. "Multi-point to multi-point traffic engineering". *In proceedings of IEEE Symposium on Computers and Communications (ISCC '08)*, pp. 1047–1055, July 2008.

[Chai 08b] M. Chaitou and J.-L. Le Roux. "On Optimizing Leaf Initiated Point to Multi Point Trees in MPLS". *In proceedings of the 4th International Conference on Networking and Services (ICNS '08)*, pp. 53–60, March 2008.

[Chai 08c] M. Chaitou and J.-L. L. Roux. "On Optimizing Leaf Initiated Point to Multi Point Trees in MPLS". *In proceedings of the 4th IEEE International Conference on Networking and Services (ICNS '08)*, pp. 53–60, 2008.

[Cham 05] S. Chamberland, D. O. Khyda, and S. Pierre. "Joint routing and wavelength assignment in wavelength division multiplexing networks for permanent and reliable paths". *Elsevier Journal of Computer Operations and Research*, Vol. 32, No. 5, pp. 1073–1087, 2005.

[Chen 95] S. wing Cheng and C. keung Tang. "A Fast Algorithm for Computing Optimal Rectilinear Steiner Trees for Extremal Point Sets". *In proceedings of the 6th Annual International Symposium on Algorithms and Computation (ISAAC '95)*, pp. 322–331, 1995.

[Cink 07a] T. Cinkler, P. Hegyi, G. Geleji, and J. Szigeti. "Fairness Issues of AMLTE: Adaptive Multi-Layer Traffic Engineering with Grooming". *In proceedings of the 9th International Conference on Transparent Optical Networks (ICTON '07)*, Vol. 1, No. , pp. 63–66, July 2007.

[Cink 07b] T. Cinkler, J. Szigeti, and L. Gyarmati. "Multi-Domain Resilience: Can I Share Protection Resources with my Competitors?". *In proceedings of the 9th IEEE International Conference on Transparent Optical Networks (ICTON '07)*, Vol. 3, pp. 138–141, July 2007.

[Cnod 03] S. D. Cnodder and C. Pelsser. "Protection for inter-AS MPLS tunnels". September 2003. Work in progress, draft-decnodder-mpls-interas-protection-01.txt.

[Cugi 07] F. Cugini, A. Giorgetti, N. Andriolli, I. Paolucci, L. Valcarenghi, and P. Castoldi. "Multiple Path Computation Element (PCE) Cooperation for Multilayer Traffic Engineering". *In proceedings of the Optical Fiber Communication and the National Fiber Optic Engineers Conference (OFC/NFOEC '07)*, pp. 1–3, March 2007.

[Dasg 07] S. Dasgupta, J. de Oliveira, and J.-P. Vasseur. "Path-Computation-Element-Based Architecture for Interdomain MPLS/GMPLS Traffic Engineering: Overview and Performance". *IEEE Network Journal*, Vol. 21, No. 4, pp. 38–45, July-August 2007.

[Dasg 08] S. Dasgupta, J. Oliveira, and J. Vasseur. "Performance Analysis of Inter-Domain Path Computation Methodologies". Internet-Draft draft-dasgupta-ccamp-path-comp-analysis-02.txt,, Internet Engineering Task Force, July 2008. Work in progress.

[Deme 99] P. Demeester and T. Wu. "Survivable communications networks". *IEEE Communication Magazine,*, Vol. 37, No. 8, p. , August 1999.

[Dono 04] Y. Donoso, R. Fabregat, and J. Marzo. "Multiobjective optimization model and heuristic algorithm for dynamic multicast routing". *In proceedings of the 11th International Telecommunications Network Strategy and Planning Symposium (Networks 2004)*, pp. 423–428, June 2004.

[Douv 08] R. Douville, J.-L. Le Roux, J.-L. Rougier, and S. Secci. "A service plane over the PCE architecture for automatic multidomain connection-oriented services". *IEEE Communications Magazine*, Vol. 46, No. 6, pp. 94–102, June 2008.

[Elsa 05] Elsayed and K.M.F. "A framework for end-to-end deterministic-delay service provisioning in multiservice packet networks". *IEEE Transactions on Multimedia*, Vol. 7, No. 3, pp. 563–571, June 2005.

[Fang 05] L. Fang, N. Bita, J.-L. Le Roux, and J. Miles. "Interprovider IP-MPLS services: requirements, implementations, and challenges". *IEEE Communications Magazine*, Vol. 43, No. 6, pp. 119–128, June 2005.

[Fark 05] A. Farkas, J. Szigeti, and T. Cinkler. "P-cycle based protection schemes for multi-domain networks". *In proceedings of the 5th IEEE International Workshop on Design of Reliable Communication Networks (DRCN '05)*, pp. 8–16, Oct. 2005.

[Farr 06a] A. Farrel, J.-P. Vasseur, and J. Ash. "A Path Computation Element (PCE)-Based Architecture". RFC 4655, Internet Engineering Task Force, Aug. 2006.

[Farr 06b] A. Farrel, J.-P. Vasseur, and A. Ayyangar. "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering". RFC 4726 (Informational), Nov. 2006.

[Fenn 03] B. Fenner and D. Meyer. "Multicast Source Discovery Protocol (MSDP)". RFC 3618 (Experimental), Oct. 2003.

[Ferr 06] C. E. Ferreira and F. M. de Oliveira Filho. "Some formulations for the group steiner tree problem". *Elsevier Journal of Discrete Application Mathematics*, Vol. 154, No. 13, pp. 1877–1884, 2006.

[Fuma 00] A. Fumagalli and L. Valcarenghi. "IP restoration vs. WDM protection: is there an optimal choice?". *IEEE Journal on Networks*, Vol. 14, No. 6, pp. 34–41, Nov/Dec 2000.

[Gao 08] Z. Gao and H. Naser. "End-to-end shared restoration algorithms in multi-domain mesh networks". *In proceedings of the IEEE Symposium on Computers and Communications (ISCC '08)*, pp. 411–416, July 2008.

[Gare 79] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness.* W. H. Freeman and Company, New York, NY, USA, 1979.

[Ghan 08] N. Ghani, Q. Liu, D. Benhaddou, N. Rao, and T. Lehman. "Control plane design in multidomain/multilayer optical networks". *IEEE Communications Magazine*, Vol. 46, No. 6, pp. 78–87, June 2008.

[Ghan 99] A. Ghanwani, B. Jamoussi, D. Fedyk, P. Ashwood-Smith, L. Li, and N. Feldman. "Traffic Engineering Standards in IP Networks using MPLS". *IEEE Communications Magazine*, Vol. 32, No. 5, pp. 1073–1087, 1999.

[Gram 07] E. Grampin, A. Castro, M. German, F. Rodriguez, G. Tejera, and M. Sanguinetti. "A PCE-based Connectivity Provisioning Management Framework". *In proceedings of the Conference of Network Operations and Management Symposium (LANOMS '07)*, pp. 76–83, Sept. 2007.

[Grou 09] I. W. Group. "Common Control and Measurement Plane". 2009. This is an electronic document. Date retrieved: February 2, 2009.

[Guo 98a] L. Guo and I. Matta. "On state aggregation for scalable QoS routing". *In proceedings of the IEEE ATM Workshop, 1998*, pp. 306–314, May 1998.

[Guo 98b] L. Guo and I. Matta. "On State Aggregation for Scalable QoS Routing". *In proceedings of the ATM Workshop'98*, Vol. 6, pp. 306–314, 1998.

[Hank 94] S. Hanks, T. Li, D. Farinacci, and P. Traina. "Generic Routing Encapsulation (GRE)". RFC 1701 (Informational), Oct. 1994.

[HAYA 08] R. HAYASHI, E. OKI, and K. SHIOMOTO. "Inter-Domain Redundancy Path Computation Methods Based on PCE". *IEICE Transactions on Communications*, Vol. E91-B, No. 10, pp. 3185–3193, March 2008.

[Ho 02] P.-H. Ho and H. Mouftah. "An approach for enhancing fixed alternate routing in dynamic wavelength-routed WDM networks". *In proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '02)*, Vol. 3, pp. 2792–2797, Nov. 2002.

[Ho 04] K.-H. Ho, N. Wang, P. Trimintzios, G. Pavlou, and M. Howarth. "On egress router selection for inter-domain traffic with bandwidth guarantees". *In proceedings of the Workshop on High Performance Switching and Routing (HPSR '04)*, pp. 337–342, 2004.

[Howa 05] M. Howarth, P. Flegkas, G. Pavlou, N. Wang, P. Trimintzios, D. Griffin, J. Griem, M. Boucadair, P. Morand, A. Asgari, and P. Georgatsos. "Provisioning for interdomain quality of service: the MESCAL approach". *IEEE Communications Magazine*, Vol. 43, No. 6, pp. 129–137, June 2005.

[Huan 04a] C. Huang and D. Messier. "A Fast and Scalable Inter-Domain MPLS Protection Mechanism". *Elsevier Journal of Communications and Networks*, Vol. 6, No. 1, pp. 60–67, March 2004.

[Huan 04b] Y. Huang, J. Heritage, B. Mukherjee, and W. Wen. "Availability-guaranteed service provisioning with shared-path protection in optical WDM networks". *In proceedings of the Conference on Optical Fiber Communication (OFC '04)*, Vol. 1, pp. –, Feb. 2004.

[Huan 04c] Y. Huang, W. Wen, J. Zhang, J. Heritage, and B. Mukherjee. "A new link-state availability model for reliable protection in optical WDM networks". *In proceedings of the IEEE International Conference on Communications (ICC '04)*, Vol. 3, pp. 1649–1653, June 2004.

[Hwan 92] F. Hwang, D. Richards, and P. Winter. *The Steiner tree problem.* Vol. 53 of *Ann. Discrete Math.*, North-Holland, Ed., 1992.

[Iova 04] P. Iovanna, M. Settembre, R. Sabella, G. Conte, and L. Valentini. "Performance analysis of a traffic engineering solution for multilayer networks based on the GMPLS paradigm". *IEEE Journal on Selected Areas in Communications*, Vol. 22, No. 9, pp. 1731–1740, Nov. 2004.

[Iwat 98] A. Iwata, H. Suzuki, R. Izmailov, and B. Sengupta. "QOS aggregation algorithms in hierarchical ATM networks". *In proceedings of the IEEE Interna-*

tional Conference on Communications (ICC '98), Vol. 1, No. , pp. 243–248 vol.1, Jun 1998.

[Jaff 84] J. M. Jaffe. "Algorithms for finding paths with multiple constraints". *Wiley Periodicals Journal on Networks*, Vol. 14, No. 1, pp. 95–116, 1984.

[Jutt 01] A. Juttner, B. Szviatovski, I. Mecs, and Z. Rajko. "Lagrange relaxation based method for the QoS routing problem". *In proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'01)*, Vol. 2, pp. 859–868, Aug. 2001.

[Kara 98] E. Karasan and E. Ayanoglu. "Effects of wavelength routing and selection algorithms on wavelength conversion gain in WDM optical networks". *IEEE/ACM Transactions on Networking*, Vol. 6, No. 2, pp. 186–196, 1998.

[Kim 99] W. Kim and Y. Park. "DCBT: an efficient multicast architecture for wide scale and large group multimedia communications". *In proceedings of the IEEE International Conference on Communications (ICC '99)*, Vol. 1, No. , pp. 665–670 vol.1, 1999.

[King 08] D. King, Y. Lee, H. Xu, and A. Farrel. "Path Computation Architectures Overview in Multi-Domain Optical Networks Based on ITU-T ASON and IETF PCE". *In proceedings of the IEEE/IFIP Network Operations and Management Symposium Workshops (NOMS '08)*, pp. 219–226, April 2008.

[Komp 05] K. Kompella and Y. Rekhter. "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)". RFC 4206, Internet Engineering Task Force, Oct. 2005.

[Kork 03] T. Korkmaz and M. Krunz. "Bandwidth-delay constrained path selection under inaccurate state information,". *IEEE/ACM Transactions on Networking*, Vol. 11, No. 3, pp. 384–398, June 2003.

[Kuma 08] K. Kumaki, I. Nakagawa, K. Nagami, T. Ogishi, and S. Ano. "Design and evaluation of a P2MP MPLS-based hierarchical service management system". *In proceedings of the IEEE Symposium on Computers and Communications (ISCC '08)*, pp. 794–799, July 2008.

[Lai 02] W. Lai and D. McDysan. "Network Hierarchy and Multilayer Survivability". RFC 3386 (Informational), Nov. 2002.

[Lang 05] J. Lang. "Link Management Protocol (LMP)". RFC 4204 (Proposed Standard), Oct. 2005.

[Larr 05] D. Larrabeiti, R. Romeral, I. Soto, M. Uruena, T. Cinkler, J. Szigeti, and J. Tapolcai. "Multi-domain issues of resilience". *In proceedings of 7th International Conference Transparent Optical Networks (ICTON '05)*, Vol. 1, pp. 375–380, July 2005.

[Lass 07] M. Lasserre and V. Kompella. "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling". RFC 4762 (Proposed Standard), Jan. 2007.

[Lau 05] J. Lau, M. Townsley, and I. Goyret. "Layer Two Tunneling Protocol - Version 3 (L2TPv3)". RFC 3931 (Proposed Standard), March 2005.

[Lee 03] S. S. Lee, S. Das, G. Pau, and M. Gerla. "A hierarchical multipath approach to QoS routing: Performance and evaluation". *In proceedings of International Conference on Communications, ICC'03*, pp. 11–15, 2003.

[Lee 95] W. C. Lee, M. G. Hluchyi, and P. A. Humblet. "Routing Subject to Quality of Service Constraints Integrated Communication Networks". *IEEE Network Magazine*, Vol. 9, No. 4, pp. 46–55, July/August 1995.

[Liu 05] G. Liu, C. Ji, and V. Chan. "On the scalability of network management information for inter-domain light-path assessment". *IEEE/ACM Transactions on Networking*, Vol. 13, No. 1, pp. 160–172, Feb. 2005.

[Liu 06a] Q. Liu, M. A. Kök, N. Ghani, and A. Gumaste. "Hierarchical Inter-Domain Routing and Lightpath Provisioning in Optical Networks". *OSA Journal of Optical Networking*, Vol. 5, No. 10, pp. 764–774, Oct. 2006.

[Liu 06b] Q. Liu, M. A. Kök, N. Ghani, and A. Gumaste. "Hierarchical routing in multi-domain optical networks". *Elsevier Journal of Computer Communications*, Vol. 30, No. 1, pp. 122–11, December 2006.

[Liu 07a] Q. Liu, N. Ghani, and T. Frangieh. "Topology abstraction schemes in multi-domain full wavelength conversion DWDM networks". *In proceedings of the*

International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET '07), pp. 1–6, Nov. 2007.

[Liu 07b] Q. Liu, N. Ghani, N. S. V. Rao, A. Gumaste, and M. L. Garcia. "Distributed inter-domain lightpath provisioning in the presence of wavelength conversion". *Elsevier Journal of Computer Communications*, Vol. 30, No. 18, pp. 3662–3675, 2007.

[Liu 07c] Q. Liu, N. Ghani, N. Rao, and T. Lehman. "Multi-Domain Multi-Granularity Service Provisioning in Hybrid DWDM/SONET Networks". *In proceedings of the IEEE Conference on High-Speed Networks (HSN '07)*, pp. 26–30, May 2007.

[Loja 05] K. Loja, J. Szigeti, and T. Cinkler. "Inter-domain routing in multiprovider optical networks: game theory and simulations". *In proceedings of the IEEE Conference on Next Generation Internet Networks (NGI '05)*, pp. 157–164, April 2005.

[Lui 00] K.-S. Lui, K. Nahrstedt, and S. Chen. "Hierarchical QoS routing in delay-bandwidth sensitive networks". *In proceedings of the Annual IEEE Conference on Local Computer Networks*, p. 579, 2000.

[Maie 08] G. Maier, C. Busca, and A. Pattavina. "Multi-domain routing techniques with topology aggregation in ASON networks". *In proceedings of the International Conference on Optical Network Design and Modeling (ONDM '08)*, pp. 1–6, March 2008.

[Mann 04] E. Mannie. "Generalized Multi-Protocol Label Switching (GMPLS) Architecture". RFC 3945 (Proposed Standard), Oct. 2004.

[Mats 06] S. Matsushima, T. Murakami, and K. Nagami. "BGP Extension for MPLS P2MP-LSP". *IEICE Transactions on Information Systems*, Vol. E89-D, No. 1, pp. 211–218, 2006.

[Mats 07] H. Matsuura, N. Morita, I. Nakajima, and K. Takami. "Hierarchically Distributed PCE for Flexible Multicast Traffic Engineering". *In proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 2439–2444, Nov. 2007.

[Mats 08] R. Matsumura, M. Inoue, M. Tsujino, and M. Iwashita. "Evaluation of MPLS P2MP Distribution Tree Algorithms". *In proceedings of the 13th International Conference on Telecommunications Network Strategy and Planning Symposium (Networks 2008)*, pp. 1–15, 28 2008-Oct. 2 2008.

[Mesk 06] D. Mesko, G. Viola, and T. Cinkler. "A Hierarchical and a Non-Hierarchical European Multi-Domain Reference network: Routing and Protection". *In proceedings of the 12th International Conference on Telecommunications Network Strategy and Planning Symposium (Networks 2006)*, pp. 1–5, Nov. 2006.

[Moha 05] A. Mohamad and S. Asano. "Deployment Strategies on PCE-based Inter-AS LSP Path Computation". *In proceedings of the 9th IEEE International Multitopic Conference (INMIC '05)*, pp. 1–6, Dec 2005.

[MRai 05] D. M'Raihi, M. Bellare, F. Hoornaert, D. Naccache, and O. Ranen. "HOTP: An HMAC-Based One-Time Password Algorithm". RFC 4226 (Informational), Dec. 2005.

[Noro 94] J. Noronha, C.A. and F. Tobagi. "Optimum routing of multicast streams". *In proceedings of the 13th IEEE Conference on Networking for Global Communications (INFOCOM)*, Vol. 2, pp. 865–873, Jun 1994.

[Oki 08] E. Oki, J. L. Roux, and A. Farrel. "Extensions to the Path Computation Element communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering". Internet-Draft draft-ietf-pce-inter-layer-ext-01, Internet Engineering Task Force, June 2008. Work in progress.

[Okum 01] I. T. Okumus, J. Hwang, H. A. Mantar, and S. J. Chapin. "Inter-domain LSP setup using bandwidth management points". *In proceedings of the IEEE Global Telecommunications Conference (Globecom '01)*, pp. 25–29, 2001.

[Oliv 05] C. A. S. Oliveira and P. M. Pardalos. "A survey of combinatorial optimization problems in multicast routing". *Elsevier Journal of Computer Operations Research*, Vol. 32, No. 8, pp. 1953–1981, 2005.

[Otan 08] T. Otani and K. Ogaki. "Requirements for GMPLS applications of PCE". Internet-Draft draft-otani-pce-gmpls-aps-req-02, Internet Engineering Task Force, July 2008. Work in progress.

[Pand 06] Z. Pandi, M. Tacca, A. Fumagalli, and L. Wosinska. "Dynamic provisioning of availability-constrained optical circuits in the presence of optical node failures". *IEEE Journal of Lightwave Technology*, Vol. 24, No. 9, pp. 3268–3279, Sept. 2006.

[Papa 05] D. Papadimitriou and D. Verchere. "GMPLS user-network interface in support of end-to-end rerouting". *IEEE Communications Magazine*, Vol. 43, No. 7, pp. 35–43, July 2005.

[Papa 08] D. Papadimitriou, M. Vigoureux, K. Shiomoto, D. Brungard, J. Roux, E. Oki, I. Inoue, E. Dotaro, and G. Grammel. "Generalized Multi-Protocol Label Switching (GMPLS) Protocol Extensions for Multi-Layer and Multi-Region Networks (MLN/MRN)". Internet-Draft draft-ietf-ccamp-gmpls-mln-extensions-02, Internet Engineering Task Force, July 2008. Work in progress.

[Pels 03] C. Pelsser and O. Bonaventure. "Extending RSVP-TE to support inter-AS LSPs". *In proceedings of the Workshop on High Performance Switching and Routing (HPSR '03)*, pp. 79–84, June 2003.

[Pels 06] C. Pelsser and O. Bonaventure. "Path Selection Techniques to Establish Constrained Interdomain MPLS LSPs". *In proceedings of 5th International IFIP-TC6 Networking Conference (Networking 2006)*, May 15-19th 2006.

[Pick 06] M. Pickavet, P. Demeester, D. Colle, D. Staessens, B. Puype, L. Depre, and I. Lievens. "Recovery in multilayer optical networks". *IEEE Journal of Lightwave Technology*, Vol. 24, No. 1, pp. 122–134, Jan. 2006.

[Pong 04] P. Pongpaibool and H. S. Kim. "Providing end-to-end service level agreements across multiple ISP networks". *Elsevier Journal of Computer Networks*, Vol. 46, No. 1, pp. 3–18, 2004.

[Prin 05] R. Prinz, A. Autenrieth, and D. Schupke. "Dual failure protection in multi-layer networks based on overlay or augmented model". *In proceedings of the 5th International Workshop on Design of Reliable Communication Networks (DRCN 2005)*, pp. 8–12, Oct. 2005.

[Puyp 04] B. Puype, Q. Yan, S. De Maesschalck, D. Colle, M. Pickavet, and P. Demeester. "Optical cost metrics in multi-layer traffic engineering for IP-over-

optical networks". *In proceedings of the 6th IEEE International Conference on Transparent Optical Networks (ICTON '04)*, Vol. 1, pp. 75–80, July 2004.

[Puyp 05] B. Puype, J.-P. Vasseur, A. Groebbens, S. De Maesschalck, D. Colle, I. Lievens, M. Pickavet, and P. Demeester. "Benefits of GMPLS for multilayer recovery". *IEEE Communications Magazine*, Vol. 43, No. 7, pp. 51–59, July 2005.

[Qin 03] Y. Qin, L. Mason, and K. Jia. "Study on a joint multiple layer restoration scheme for IP over WDM networks". *IEEE Network Journal*, Vol. 17, No. 2, pp. 43–48, Mar/Apr 2003.

[Quoi 05a] B. Quoitin and O. Bonaventure. "A cooperative approach to interdomain traffic engineering". *In proceedings of the Network of Excellence (NoE) Next Generation Internet Networks (NGI '05)*, pp. 450–457, April 2005.

[Quoi 05b] B. Quoitin, C. Pelsser, O. Bonaventure, and S. Uhlig. "A performance evaluation of BGP-based traffic engineering". *Wiley International Journal on Networking Management*, Vol. 15, No. 3, pp. 177–191, 2005.

[Raja 04] B. Rajagopalan, J. Luciani, and D. Awduche. "IP over Optical Networks: A Framework". RFC 3717 (Informational), March 2004.

[Rama 03] S. Ramamurthy, L. Sahasrabuddhe, and B. Mukherjee. "Survivable WDM mesh networks". *IEEE Journal of Lightwave Technology*, Vol. 21, No. 4, pp. 870–883, April 2003.

[Rama 99] S. Ramamurthy and B. Mukherjee. "Survivable WDM mesh networks. II. Restoration". *In proceedings of the IEEE International Conference on Communications (ICC '99)*, Vol. 3, pp. 2023–2030, Jun. 1999.

[Rekh 01] Y. Rekhter and E. Rosen. "Carrying Label Information in BGP-4". RFC 3107 (Proposed Standard), May 2001.

[Rekh 94] Y. Rekhter and T. Li. "A Border Gateway Protocol 4 (BGP-4)". RFC 1654 (Proposed Standard), July 1994. Obsoleted by RFC 1771.

[Ricc 05] F. Ricciato, U. Monaco, and D. Ali. "Distributed schemes for diverse path computation in multidomain MPLS networks". *IEEE Communications Magazine*, Vol. 43, No. 6, pp. 138–146, June 2005.

[Rome 06] R. Romeral and D. Larrabeiti. "Combining Border Router Policies for Disjoint LSP computation". *In proceedings of the 5th Workshop of GMPLS Networks (WGN5)*, March 2006.

[Saad 04a] T. Saad and H. T. Mouftah. "Constraint-based Routing Across Multi-domain Optical WDM Networks". *In proceedings IEEE Canadian Conference on Electrical and Computer Engineering (CCECE '04)*, pp. 2065–2068, May 2004.

[Saad 04b] T. Saad and H. T. Mouftah. "Inter-Domain Wavelength Routing in Optical WDM Networks". *In proceedings of the 11th International Telecommunications Network Strategy and Planning Symposium (Networks '04)*, pp. 391–396, June 2004.

[Saad 04c] T. Saad, H. Naser, and H. T. Mouftah. "A Hierarchical SRLG Organization Scheme for Multi-domain Multi-layered Transport Networks". *In proceedings of the 7th International Symposium on Communications Interworking (Interworking '04)*, pp. 301–338, September 2004.

[Saad 04d] T. Saad and H. T. Mouftah. "A Multidomain Differentiated Resilience Scheme Using SRLG Aggregation in Multi-layered Transport Networks". CITO Innovators Showcase, Meet Tomorrow's Technology Leader's Award 2004, November 2004.

[Saad 05a] T. Saad, B. Alawieh, S. Gulder, and H. T. Mouftah. "Tunneling Techniques for End-to-End VPNs: Generic Deployment in an Optical Testbed Environment". *In proceedings of the 2nd International Conference on Broadband Networks, (BroadNets '05)*, pp. 924–930, October 2005.

[Saad 05b] T. Saad, B. Alawieh, and H. T. Mouftah. "Inter-VLAN VPNs Over a High Performance Optical Testbed". *In proceedings of the 1st IEEE International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM '05)*, pp. 221–229, February 2005.

[Saad 05c] T. Saad and H. T. Mouftah. "End-to-End Inter-domain Routing and Signaling for Wavelength Routed WDM Optical Networks". *In proceedings of the Optical Fiber Communication and the National Fiber Optic Engineers Conference (NFOEC/OFC '05)*, pp. 293–298, March 2005.

[Saad 06] T. Saad, B. Alawieh, S. Gulder, and H. T. Mouftah. "Tunneling Techniques for End-to-End VPNs: Generic Deployment in an Optical Testbed Environment". *IEEE Communications Magazine*, Vol. 44, No. 55, pp. 124–132, 2006.

[Saad 07] T. Saad, J. Israr, S. Sivabalan, and H. Mouftah. "An Evaluation for PCE Selection Schemes for Inter-Domain Path Computation". *In proceedings of the 9th IEEE International Conference on Transparent Optical Networks (ICTON '07)*, Vol. 3, No. , pp. 187–187, July 2007.

[Sabe 03] R. Sabella, M. Settembre, G. Oriolo, F. Razza, F. Ferlito, and G. Conte. "A multilayer solution for path provisioning in new-generation optical/MPLS networks". *IEEE Journal of Lightwave Technology*, Vol. 21, No. 5, pp. 1141–1155, May 2003.

[Sanc 05] S. Sanchez-Lopez, X. Masip-Bruin, E. Marin-Tordera, J. Sole-Pareta, and J. Domingo-Pascual. "A hierarchical routing approach for GMPLS based control plane for ASON". *In proceedings of the IEEE International Conference on Communications (ICC '05)*, Vol. 3, pp. 1683–1687, May 2005.

[Secc 08a] S. Secci, J.-L. Rougier, and A. Pattavina. "AS Tree Selection for Inter-Domain Multipoint MPLS Tunnels". *In proceedings of the IEEE International Conference on Communications (ICC '08)*, pp. 5863–5868, May 2008.

[Secc 08b] S. Secci, J.-L. Rougier, and A. Pattavina. "On the selection of optimal diverse AS-paths for inter-domain IP/(G)MPLS tunnel provisioning". *In proceedings of the 4th IEEE Telecommunication Networking Workshop on QoS in Multiservice IP Networks (IT-NEWS '08)*, pp. 2350–241, Feb. 2008.

[Shio 08] K. Shiomoto, D. Papadimitriou, J. L. Roux, M. Vigoureux, and D. Brungard. "Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)". RFC 5212 (Informational), July 2008.

[Siga 03] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos. "Power laws and the AS-level Internet topology". *IEEE/ACM Transactions on Networking*, Vol. 11, No. 4, pp. 514–524, Aug. 2003.

[Sort 09] D. D. Sorte, M. Femminella, and G. Reali. "QoS-enabled multicast for delivering live events in a Digital Cinema scenario". *Elsevier Journal on Network Computing Applications*, Vol. 32, No. 1, pp. 314–344, 2009.

[Spat 00] J. Spath. "Dynamic routing and resource allocation in WDM transport network". *Elsevier Journal of Computer Networks*, Vol. 32, No. 5, pp. 519–538, 2000.

[Spri 07] A. Sprintson, M. Yannuzzi, A. Orda, and X. Masip-Bruin. "Reliable Routing with QoS Guarantees for Multi-Domain IP/MPLS Networks". *In proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 1820–1828, May 2007.

[Subr 97] S. Subramaniam and R. Barry. "Wavelength assignment in fixed routing WDM networks". *In proceedings of the IEEE International Conference on Communications (ICC '97)*, Vol. 1, pp. 406–410, Jun 1997.

[Suks 08] K. Suksomboon, P. Pongpaibool, and C. Aswakul. "An Equilibrium Policy for Providing End-to-End Service Level Agreements in Interdomain Network". *In proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '08)*, pp. 2963–2968, April 2008.

[Swal 05] G. Swallow, J. Drake, H. Ishimatsu, and Y. Rekhter. "Generalized Multiprotocol Label Switching (GMPLS) User-Network Interface (UNI): Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Support for the Overlay Model". RFC 4208 (Proposed Standard), Oct. 2005.

[Szeg 07] P. Szegedi, J. Szigeti, and T. Tinkler. "Reliable Control and Management Plane Design in Multi-Domain Optical Networks". *In proceedings of the 9th IEEE International Conference on Transparent Optical Networks, (ICTON '07)*, Vol. 3, pp. 193–196, July 2007.

[Szig 04] J. Szigeti, J. Tapolcai, T. Cinkler, T. Henk, and G. Sallai. "Stalled information based routing in multidomain multilayer networks". *In porceedings of the 11th International Telecommunications Network Strategy and Planning Symposium (Networks '04)*, pp. 297–302, June 2004.

[Szig 05] J. Szigeti, I. Ballok, and T. Cinkler. "Efficiency of information update strategies for automatically switched multi-domain optical networks". *In proceed-*

ings of 7th IEEE International Conference on Transparent Optical Networks (ICTON '05), Vol. 1, pp. 445–454, July 2005.

[Taka 05] M. Takagi, K. Mochizuki, K. Takahashi, M. Shimizu, and S. Yasukawa. "P2P and P2MP TE-LSPs packing algorithms for MPLS networks". *In proceedings of the 6th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT '05)*, pp. 409–414, Nov. 2005.

[Take 07] T. Takeda. "Framework and Requirements for Layer 1 Virtual Private Networks". RFC 4847 (Informational), Apr. 2007.

[Tang 07] Y. Tang, S. Chen, and Y. Ling. "State aggregation of large network domains". *Elsevier Journal of Computer Communications*, Vol. 30, No. 4, pp. 873–885, 2007.

[Thal 04] D. Thaler. "Border Gateway Multicast Protocol (BGMP): Protocol Specification". RFC 3913 (Historic), Sep. 2004.

[Tiru 04] A. Tirumala, M. Gates, F. Qin, J. Dugan, and J. Ferguson. "IPerf: IP Performance and Measurement Tool". 2004. [Online]. Available at: http://sourceforge.net/projects/iperf.

[Tora 06] P. Torab, B. Jabbari, Q. Xu, S. Gong, X. Yang, T. Lehman, C. Tracy, and J. Sobieski. "On Cooperative Inter-Domain Path Computation". *In proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC '06)*, pp. 511–518, 2006.

[Truo 06] D. L. Truong and B. Thiongane. "Dynamic routing for shared path protection in multidomain optical meshnetworks". *OSA Journal of Optical Networks*, Vol. 5, No. 1, p. , 2006.

[Truo 07] D.-L. Truong and B. Jaumard. "Using Topology Aggregation for Efficient Shared Segment Protection Solutions in Multi-Domain Networks". *IEEE Journal on Selected Areas in Communications*, Vol. 25, No. 9, pp. 96–107, December 2007.

[Ulud 07] S. Uludag, K.-S. Lui, K. Nahrstedt, and G. Brewster. "Analysis of Topology Aggregation techniques for QoS routing". *ACM Journal of Computing Surveys*, Vol. 39, No. 3, p. 7, 2007.

[Vass 06] J. Vasseur, Y. Ikejiri, and R. Zhang. "Reoptimization of Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Loosely Routed Label Switched Path (LSP)". RFC 4736, Internet Engineering Task Force, Nov. 2006.

[Vass 08a] J. Vasseur, R. Zhang, N. Bitar, and J. Roux. "A Backward Recursive PCE-based Computation (BRPC) Procedure To Compute Shortest Constrained Inter-domain Traffic Engineering Label Switched Paths". Internet-Draft draft-ietf-pce-brpc-09, Internet Engineering Task Force, Apr. 2008. Work in progress.

[Vass 08b] J. Vasseur, A. Ayyangar, and R. Zhang. "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)". RFC 5152 (Proposed Standard), Feb. 2008.

[Vela 06] L. Velasco, S. Spadaro, J. Comellas, and G. Junyent. "Failure Aware Diverse Routing: A Novel Algorithm to Improve Availability in ASON/GMPLS Networks". *In proceedings of the IEEE International Conference on Transparent Optical Networks (ICTON '06)*, Vol. 3, pp. 195–198, June 2006.

[Vela 08] L. Velasco, R. Romeral, F. Agraz, S. Spadaro, J. Comellas, G. Junyent, and D. Larrabeiti. "On the design of MPLS-ASON/GMPLS interconnection mechanisms". *In the proceedings of the International Conference on Optical Network Design and Modeling (ONDM '08)*, pp. 1–6, March 2008.

[Veri 09] Verizon. "Internet Dedicated Service Service Level Agreement (SLA)". Online, Jan. 2009. http://www.verizonbusiness.com/terms/us/products/internet/sla/.

[Vija 08] K. Vijayalakshmi and S. Radhakrishnan. "Artificial immune based hybrid GA for QoS based multicast routing in large scale networks (AISMR)". *IEEE Journal of Computer Communications*, Vol. 31, No. 17, pp. 3984–3994, 2008.

[Vogt 03] M. Vogt, R. Martens, and T. Andvaag. "Availability modeling of services in IP networks". *In proceedings of 4th International Workshop on Design of Reliable Communication Networks (DRCN '03)*, pp. 167–172, Oct. 2003.

[Wang 00] B. Wang and J. C. Hou. "Multicast routing and its QoS extension: problems, algorithms, and protocols". *IEEE Network Journal*, Vol. 14, pp. 22–36, 2000.

[Wang 05] H. Wang, H. Xie, Y. R. Yang, A. S. Silberschatz, L. E. Li, and Y. Liu. "Stable Egress Route Selection for Interdomain Traffic Engineering: Model and Analysis". *In proceedings of the 13TH IEEE International Conference on Network Protocols (ICNP '05)*, pp. 16–29, 2005.

[Wint 87] P. Winter. "Steiner problem in networks: a survey". *Wiley Journal on Networking*, Vol. 17, No. 2, pp. 129–167, 1987.

[Xie 97] H. Xie and J. Baras. "Performance analysis of PNNI routing in ATM networks: hierarchical reduced load approximation". *In proceedings of IEEE Conference on Military Communications (MILCOM '97)*, Vol. 2, No. , pp. 998–1002 vol.2, Nov 1997.

[Xue 03] G. Xue, L. Chen, and K. Thulasiraman. "Quality-of-service and quality-of-protection issues in preplanned recovery schemes using redundant trees". *IEEE Journal on Selected Areas in Communications*, Vol. 21, No. 8, pp. 1332–1345, Oct. 2003.

[Yang 03] X. Yang and B. Ramamurthy. "Inter-domain dynamic routing in multi-layer optical transport networks". *In proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '03)*, Vol. 5, pp. 2623–2627, Dec. 2003.

[Yann 05] M. Yannuzzi, S. Sanchez-Lopez, X. Masip-Bruin, J. Sole-Pareta, and Jordi-Domingo-Pascua. "A combined intra-domain and inter-domain QoS routing model for optical networks". *In proceedings of the IFIP Conference on Optical Network Design and Modeling (ONDM '05)*, pp. 197–203, 7-9, 2005.

[Yann 06a] M. Yannuzzi, X. Masip-Bruin, S. Sanchez-Lopez, E. Tordera, J. Sole-Pareta, and J. Domingo-Pascual. "Interdomain RWA Based on Stochastic Estimation Methods and Adaptive Filtering for Optical Networks". *In proceedings of the Global Telecommunications Conference (GLOBECOM '06)*, pp. 1–6, Dec 2006.

[Yann 06b] M. Yannuzzi, X. Masip-Bruin, S. Sanchez, J. Domingo-Pascual, A. Orda, and A. Sprintson. "On the challenges of establishing disjoint QoS IP/MPLS paths across multiple domains". *IEEE Communications Magazine*, Vol. 44, No. 12, pp. 60–66, Dec. 2006.

[Yann 08] M. Yannuzzi, X. Masip-Bruin, G. Fabrego, S. Sanchez-Lopez, A. Sprintson, and A. Orda. "Toward a new route control model for multidomain optical networks". *IEEE Communications Magazine*, Vol. 46, No. 6, pp. 104–111, June 2008.

[Yasu 06] S. Yasukawa. "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)". RFC 4461 (Informational), Apr. 2006.

[Yous 08] M. Youssef, B.-Y. Choi, C. Scoglio, and E. K. Park. "Dynamic Hybrid Topology Design for Multicast in Constrained WDM Networks". *In proceedings of 17th International Conference on Computer and Communications and Networks (ICCCN '08)*, pp. 1–8, Aug. 2008.

[Zhu 03] Y. Zhu, A. Jukan, and M. Ammar. "Multi-segment wavelength routing in large-scale optical networks". *In proceedings of the IEEE International Conference on Communications (ICC '03)*, Vol. 2, No. , pp. 1381–1385, May 2003.

[Zhu 95] Q. Zhu, M. Parsa, and J. J. Garcia-Luna-Aceves. "A source-based algorithm for delay-constrained minimum-cost multicasting". *In proceedings of the 14th Annual Joint Conference of the IEEE Computer and Communication Societies (INFOCOM '95)*, p. 377, 1995.

Appendix A

Intersim: An Inter-domain Discrete Event Simulator

A.1 Introduction

The availability of several simulation packages, either commercial or free, makes simulation one of the most commonly used techniques for performance evaluation of computer systems and networks. In general, evaluation techniques, can be subdivided into two main categories: measurement (or empirical) techniques and modeling techniques.

Empirical techniques require that the system, or network to be evaluated, to exist and direct measurements of the evaluation target be taken. On the other hand, modelling techniques only require a model of the system. Modeling techniques are of two types: simulation and analytic. Among them, simulation is the most popular, as it applies to a wider variety of systems and does not require restrictive assumptions.

In this appendix, we present the design and implementation of the Inter-domain Simulation (**intersim**) toolkit. Today many different simulation tools have been developed for *Discrete Event Simulations* (DES), *e.g.*, **OPNET Modeler**^{TM}, **ns2**, and **OMNET**, *etc.*. However, most lacked the support for several inter-domain related protocols such as PCEP for P2P and P2MP paths, as well as extensions of RSVP-TE for P2MP inter-domain LSPs. **intersim** was purposely developed to investigate the path computation and performance issues in multi-domain and hierarchical networks. The simulator is a high level implementation of the P2P and P2MP RSVP-TE signalling protocol for

GMPLS-TE networks, as well as PCE Protocol for collaborative inter-domain path computation. The tool also included TE-related components such as resource management (RM), QoS path computation algorithms, traffic scheduling and real-time dynamic traffic workload.

intersim allows the configuration of different multi-domain network topology by defining topology files per-domain as shown in Tables A.1, A.2, A.3, and A.4. Several link the parameters such as propagation delay, availability, cost are also configurable through the configuration files.

intersim was implemented using the **C++** Object-oriented language which provides a modular architecture allowing rapid "plug-in" of new code and quantitatively assess the impact of different proposed strategies on network performance. In total the intersim was composed of around 10,000 lines of code.

A.2 The model

Intersim is designed so that the network can be studied under dynamic state where LSP connection requests arrive and leave randomly according to certain traffic rate distributions. Alternatively, it is possible to set the desired number of LSP requests that should be attempted, and allow the LSP's lifetime to span throughout the simulation time (*i.e.*, never torn). In our design the we model the arrival of LSP requests according to a Poission process, and LSP lifetime (holding time) to be exponentially distributed. The inter-arrival time for LSP requests in this case are exponentially distributed with a mean λ.

In order to be able to test blocking of requests as a measure of network performance, we assume that LSP requests are not persistent. In other words, if an LSP request fails to reach its destination, it does not re-initiate another attempt and is counted towards the blocked requests.

A.3 Design and class hierarchy

DomainController class reference. The DomainController class defines the control entity that is responsible for generating and processing events within a certain domain.

The DomainController contains a Topology object that defines the state of the traffic engineering database TED of the domain at a certain time of the network simulation. Events generated by the DomainController object are queued in an event queue that belongs to the domain. When running in multi-domain mode, events are collected by the inter-domain controller from each domain and processed based on the earliest domain event time.

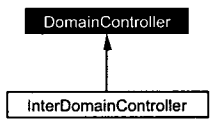


Figure A.1: Inheritance diagram for DomainController

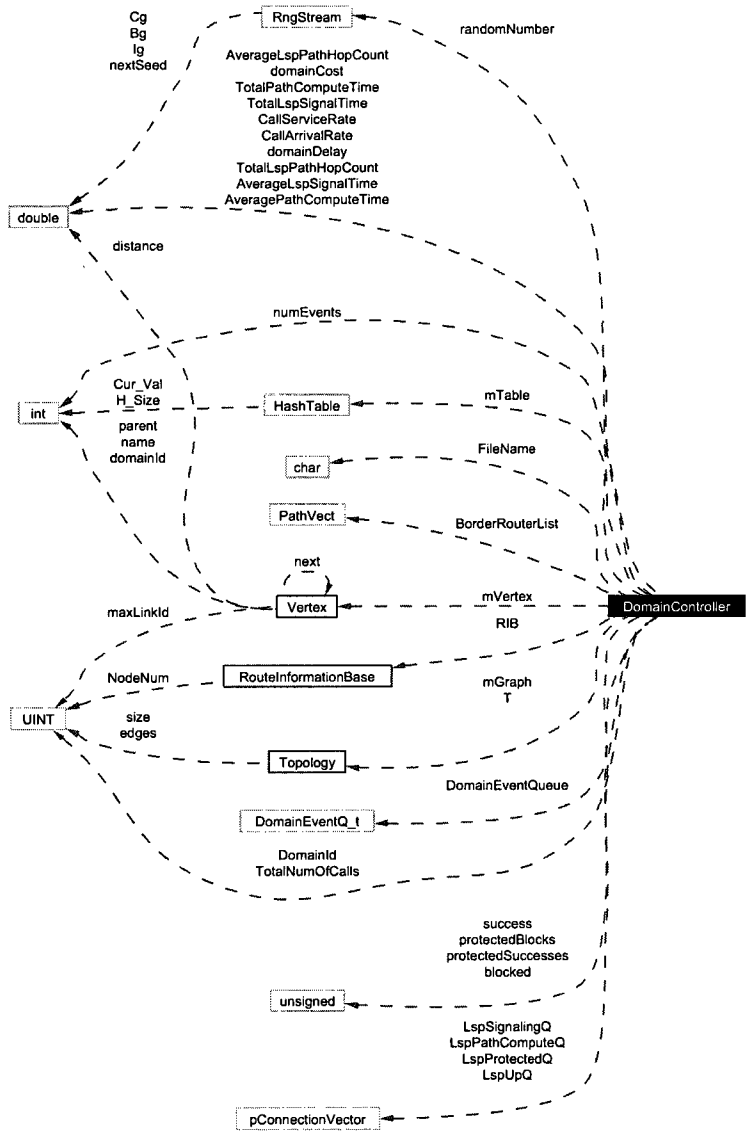


Figure A.2: Collaboration diagram for DomainController

InterDomainController class reference. The InterDomainController class is a subclass of DomainController and inherits all it's public/protected members. In addition to domain attributes, it contains inter-domain specifics *e.g.*, related inter-domain traffic such as average inter-domain LSP blocking, average LSP signaling time, average inter-domain LSP hop count, *etc.*.

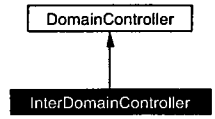


Figure A.3: Inheritance diagram for InterDomainController

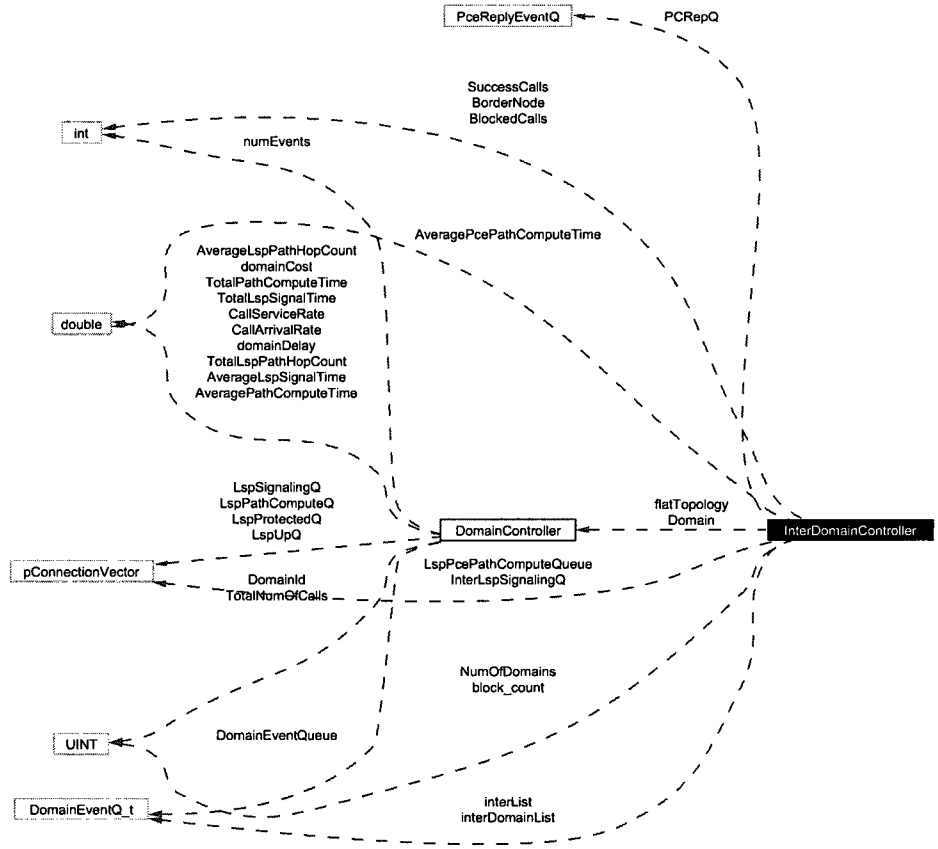


Figure A.4: Collaboration diagram for InterDomainController

Link class reference. The class Link represents TE link attributes of an edge connecting two vertices/nodes in the TE topology.

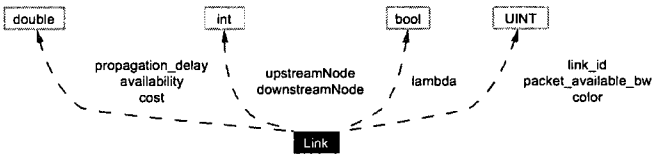


Figure A.5: Collaboration diagram for Link

Vertex class reference The class vertex represent a node in the topology.

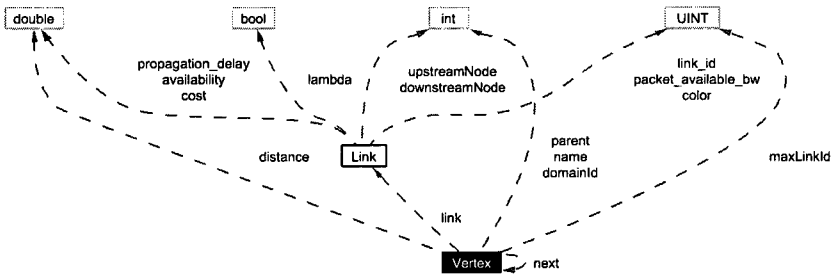


Figure A.6: Collaboration diagram for Vertex

Topology class reference. Represents the topology or directed graph of connected vertices. The topology is a member of a domain.

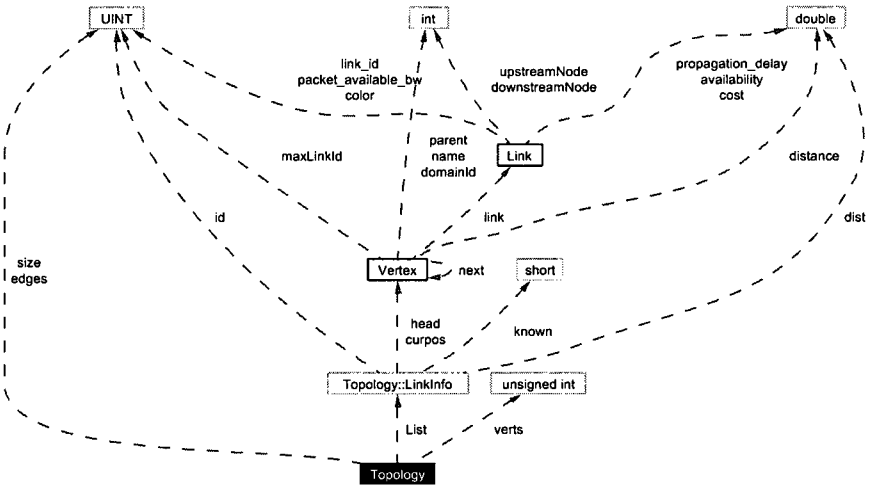


Figure A.7: Collaboration diagram for Topology

RouteInformationBase class reference Represents the inter-domain Routing Information database— for example, that learnt from BGP.

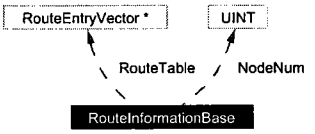


Figure A.8: Collaboration diagram for RouteInformationBase

Event class reference Represents an event that is generated by a node in a domain in the network.

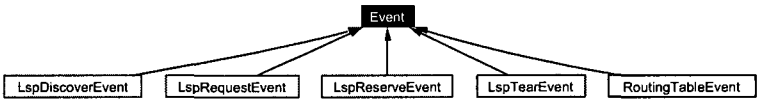


Figure A.9: Inheritance diagram for Event

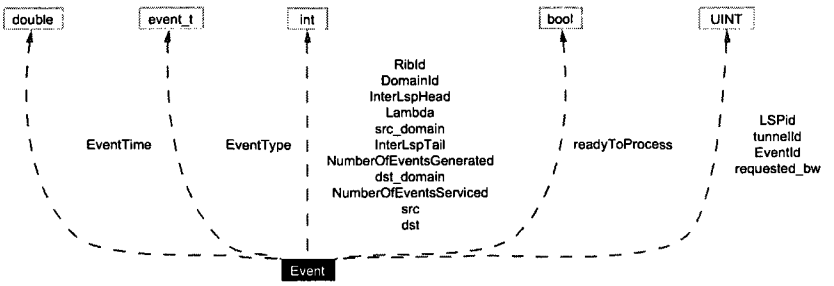


Figure A.10: Collaboration diagram for Event

RoutingTableEvent class reference Represents a routing event in generated by node in the inter-domain network.

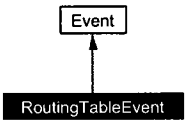


Figure A.11: Inheritance diagram for RoutingTableEvent

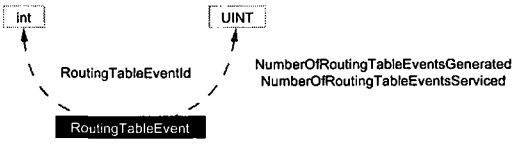


Figure A.12: Collaboration diagram for RoutingTableEvent

LspTearEvent class reference Emulates an LSP RSVP PathTear message.

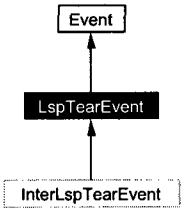


Figure A.13: Inheritance diagram for LspTearEvent

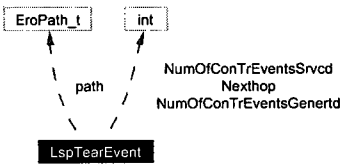
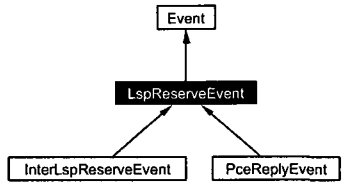


Figure A.14: Collaboration diagram for LspTearEvent

LspReserveEvent class reference Emulates an LSP RSVP Resv message.



captionInheritance diagram for LspReserveEvent

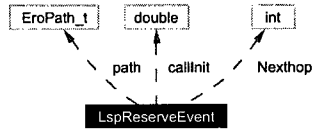


Figure A.15: Collaboration diagram for LspReserveEvent

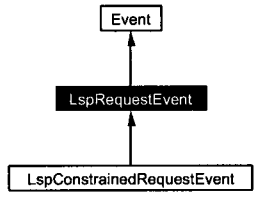


Figure A.16: Inheritance diagram for LspRequestEvent

LspRequestEvent class reference

LspDiscoverEvent class reference Emulates an LSP RSVP Path message.

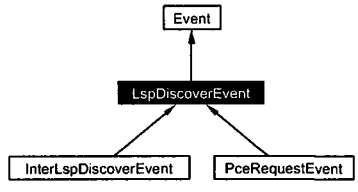


Figure A.17: Inheritance diagram for LspDiscoverEvent

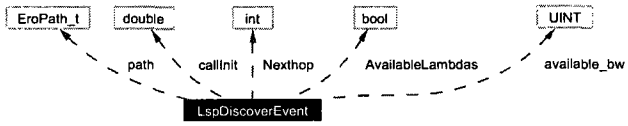


Figure A.18: Collaboration diagram for LspDiscoverEvent

InterLspDiscoverEvent class reference Emulates an inter-domain LSP RSVP Path message.

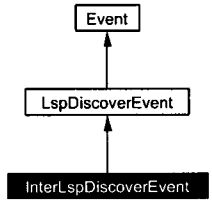


Figure A.19: Inheritance diagram for InterLspDiscoverEvent

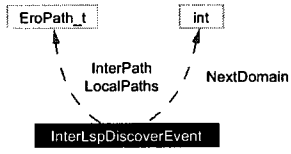


Figure A.20: Collaboration diagram for InterLspDiscoverEvent

InterLspReserveEvent class reference Emulates an inter-domain LSP RSVP Resv message.

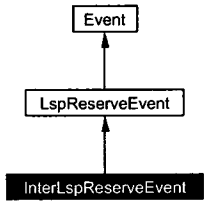


Figure A.21: Inheritance diagram for InterLspReserveEvent

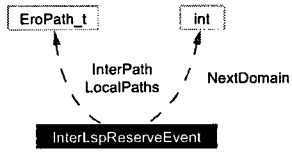


Figure A.22: Collaboration diagram for InterLspReserveEvent

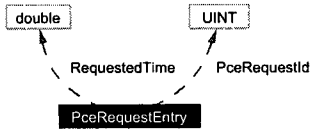


Figure A.23: Collaboration diagram for PceRequestEntry

PceRequestEntry class reference

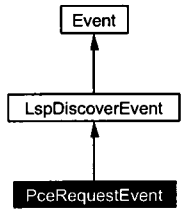


Figure A.24: Inheritance diagram for PceRequestEvent

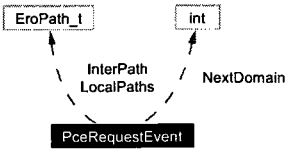


Figure A.25: Collaboration diagram for PceRequestEvent

PceRequestEvent class reference

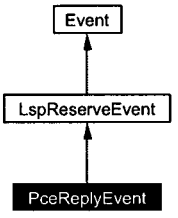


Figure A.26: Inheritance diagram for PceReplyEvent

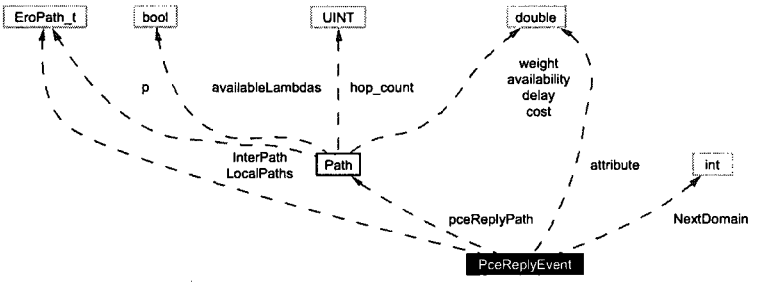


Figure A.27: Collaboration diagram for PceReplyEvent

PceReplyEvent class reference

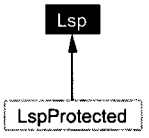


Figure A.28: Inheritance diagram for Lsp

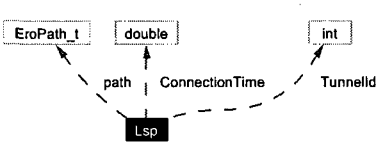


Figure A.29: Collaboration diagram for Lsp

Lsp class reference

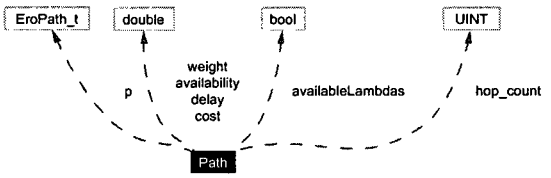


Figure A.30: Collaboration diagram for Path

Path class reference

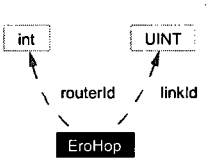


Figure A.31: Collaboration diagram for EroHop

EroHop Class Reference

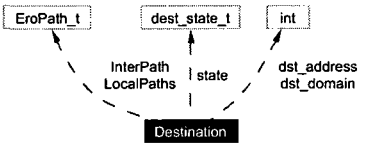


Figure A.32: Collaboration diagram for Destination

Destination Class Reference

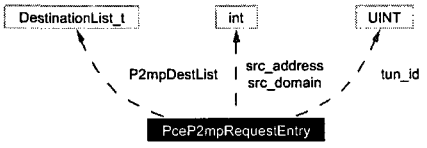


Figure A.33: Collaboration diagram for PceP2mpRequestEntry

PceP2mpRequestEntry Class Reference

A.4 Flow-chart design

Figure A.34 shows the main event loop of the inter-domain controller that is responsible for generating inter and intra-domain traffic, collecting events from all domains in the network, dispatching events to corresponding domain for processing, collecting statistics and terminating the simulation when done.

A.5 Topology files

Tables show the topology files for the inter-domain network that were used in the simulation experiments in Chapters 4, 5, and 6.

Table A.1: Topology file for domain 1

LSR *(from/to)*	LSR *(from/to)*	Link delay *(sec)*
PE2	PE3	0.00509766
PE2	P1	0.00747803
PE2	P2	0.00343445
PE3	P1	0.00753723
PE3	P2	0.00519409
P1	P2	0.00727234
P1	PE7	0.00700000
P2	PE7	0.00700000
P2	PE8	0.00700000
P1	PE8	0.00700000
P1	PE10	0.00700000
P2	PE10	0.00710000
P3	PE10	0.00720000

Table A.2: Topology file for domain 2

LSR *(from/to)*	LSR *(from/to)*	Link delay *(sec)*
PE2	PE1	0.00509766
PE2	P1	0.00747803
P1	P2	0.00727234
P1	P4	0.00314636
P1	PE1	0.00678375
PE1	P2	0.00239594
PE1	PE4	0.00477478
PE4	P3	0.00753723
PE4	P1	0.00197449
P2	P3	0.00397449
P4	P2	0.00597449
PE8	P1	0.00497449
PE8	P4	0.00507449
PE9	P2	0.00497449
PE9	P4	0.00497000

Table A.3: Topology file for domain 3

LSR *(from/to)*	LSR *(from/to)*	Link delay *(sec)*
P1	PE1	0.00112305
P1	PE3	0.00365112
P1	PE5	0.00710419
PE3	PE1	0.0024411
PE1	PE5	0.00463287
PE7	PE5	0.00700000
PE7	P1	0.00700000
PE6	PE1	0.00700000
PE6	P1	0.00700000
PE9	P1	0.00700000
PE9	PE6	0.00700000
PE9	PE1	0.00700000

Table A.4: Topology file for domain 4

LSR *(from/to)*	LSR *(from/to)*	Link delay *(sec)*
P1	PE4	0.00112305
P1	PE5	0.00365112
PE4	PE5	0.0024411
PE6	PE4	0.00700000
PE6	P1	0.00700000
PE10	P1	0.00700000
PE10	PE4	0.00700000
PE10	PE5	0.00700000

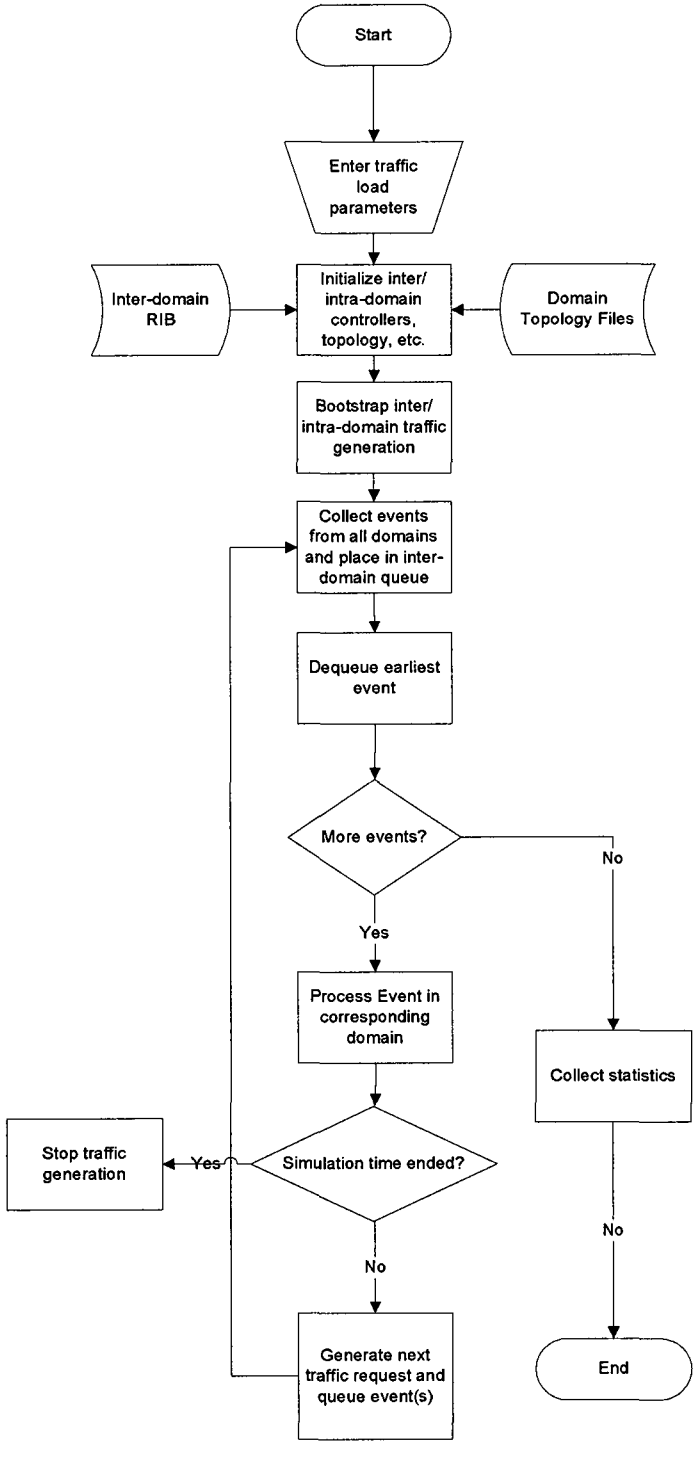


Figure A.34: Flow-chart diagram **intersim** simulator

Appendix B

Confidence Intervals

Simulated quantities such as blocking probability are measured by taking the mean of a succession of n runs, each of long enough time to ensure uncorrelated results. All runs are identical and independent from each other. The n independent results will be represented by $B_1, B_2, B_3, \cdots, B_{n-1}, B_n$.

$$\text{The Mean } \bar{B} = \frac{1}{n}\sum_{i=1}^{n} B_i \tag{B.1}$$

However, the mean of the independent simulation runs $\bar{B}$ provide us with a single numerical value for the estimate of the expected value $E|B| = \mu$. In order to know how good is the estimate provided by $\bar{B}$ for the simulation results, it is necessary to compute the variance of V_b^2.

$$V_b^2 = \frac{1}{n-1}\sum_{i}^{n}(B_i - \bar{B}) \tag{B.2}$$

Small V_b^2 indicates that the results are tightly clustered around $\bar{B}$, and we can be confident that $\bar{B}$ is close to the $E|B|$. On the other hand, if V_b^2 is large, the results are widely dispersed about $\bar{B}$ and we can not be confident that $\bar{B}$ is close to the $E|B|$. Instead of seeking a single value to estimate the $E|B|$, we can specify the interval of values that is highly likely to contain the true value of the parameter. We begin by specifying some high probability, say $1 - \alpha$. We then find the interval $[L(B), U(B)]$ such that the probability:

$$\Pr[L(B) \leq \mu \leq U(B)] = 1 - \alpha \tag{B.3}$$

This interval contains the true value of the parameter with probability $1 - \alpha$. Such an interval is $(1 - \alpha) \times 100\%$ confidence interval.

Using the standard deviation and the t distribution table, the lower and upper limits of the 95% confidence interval can be calculated as follows:

$$\text{Lower Limit } L(B) = \bar{B} - \frac{\sigma t_{[\frac{\alpha}{2},n-1]}}{\sqrt{n}} \tag{B.4}$$

$$\text{Upper Limit } U(B) = \bar{B} + \frac{\sigma t_{[\frac{\alpha}{2},n-1]}}{\sqrt{n}} \tag{B.5}$$

$$\tag{B.6}$$

where:

$$\begin{aligned} \alpha &= 0.05 \\ n &= \text{number of observations} \\ \bar{B} &= \text{sample average} \\ \sigma &= \text{sample standard deviation} \\ &= \sqrt{V_b^2} \\ &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(B_i - \bar{B})} \end{aligned}$$

The confidence interval means that 95% of the simulation results falls within the interval. Throughout this thesis, the confidence interval is computed based on seven independent runs. From the table of the t distribution, the $t_{[\frac{\alpha}{2},6]}$ is found to be 2.447. It was observed that more than 95% of the results were within the calculated confidence interval for each experiment. Table B.1 shows an example of how confidence interval is calculated. The table shows the blocking probability performance of the availability-bounded heuristic presented in Chapter 4 when simulated on the network topology in Figure 5.6 with a link capacity of OC-192 for all links in the network. The average blocking probabilities for the seven independent runs are shown together with the calculated mean $\bar{B}$, the upper and lower values of the interval $U(B) - L(B)$. The confidence intervals were not shown in the figures, however, since the intervals were relatively small as illustrated in the example below.

Load (Erlang)	Simulation run averages							$\bar{B}$	σ	$U(B)$	$L(B)$	Interval
	B_1	B_2	B_3	B_4	B_5	B_6	B_7					
1	2.854	2.852	2.850	2.852	2.840	2.843	2.854	2.849	0.006	2.854	2.844	0.010
5	3.158	3.212	3.232	3.235	3.241	3.238	3.239	3.222	0.030	3.250	3.194	0.055
10	7.125	7.128	7.124	7.131	7.141	7.142	7.142	7.133	0.008	7.141	7.126	0.015
12	10.134	10.134	10.134	10.043	10.053	10.063	10.083	10.092	0.041	10.130	10.054	0.076
14	25.035	25.006	25.057	25.080	25.055	25.052	25.056	25.049	0.023	25.070	25.028	0.042
16	30.327	30.314	30.293	30.334	30.324	30.320	30.322	30.319	0.013	30.331	30.307	0.024
18	32.137	32.101	32.169	32.126	32.129	32.130	32.198	32.142	0.032	32.171	32.112	0.059
20	35.734	35.716	35.730	35.709	35.723	35.732	35.780	35.732	0.023	35.754	35.711	0.043
25	54.326	54.315	54.400	54.352	54.393	54.330	54.346	54.352	0.033	54.382	54.321	0.061
30	61.045	61.036	61.060	61.075	61.042	61.069	61.078	61.058	0.017	61.073	61.042	0.031

Table B.1: Example of confidence interval calculations