Overleefbaarheidsaspecten
van toekomstige optische transportnetwerken

Survivability Aspects of Future Optical Backbone Networks

Dimitri Staessens

Promotoren: prof. dr. ir. D. Colle, prof. dr. ir. M. Pickavet
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Computerwetenschappen

Vakgroep Informatietechnologie
Voorzitter: prof. dr. ir. D. De Zutter
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2012 - 2013

UNIVERSITEIT
GENT

ISBN 978-90-8578-538-5
NUR 986
Wettelijk depot: D/2012/10.500/64

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur
Vakgroep Informatietechnologie

Promotoren:

Prof. Dr. Ir. Didier Colle
Prof. Dr. Ir. Mario Pickavet

Leden van de Examencommissie:

Prof. Dr. Ir. Daniël De Zutter (voorzitter)
Prof. Dr. Ir. Piet Demeester (secretaris)
Prof. Dr. Ir. Didier Colle (promotor)
Prof. Dr. Ir. Mario Pickavet (promotor)
Prof. Dr. Manos Varvarigos (University of Patras)
Prof. Dr. Ir. Kris Steenhaut (Vrije Universiteit Brussel)
Prof. Dr. Ir. Joris Walraevens (UGent, TELIN)
Dr. Ir. Bart Puype

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur

Vakgroep Informatietechnologie
Gaston Crommenlaan 8 bus 201, B-9050 Gent, België

Tel. +32-9-331.49.00
Fax. +32-9-331.48.99

Dankwoord

Dankuwel. Merci. Thank you. De eerste woorden die in dit proefschrift staan zijn ontegensprekelijk de belangrijkste. Ze zijn gericht aan iedereen die mij de laatste jaren nauw aan het hart heeft gelegen, en ook aan U, de lezer, om dit werk ter hand te nemen.

Vooreerst wens ik mijn promotoren, prof. Dr. ir. Didier Colle en prof. Dr. ir. Mario Pickavet te bedanken voor hun toewijding en inzet tijdens de jaren die ik bij de IBCN onderzoeksgroep heb doorgebracht. Ik dank hen voor de duidelijke uiteenzetten die mij in de juiste richting wezen terwijl ik mijn eerste stapjes zette in dit voor mij toen onbekende onderzoeksdomein, voor de vruchtbare gesprekken en vergaderingen die mij bleven inspireren en motiveren, en voor hun ervaren hulp bij het schrijven van de publicaties die uit ons onderzoek voortvloeiden en die samen de materie vormen die in dit proefschrift is samengevat. Ik dank hen ook voor het vele lezen, nalezen en corrigeren van de inhoud en de vormgeving van dit doctoraatsproefschrift, ondanks hun drukke agenda. Het uiteindelijke proefschrift dat U nu in handen heeft is dan ook grotendeels hun verdienste.

Verder dank ik ook prof. Dr. ir. Luc Taerwe, Decaan van de Faculteit Ingenieurswetenschappen en Architectuur, prof. Dr. ir. Daniël De Zutter voorzitter van de vakgroep Informatietechnogie, en prof. Dr. ir. Piet Demeester, de enthousiaste leider van onze IBCN onderzoeksgroep, voor het stimulerende onderzoeksklimaat waar ze samen dag na dag verantwoordelijk voor zijn. En dan mag ik ook zeker prof. Dr. ir. Peter Van Daele niet vergeten, die mij met zijn uitgebreide ervaring wegwijs maakte in enkele praktische aspecten van het conferentiegaan. Dankzij zijn tips heb ik nooit langer dan nodig aan een buffet staan wachten. De conferentiediners van ECOC waar ik aan tafel mocht aanschuiven bij Peter en zijn lieve vrouw An zullen mij altijd bijblijven.

Dit onderzoek werd mogelijk gemaakt door mijn werkgever, het Interdisciplinair Instituut voor Breedband Technologie, IBBT, dat ondertussen is uitgegroeid tot een wereldwijd erkend en gerespecteerd onderzoekscentrum, dat bestaat uit onderzoeksgroepen aan de Vlaamse universiteiten. Ik dank ook de Europese Commissie voor hun financiële steun aan de talrijke projecten waar onze onderszoeksgroep de laatste jaren heb aan meegewerkt.

De mensen waarmee ik nauw samenwerk zijn natuurlijk de belangrijkste voor het dagelijkse leven aan de Universiteit: Wouter, Sachin, Sander, Abhishek, Ward, Bram, Bart, Sofie en Pieter zijn ondertussen zoveel meer dan collega's, en ook mijn andere fijne bureaugenoten, Sahel, Sofie, Thijs, Maarten en Sofie mogen natuurlijk niet ontbreken. Zij zorgen er al jaren voor dat ik dagelijks met een glimlach naar ons bureau kom, en telkens met een glimlach vertrek. Ook mijn vorige bureaugenoten, Bart, Ruth, Jan, Adelbrecht, Evy, Lien, Lieven, Stijn, Koen, Florian, Maarten, Willem en Bart bedank ik voor de mooie dagen in de "oude" 3.13.

Ook de vele mensen die mij tijdens het onderzoek hebben bijgestaan mogen in dit dankwoord niet ontbreken. Bij respectievelijk prof. Dr. Manos Varvarigos van de Universiteit van Patras en Dr. Uri Mahlab van het ECI Telecom en het Holon Institute of Technology ben ik enige tijd ontvangen om samen te werken rond het probleem van foutdetectie in volledige transparante netwerken. Hier dank ik ook Kostas Manousakis en Panagiotis Kokkinos, die zowel tijdens als na het onderzoek van Rio en Patras voor mij een tweede thuis hebben gemaakt. Matthias Gunkel van Deutsche Telekom heeft grote bijdragen geleverd aan dit proefschrift, in het bijzonder was hij de grote bezieler van de kostenmodellen die werden gebruikt. De samenwerking met Ricardo Romeral Ortega leverde een substantiële bijdrage tot de multidomein oplossing. Verder dank ik ook alle mensen waarmee ik samen heb gewerkt en van ideeën heb gewisseld tijdens de talrijke Europese projecten en conferenties, en in het bijzonder Thierry Zami en Annalisa Morea, Salvatore Spadaro, Jordi Perelló, Davide Careglio, Josep Solé Pareta, Michael Eiselt, Anna Manolova, Sarah Ruepp, Jose Luis Marzo, Maurice Gagnaire, David Dahan, Marianna Angelou, Siamak Azodolmolky, Reza Nejabati, Eduard Escalona, Dimitra Simeonidou, Yabin Ye, Saradhi Chava, Ioannis Tomkos, Tibor Cinkler, Andreas Gladisch, Fritz-Joachim Westphal, Steffen Topp, Mario Kind, Wolfgang John, Alisa Devlic, Pontus Sköldström, Attila Takacs, András Kern, David Jocha, Hagen Woesner and Andreas Koepsel. Ook bedankt aan de nog niet vermelde leden van de lees- en examencommissie, prof. Dr. ir. Kris Steenhaut en prof. Dr. ir. Joris Walraevens.

Bij het verrichten van onderzoek komen natuurlijk ook heel wat administratieve zaken kijken, zoals het inschrijven voor conferenties, het boeken van reizen, het regelen van vergaderzalen voor projecten en het verwerken van de onkostennota's. Dat dit voor ons geen onoverkomelijke rompslomp is, is vooral te danken aan het efficiënte secretariaat, waar Joke, Nathalie, Martine, Davinia, Karien, Ilse en Bernadette altijd klaar staan om ons met de glimlach verder te helpen.

Natuurlijk mag ik de belangrijkste mensen in mijn leven niet vergeten, Mama, mijn broer Kevin, mijn fantastische schoonzus Doorke, Meter, Tante Anne, Nonkel Peter, Nonkel Bernard en Tante Christianne, mijn nichtje Lindsay. En dan zijn er die gekke mensen die mijn vrienden zijn en elke dag kleuren en kruiden tot het zinneprikkelende spektakel waar ze samen voor zorgen , Gisela, Anthony, Lindsay, Mos, Kinga, Nico, Bert, Leonore, Seppe, Jeroen, Pieter, Thomas, David, Frederik, Benjamin, Elke, Birger, Steffen, Fredje, Eva, Peter, Bart, Stef, Christian, Anne-Sofie, Isabel, Yves, Svetlana, Yoshi, Katrien, Sandra, Guy, Tim, Geert, Rozemie, Maarten, Michiel, Kurt, Hademou, Charlotte en de oude bende van Vlissegem, Stijn, Kurt, Sarah, Nick, Bart. En natuurlijk ook iedereen die ik moest bedanken en die ik tot mijn scha en schande vergeten ben.

Love you guys!

Amsterdam, 18 september 2012
Dimitri Staessens

Table of Contents

List of Figures

List of Tables

List of Acronyms

A

ABR	Area Border Router
AS	Autonomous System
ASBR	Autonomous System Boundary Router
ASON	Automatically Switched Optical Network
ATM	Asynchronous Transfer Mode
AWG	Arrayed Waveguide Grating

B

BCBC	Backup Connection Backup Capacity
BCWC	Backup Connection Working Capacity
BER	Bit Error Rate
BGP	Border Gateway Protocol
BR	Backbone Router

C

CapEx	Capital Expenditures
CAGR	Compound Annual Growth Rate
CCD	Charge Coupled Device
CDC	Colorless Directionless Contentionless
CR-LDP	Constraint-based Routing LDP
CU	Cost Units

D

DCC	Digital Communication Channel
DCF	Dispersion Compensation Fiber
DFA	Doped Fiber Amplifier
DGE	Dynamic Gain Equalizer
DICONET	Dynamic Impairment COnstraint NEtworking for Transparent Optical Networks
DWDM	Dense Wavelength Division Multiplexing

E

E/O	Electronic/Optical
EDFA	Erbium Doped Fiber Amplifier
EGP	Exterior Gateway Protocol
ERO	Explicit Route Object

F

FC	Filter Concatenation
FSC	Fiber Switch Capable
FWO	Fonds voor Wetenschappelijk Onderzoek, the Flemish NSF

G

GCC	General Communication Channel
GMPLS	Generalized Multiprotocol Label Switching
GSRO	Gateway Selection Route Object

H

HD	High Definition
HiTA	Hierarchical Routing with Topology Aggregation
HSDPA	High-Speed Downlink Packet Access

I

IA-RWA	Impairment Aware Routing and Wavelength Assignment
IETF	Internet Engineering Task Force
IGP	Interior Gateway Protocol
ILP	Integer Linear Program
ION	Intelligent Optical Network
IP	Internet Protocol
ISIS	Intermediate System To Intermediate System
ISIS-TE	Intermediate System To Intermediate System, Traffic Engineering Extensions
ISP	Internet Service Provider
ITU	International Telecommunication Union
ITU-T	ITUs Telecommunication Standardization Sector

L

L2SC	Layer 2 Switch Capable
LASER	Light Amplification by Stimulated Emission of Radiation
LCoS	Liquid Crystal on Silicon
LDP	Label Distribution Protocol
LMP	Link Management Protocol
LSC	Lambda Switch Capable
LSP	Label Switched Path
LSR	Label Switched Router

M

MEMS	Micro Electro Mechanical System
MIDP	Multiple Intra-Domain Protection
MLTE	Multilayer Traffic Engineering
MPLS	MultiProtocol Label Switching
MTL	Maximum Transparent Length

N

NMS	Network Management System
NPOT	Network Planning and Operation Tool
NSF	National Science Foundation

O

O/E	Optical/Electronic
O/E/O	Optical/Electronic/Optical
OADM	Optical Add-Drop Multiplexer
OAMP	Operations, Administration, Maintenance and Provisioning
OC	Optical Channel (SONET)
OCC	Optical Communication Controller
OCh	Optical Channel (OTN)
ODU	Optical Data Unit
OFELIA	OpenFlow in Europe, Linking Infrastructure and Applications
OMS	Optical Multiplex Section
OPU	Optical Channel Payload Unit
OSC	Optical Supervisory Channel
OSNR	Optical Signal-to-Noise Ratio
OSPF	Open Shortest Path First
OSPF-TE	Open Shortest Path First, Traffic Engineering extensions
OTN	Optical Transport Network
OTS	Optical Transmission Section
OTU	Optical Transport Unit
OXC	Optical Cross-Connect

P

PCC	Path Computation Client
PCE	Path Computation Element
PLI	Physical Layer Impairment
PMD	Polarization Mode Dispersion
PP	Path Protection
PXC	Photonic Cross Connect

Q

QoS	Quality of Service
QoT	Quality of Transmission

R

RFC	Request for Comments
ROADM	Reconfigurable Optical Add-Drop Multiplexer
RSVP-TE	Reservation Protocol, Traffic Engineering extensions
RWA	Routing and Wavelength Assignment

S

SDH	Synchronous Digital Hierarchy
SHDC	Shared Capacity
SLA	Service Level Agreement
SONET	Synchronous Optical Networking
SPARC	Split Architecture
SRLG	Shared Risk Link Group
STC	Standard Telephone and Cables
STM	Synchronous Transport Module
STS	Synchronous Transport Signal

T

TCO	Total Cost of Ownership
TCP	Transport Control Protocol
TDM	Time Division Multiplexing
TE	Traffic Engineering
TED	Traffic Engineering Database

U

UDP	User Datagram Protocol
UDWDM	Ultra Dense Wavelength Division Multiplexing

V

VPN	Virtual Private Network
VOA	Variable Optical Attenuator
VoD	Video-on-Demand
VoIP	Voice over IP

W

WCBC	Working Connection Backup Capacity
WCWC	Working Connection Working Capacity
WDM	Wavelength Division Multiplexing
WL	Wavelength-Link product
WSS	Wavelength Selective Switch

X

XT	Crosstalk

Nederlandse samenvatting –Summary in Dutch–

Transportnertwerken hebben in de laatste 2-3 decennia radicale veranderingen ondergaan. Het allereerste datatransportnetwerk was het NSFNet, ontrold door het National Science Foundation, het Amerikaanse Fonds voor Wetenschappelijk Onderzoek, in 1987. Het was gebaseerd op een T1 lijn, met een capaciteit van 1.5 Mb/s. Huidige smartphones hebben draadloos een bandbreedte beschikbaar die een orde groter is, meer dan 10 Mb/s, gebruikmakend van HSDPA. De huidige datasnelheid voor een enkel kanaal in een transportnetwerk bedraagt 10Gb/s, met 40Gb/s en zelfs 100Gb/s in de nieuwste optische systemen. Waar de eerste optische netwerksystemen een enkel kanaal per glasvezel overdroegen, gebruiken huidige systemen geavanceerde technieken (Wavelength Division Multiplexing, WDM) om tot 160 zulke kanalen, elk op een eigen specifieke golflengte, over een enkele vezel te sturen. Ook evolueerde het optische netwerk van een louter punt-tot-punt systeem tot een Automatisch geSchakeld Optisch Netwerk (ASON) dat zijn eigen systemen heeft om golflengten te schakelen van een bepaalde vezel naar een bepaalde vezel.

Dergelijke transportnetwerken voorzien breedbandverbindingen voor bijvoorbeeld het Internet. In tegenstelling tot transportnetwerken, waar de gegevens verstuurd worden in min-of-meer vaste circuits, is het internet een pakketgeschakeld netwerk. In pakketgeschakelde netwerken worden de gegevens verdeeld in kleine pakketjes die een etiket opgespeld krijgen met het adres van de bestemming en verstuurd tussen routers in zogenaamde datagrammen. De methode vertoont grote gelijkenissen met het versturen van een brief met de post. De brief wordt in een enveloppe verpakt, waar het adres van de bestemming wordt op vermeld en van postkantoor naar postkantoor gestuurd tot de bestemming is bereikt. Het resulterende netwerk, waar enerzijds pakketjes tussen routers verstuurd worden op basis van een adres, en anderzijds het transport tussen deze routers onderling verzorgd wordt door een optisch geschakeld netwerk, is een voorbeeld van een meerlaags netwerk.

Over één enkele optische vezel wordt tegenwoordig een gigantische hoeveelheid data verstuurd, ruwweg het equivalent van 25 miljoen gelijktijdige telefoongesprekken. Hierdoor zullen enkelvoudige netwerkstoringen, zoals breuken van een glasvezelkabel, de communicatie van een groot aantal eindgebruikers verstoren. Netwerkoperatoren kiezen er dan ook voor om hun netwerk zo te bouwen dat zulke grote storingen automatisch opgevangen worden. Dergelijke automatische mecha-

nismen worden overleefbaarheidsmechanismen genoemd. Alles komt tegen een prijs, en de prijs voor een hoge tolerantie tegen netwerkfouten is een sterk verhoogde netwerkkost. In geval van een fout moeten de gegevens immers over een alternatieve route worden verstuurd. Nu zullen overleefbaarheidsmechanismen in de transportlaag de betrouwbaarheid van alle verbindingen die gebruik maken van de transportlaag verhogen. Dit is een van de hoekstenen van meerlaags netwerkherstel, dat tot doel heeft om fouten in elke laag van het netwerk het hoofd te kunnen bieden en tegelijkertijd zo efficiënt mogelijk gebruik te maken van de aanwezige middelen.

Het onderzoek dat in deze thesis wordt uiteengezet spitst zich toe op twee aspecten rond overleefbaarheid in toekomstige optische netwerken. De eerste doelstelling die beoogd wordt in dit werk is het tot stand brengen van zeer robuuste dataverbindingen over meerdere netwerken om veeleisende toepassingen mogelijk te maken over langere afstand. Voorbeelden van dergelijke toepassingen zijn kritische verbindingen tussen grote datacenters, chirurgische ingrepen waarbij een robot wordt bestuurd op afstand of, dichter bij de thuisgebruiker, een online videotheek waar op aanvraag films kunnen worden afgespeeld in hoge resolutie (High Definition Video-on-Demand, HD VoD). Vandaag zijn VoD diensten gebonden aan het netwerk van de operator. In ons land hebben de grote telecomoperatoren (Belgacom en Telenet) elk hun eigen Video-on-Demand dienst, met elk hun eigen catalogus van aangeboden bioscoopprenten. De klanten van de ene operator kunnen niet gebruik maken van de videotheekdienst van de andere operator. Dit is niet enkel door de auteurswetgeving, maar ook door de technische moeilijkheden om voldoende betrouwbare verbindingen (die een positieve gebruikerservaring kunnen garanderen) tot stand te brengen over een infrastructuur die niet door een enkele entiteit wordt beheerd. Een enkele onafhankelijke aanbieder van Video-on-Demand voor Europa, met servers op een paar strategische locaties, kan een technische mogelijkheid worden door gebruik te maken van de technieken die bestudeerd werden in dit proefschrift.

De bestudeerde oplossing heeft niet enkel tot doel om een zeer betrouwbare verbinding tot stand te brengen, maar ook dit te bewerkstelligen met een minimum aan gebruikte netwerkcapaciteit. Om dit te verwezenlijken hebben we bestaande mechanismen voor meerlaagse netwerken toegepast op netwerken die opgedeeld zijn in meerdere onafhankelijke domeinen en geëvalueerd met betrekking tot het capaciteitsverbruik. Net als in het geval van een enkel domein was het zogenaamde ”common pool”concept het meest efficient voor dergelijke meerdomeinsnetwerken. Het originele concept wordt verder uitgebreid met de specifieke eigenschappen van een verbinding over meerdere domeinen om nog beter gebruik te maken van de beschikbare capaciteit, en we tonen aan dat de voorgestelde oplossing altijd kan worden toegepast in een netwerk waarvoor de topologie 2 onafhankelijke paden toelaat tussen eender welk knopenpaar. Dit is van groot belang omdat het toelaat om de verbinding lineair (domein per domein) tot stand te brengen. Het geleverde bewijs was constructief, wat onmiddelijk een heuristiek opleverde om de oplossing te benaderen. We hebben ook een optimale oplossing geformuleerd, gebruik makend van de techniek van het Integer Lineair Programmeren. Onze resultaten tonen aan dat de heuristiek gemiddeld zeer goed presteert (binnen 5% van de optimale oplossing), maar in specifieke gevallen toch een marge voor verbetering heeft tot 30%. We hebben ook een nieuw object voor het RSVP-TE protocol gedefinieerd dat ons toelaat om de beoogde verbindingen tot stand te brengen door gebruik te maken van GMPLS.

De tweede doelstelling die werd gesteld was om een antwoord te formuleren op de vraag hoe het toepassen van optische schakelsystemen gebaseerd op herconfigureerbare optische multiplexers (Reconfigurable Optical Add-Drop Multiplexers, ROADMs) een impact heeft op de overleefbaarheid (en de kost) van een optisch netwerk. Ten eerste tonen we aan dat het gebruik van ROADMS met beperkte flexibiliteit een nefaste invloed heeft op de capaciteit van het netwerk om dynamisch netwerkherstel uit te voeren, waar bij kleine netwerken tot 100% van de onderbroken trafiek niet kan worden hersteld door het optische netwerk. Ten tweede hebben we een grondig onderzoek uitgevoerd naar de mogelijke kostenbesparingen door het toepassen van geoptimalizeerde hersteltechnieken. We tonen aan dat het aantal fysieke verbindingen in het netwerk een kleine invloed heeft op deze kostenbesparingen, waarbij weinig vermaasde netwerken een groter voordeel hebben van optimalisatie dan sterk vermaasde netwerken. We ontdekten ook dat de hoeveelheid data die verstuurd wordt over het netwerk een belangrijke invloed heeft bij optisch geschakelde transportnetwerken. Bij lagere volumes hebben optisch geschakelde netwerken weinig voordeel (minder dan 5%) van dergelijke gesofistikeerde methoden, vooral dankzij hun efficiëntie om grote datavolumes te schakelen. Bij hogere volumes waar zeer veel golflengtes actief zijn in het netwerk, is het voordeel van optimalizatie (tot 20 % kostreductie) gelijkaardig aan de voordelen bij elektronisch geschakelde netwerken. Elektronisch geschakelde netwerken vertonen geen afhankelijkheid van het datavolume en hebben altijd een kostenreductie die tot 25 % bedraagt. Door het toenemende aantal golflengten per vezel en de toenemende capaciteit per golflengte komen wij in dit proefwerk tot de conclusie dat de huidige transparante transportnetwerken momenteel minder baat hebben bij het doorgedreven optimalizeren van herstelmechanismen.

English summary

In the last 2-3 decades, transport networks have undergone some radical changes. The first high-speed backbone for data communications was the NSFNet, created by the American National Science Foundation (NSF) in 1987. It was based on a T1 line which had a capacity of 1.544 Mb/s. Today's smartphones have wireless bandwidths of over 10 Mb/s over HSDPA. The current data rates for a single channel (called a wavelength) in a backbone network is 10Gb/s, with 40Gb/s and 100Gb/s being deployed. Where the first optical transmission systems were carrying only a single channel over a fiber, advanced techniques (called Wavelength Division Multiplexing, WDM) can transport up to 160 such channels on a single optical fiber. Furthermore, the optical network has evolved from a point-to-point optical system into Automatically Switched Optical Networks (ASONs) with their own switching functionality at wavelength granularity.

These transport networks provide high bandwidth pipes for, for instance, the Internet. In contrast to transport networks, where data is transported in fixed circuits, the Internet is a packet-routed network, where a data bitstream is divided into smaller packets that are labeled with a destination address and sent between routers in datagrams. The operation is not unlike an envelope being delivered from one postal office to the next. The resulting network, where packets are forwarded from one router to the next based on a destination address, and the transport between routers is done by an optically switched network (ASON), is an example of a multilayer network.

Due to the high data rate carried in a single optical fiber (currently roughly the equivalent of streaming 42000 High Definition movies simultaneously), cable cuts in a transport network can cause major disruptions for the end users of the network. Therefore network operators take precautions by deploying recovery mechanisms in their networks which react quickly in case of failures. Everything comes at a price, and the price for high resilience against failures is an increased cost of the network due to additional required resources. Recovery mechanisms in the transport layer will increase the reliability of all higher layer connections. This is one of the cornerstones for multilayer recovery, which aims at distributing the spare resources as optimally as possible between all network layers.

The research presented in this thesis focuses on two aspects of survivability in future optical networks. The first question we try to answer in this work is how we can enable highly reliable connections spanning multiple networks. Examples of applications which benefit such a solution are critical interconnections of datacenters over intercontinental distances or tele-surgery. An example closer to the

consumer is High Definition Video-on-Demand (VoD). Currently, VoD solutions are tied to an operator's network. If we take Belgium as an example, our major operators Belgacom and Telenet each manage their own VoD service, with their own catalog of movies to choose from. Customers from Telenet can not use the VoD service by Belgacom or vice versa. This is not only due to copyright laws, but also due to the technical difficulty of guaranteeing a positive user experience for a high bandwidth service such as VoD when you are not directly managing the entire network from video server to end user. A single independent VoD service provider in Europe, operating from a limited number of strategic locations, could become a technical feasibility using the multidomain connection setup techniques developed in this work.

The solution presented not only targets high availability, but also tries to maximize resource efficiency. In order to do this, we extended the concepts from multilayer networks towards multilayer multidomain networks, and evaluated the schemes. Just as in the single domain case, we found that the common pool solution, a highly efficient scheme for sharing protection resources in a network, was the most efficient for multidomain networks as well. We proved that a common pool solution for multidomain networks can be calculated in any 2-connected network topology, which is quite important because this means that we can establish the connection per domain in a linear way. This greatly reduces the complexity of path establishment, removing the need for any backtracking mechanisms. The proof was constructive, which led immediately to a heuristic algorithm for calculating the common pool solution. We also developed a mathematical optimum solution using Integer Linear Programming optimization techniques. Our results show that the heuristic performs very well on average (well within 5% of the optimum), but the ILP can outperform the heuristic up to 30% in some particular cases, which shows there is definitely room for future improvement. We also designed a new routing object for RSVP-TE, the GSRO (Gateway Specification Routing Object) which allows signaling of the proposed solution in networks using a GMPLS control plane.

The second question we answer in this thesis is how the increased deployment of optical circuit switches based on ROADM designs impact the network survivability. First we show that directional ROADM designs severely limit the capability of the network to perform restoration. Due to the lack of flexibility, only a small amount of traffic can be restored in the optical layer. We also performed a thorough investigation into the possible cost savings through resource sharing in transparent and opaque transport networks. The number of links in the network have a small impact on this gain, with sparsely meshed networks having greater benefit than densely meshed networks. We found that the network load plays an important role in transparent networks, where low load (i.e. few parallel line systems) means that the network does not benefit greatly from protection sharing. However, when the average required ROADM degree increases, the cost benefits approach the same levels as for traditional opaque networks. It will depend on the relative cost evolution of WSSs compared to transponders whether the cost benefits of protection sharing for transparent networks will increase or de-

crease. Opaque networks do not show a dependency on the load and always have a similar node cost gain from protection sharing. With the ongoing trend towards higher bitrates (400Gb/s and up), denser channel spacing and more efficient spectrum usage (Flexigrid), we think the balance for transparent networks will tip over towards the low load solution, meaning protection sharing may be less interesting to implement in such networks.

1 Introduction

1.1 Background

Communications services are playing a vital role in modern private, corporate and institutional life. This prevalent role is expected to continue to grow in importance for years to come. From the corporate and institutional point of view, strategic corporate functions become more dependent on communications between different offices and sites where even minor service interruptions can result in huge production delays and revenue loss.

The transport networks that these businesses are relying on are also evolving to meet future requirements. Optical technologies such as Wavelength Division Multiplexing (WDM) have drastically increased the bandwidth capacity at a low cost, expanding the service possibilities for these networks. This cost-effectiveness has driven the competition between operators to the point that there is little revenue to be made from the classical phone/fax/data service. Looking to increase revenue, these operators are now exploiting the capacity to introduce high bandwidth services such as broadband Internet access and high definition digital television. This trend fuels the quest for a converged network architecture, able to run all voice, data and multimedia services, commonly called triple-play. The scalability and robustness of the Internet protocol (IP) suite are the main reasons for its success, therefore IP is the network layer protocol of choice for these future networks. Flexibility was an issue in IP, but recent developments with Multi-Protocol Label Switching (MPLS), have introduced very powerful traffic engineering extensions that can be used in IP networks.

The cost-effectiveness of IP networks for data communications and telephony through Voice-over-IP (VoIP) solutions attracts a lot of attention from the corporate business community. While widely used for local area networks (LAN) at a single location, having a corporate IP network with internal VoIP spanning different sites is arguably the cheapest solution for handling internal data and voice traffic. To cope with this, most operators are offering VPN services over their network, subject to a Service Level Agreement (SLA).

The past 2 decades the growth in Internet traffic is explosive, with a projected total growth of 32% in the coming years. While Web browsing and Peer-to-Peer file sharing have been the predominant bandwidth consumers until just recently, online video (such as Youtube offering High Definition clips) is now the dominant driver for Internet growth, both in terms of average traffic and peak traffic. Even if consumption of video services is still different between average users and top users, which still use a lot of Peer-to-Peer solutions for instance, real-time online video services (i.e. streaming or progressive download of live events) are now mainstream. In 2010, only 3 percent of Internet traffic originated with non-PC devices, but by 2015 the non-PC share of Internet traffic will grow to 15 percent. PC-originated traffic will grow at a CAGR of 33 percent, while TVs, tablets, smartphones, and machine-to-machine (M2M) modules will have growth rates of 101 percent, 216 percent, 144 percent, and 258 percent, respectively.[1]

The bulk of Internet traffic is transported in the backbone networks, the high bandwidth pipes responsible for transporting huge traffic volumes over large distances. These networks are based on optical transport technologies due to their cost-efficiency. The fibers in an optical network can carry huge amounts of data. Current data rates for a single channel are 10Gb/s, with 40Gb/s and 100Gb/s being deployed, and advanced techniques can multiplex up to 160 such channels on a single fiber, increasing the data volume sent over a single fiber to 1.6 Tb (1600000000000 bits) of data every second. This is roughly the equivalent of 25 million simultaneous telephone calls or streaming 42000 High Definition movies. Fibers are typically packed into cables of hundreds of fibers, so the amount of traffic that will be lost should a cable be cut ensures fast failure recovery in backbone networks is paramount.

The work in this thesis is performed in the framework of future networks: Optical backbone networks supporting large amounts of (Internet) traffic. Such networks are multilayer networks: each network technology has its own layer, and there is a client/server relationship between the layers. In our case, the IP network is viewed as a client layer of the optical network server layer. More particular, we look at resilience, the ability of the network to recover from failures such as fiber cuts and power outages. Two main aspects are elaborated in this book. First, the

[1] Cisco Visual Networking Index (VNI) 2011

provisioning of high availability connections over multiple networks, for which an efficient solution was developed and evaluated. Second, we investigate the impact of certain switch designs on the ability to reduce costs through the optimization of resilience mechanisms.

1.2 Overview

The research performed during this thesis spans a number of topics, all related to resilience in future networks, and was performed in the framework of European-funded (7th Framework Programme, FP7) and national-funded (FWO) projects. The first focus was on multilayer aspects of resilience, building on the vast expertise available at the IBCN research group. Quickly; we found a need for extending these concepts towards a network scenario where reliable services are provided which are spanning multiple networks (called a multidomain scenario), such as intercontinental Virtual Private Networks (VPNs) between corporate offices. This work started with a comparison of how the known solutions for multilayer networks [1], [9] performed in a multidomain scenario [2], [4], [10], [20], [21], [23], [24], [26] and [34]. We identified that a particular solution, called common pool multilayer protection (see Chapter 3) was particularly suited to be extended towards a multidomain scenario and further optimized the solution [5], [11], [12], [22], [25] and [29], providing proof [5], [29] that the solution was generally applicable in 2-connected networks. This work is the main topic for Chapter 4.

During the work on multidomain networks, another important question was being raised. The solutions for multilayer networks tacitly assumed a lot of flexibility in the optical layer, and also assumed that reducing the load in the network drastically reduces the cost of the network. While it is possible to build fully flexible optical switches, some limited designs are being more readily deployed in current networks because of their cost and modularity (see Chapter 2). We briefly investigated the impact of this limited design on restoration mechanisms [16], and turned our focus to the effects of these node designs on the cost of the network [13], [14], [19], [27], [28] and [33]. This research is presented in Chapter 5.

During the research on these two larger aspects of survivability, we also explored some other interesting research subjects. The first is the concept of taking the physical properties op the optical network, such as certain characteristics of the switches and properties of the optical fiber, into account when deciding how the transport channel should be routed in the network. This concept is known as Impairment-Aware (IA) networking. This highly cooperative work was performed in the framework of the EU-funded DICONET[2] project, and published in [3], [6], [17], [36]. Appendix B presents the summary paper [3]. One of the control options for Impairment Aware networking was using the Path Computation Element

[2] www.diconet.eu

(briefly discussed in Chapter 2), for which we designed some protocol extensions which were tested on a proprietary (partial) PCE implementation [15].

A second interesting question was a very practical one: how do we locate a failure in a transparent network. This joint work with the University of Patras (Greece) and the Holon Institute of Technology (Israel) led to a critical analysis of the failure modes in optical networks, and identified the failures which are the most difficult to localize. The work resulted in the initial definition of a stochastic framework in which we can analyze optical network failures [30][36], Appendix A. This work shows some promising applications, and it is definitely a worthwhile topic for further investigation.

In the last years, a new technology was gaining a lot of attention: OpenFlow[3] and our group now stands at the forefront of European research regarding OpenFlow with involvements in major European projects, SPARC[4] and OFELIA[5]. We investigate resiliency in OpenFlow, resulting in co-authored publications [7], [18], [31] and [32].

Then there is one more co-authored publication on the design of a heuristic for survivable network design based on a model for the growth of a true slime mold that builds networks to transport nutrients through its body [8]. These slime molds were shown to closely emulate road and public transport networks[6] and the model was used as the basis for a heuristic to design survivable telecommunications networks.

1.3 Organization

Chapter 2 and Chapter 3 provide the fundamentals on which the work in this thesis is built. We present the technical background on optical transmission technologies in Chapter 2, along with architectures for optical switches, and we also present the network topologies used as a reference. Chapter 2 supports all other work in this thesis. Chapter 3 builds upon Chapter 2 and explains background work on multilayer survivability which are the fundamentals for our work on multidomain survivability performed in Chapter 4 and the work on the effects of transmission technologies from Chapter 5. Chapter 6 summarizes the conclusions from this work.

Two appendices present publications on two other topics: failure localization in transparent optical networks (Appendix A) and cost-effectiveness of Impairment Aware optical networking (Appendix B).

[3] www.opennetworking.org

[4] www.fp7-sparc.eu

[5] www.fp7-ofelia.eu

[6] A Tero et al., *Rules for biologically inspired network design.* Science, 327, 2010.

1.4 Publications

1.4.1 A1: Publications indexed by the ISI Web of Science ”Science Citation Index Expanded”

1. M. Pickavet, P. Demeester, D. Colle, **D. Staessens**, B. Puype, L. Depré, I. Lievens, *Recovery in Multilayer Optical Networks*, Journal of Lightwave Technology, January 2006, Vol. 24, Issue 1. (times cited: 27)

2. **D. Staessens**, D. Colle, I. Lievens, M. Pickavet, P. Demeester, W. Colitti, A. Nowé, K. Steenhaut, R. Romeral, *Enabling High Availability over Multiple Optical Networks*, IEEE Communications Magazine, June 2008, Vol. 46, Issue 6, pp. 120-126. (times cited: 5)

3. S. Azodolmolky, D. Klonidis, I. Tomkos, Y. Ye, C. V. Saradhi, C. E. Salvadori, M. Gunkel, K. Manousakis, K. Vlachos, E. Varvarigos, R. Nejabati, D. Simeonidou, M. Eiselt, J. Comellas, J. Solé-Pareta, C. Simonneau, D. Bayart, **D. Staessens**, D. Colle, M. Pickavet, *A Dynamic Impairment-Aware Networking Solution for Transparent Mesh Optical Networks*, IEEE Communications Magazine, May 2009, pp. 38-48. (times cited: 18)

4. F. Callegati, F. Cugini, P. Ghobril, S. Gunreben, V. Lopez, B. Martini, P. Pavon-Marino, M. Perenyi, N. Sengezer, **D. Staessens**, J. Szigeti, M. Tornatore, *Optical Core Networks Research in the e-Photon-ONe+ Project*, Journal of Lightwave Technology, October 2009, pp. 4415-4423.

5. **D. Staessens**, D. Colle, M. Pickavet, A. Nowé, K. Steenhaut, P. Demeester, *Optimization of Common Pool Resource Sharing in Multidomain IP-over-WDM Networks.*, Computer Communications Vol. 35, Issue 5, March 2012, pp 531-540.

6. M. Angelou, S. Azodolmolky, I. Tomkos, J. Perello, S. Spadaro, D. Careglio, K. Manousakis, P. Kokkinos, E. Varvarigos, **D. Staessens**, D. Colle, C. V. Saradhi, M. Gagnaire, Y. Ye, *Benefits of Implementing a Dynamic Impairment Aware Optical Network: Results of EU project DICONET*, IEEE Communications Magazine, Vol 50, no.8, August 2012, pp. 79-88.

7. S. Sharma, **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *OpenFlow: Meeting Carrier-Grade Recovery Requirements*, Accepted for publication in Computer Communications, Special Issue on Reliable Networks and Services.

8. M. Houbraken, S. Demeyer, **D. Staessens**, P. Audenaert, D. Colle, M. Pickavet, *Fault Tolerant Network Design Inspired by Physarum Polycephalum*, Natural Computing, 2012, DOI: 10.1007/s11047-012-9344-7.

1.4.2 P1: Proceedings included in the ISI Web of Science ”Conference Proceedings Citation Index - Science”

9. D. Colle, B. Puype, A. Groebbens, L. Depré, **D. Staessens**, I. Lievens, M. Pickavet, P. Demeester, *Recovery Strategies in Multilayer Networks*, published in Proceedings (on CD-ROM) of Asia-Pacific Optical Communications Conference, APOC 2005, Shanghai, China, 06-10 November 2005

10. **D. Staessens**, L. Depré, D. Colle, I. Lievens, M. Pickavet, P. Demeester, *A Quantitative Comparison of some Resilience Mechanisms in a Multidomain IP-over-Optical Network Environment*, published in Proceedings (on CD-ROM) of ICC2006, the IEEE International Conference on Communications, ISBN 1-4244-0355-3, Istanbul, Turkey, 11-15 June 2006

11. M. Pickavet, P. Audenaert, J. Vanhaverbeke, **D. Staessens**, D. Colle, P. Demeester, *Optimizing Reliable Multidomain Optical Routing*, published in Proceedings (on CD-ROM) of ICTON2006, the 8th International Conference on Transparent Optical Networks, ISBN 1-4244-0236-0, Nottingham, 18-22 June 2006

12. **D. Staessens**, P. Audenaert, D. Colle, I. Lievens, M. Pickavet, P. Demeester, *Survivability over Multiple GMPLS Domains*, published in Anniversary International Conference on Transparent Optical Networks (ICTON), ISBN 978-1-4244-2625-6, Athens, Greece, 22-26 June 2008, pp. 31-33

13. **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *Path protection in WSXC switched networks*, published in European Conference and Exhibition on Optical Communication (ECOC), ISBN ISBN 978-1-4244-2229-6, Brussels, Belgium, 21-25 September 2008, pp. 89-90

14. **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *Cost Efficiency of Protection in Future Transparent Networks*, Proceedings (on CD-ROM) of ICTON2009, the 11th International Conference on Transparent Optical Networks, San Miguel, Portugal, 28 June - 02 July 2009

15. **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *Dissemination of Monitoring Information in Transparent Optical Networks*, ECOC2009, the 35th European Conference on Optical Communication, Vienna, Austria, 20-24 September 2009

16. **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *Impact of Node Directionality on Restoration in Translucent Optical Networks*, ECOC2010, the 36th European Conference on Optical Communication, Torino, Italy, 19-23 September 2009

17. **D. Staessens**, M. Angelou, M. De Groote, S. Azodolmolky, D. Klonidis, S. Verbrugge, D. Colle, M. Pickavet, I. Tomkos, *Techno-Economic Analysis of a Dynamic Impairment-Aware Optical Network*, Proceedings of Optical Fiber Communication Conference and Exposition (OFC/NFOEC), and the National Fiber Optic Engineers Conference, March 2011.

18. **D. Staessens**, S. Sharma, D. Colle, M. Pickavet, P. Demeester, *Software Defined Networking: Meeting Carrier Grade Requirements (invited)*, Proceedings of the 18th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN), October 2011.

1.4.3 C1: Other international and national publications

19. **D. Staessens**, O. Audouin, S. Verbrugge, D. Colle, M. Pickavet, P. Demeester, *Assessment of Economical Interest of Transparent Switching*, published in Proceedings of ECOC2005, 31st European Conference on Optical Communications, ISBN 0-86341-546-6, Glasgow, 25-29 September 2005

20. **D. Staessens**, L. Depré, D. Colle, I. Lievens, M. Pickavet, P. Demeester, *End-to-end Recovery in Multidomain IP-over-OTN Networks*, published in Workshop on Design of Next Generation Optical Networks, Ghent, Belgium, 06 February 2006

21. R. Van Caenegem, E. Van Breusegem, B. Puype, **D. Staessens**, D. Colle, I. Lievens, M. Pickavet, P. Demeester, *All-optical Label Swapping in Novel Optical Network Architectures: A Network Recovery Perspective*, published in Workshop on Design of Next Generation Optical Networks, Ghent, Belgium, 06 February 2006

22. P. Audenaert, **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *ILP for Reliability of Multidomain Optical Links*, published in Proceedings (on CD-ROM) of the 2006 COST-ADONET Spring School: Combinatorial Optimization, Communication Networks, Budapest, Hungary, 20-24 March 2006

23. **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *An MPLS Extension for Fast Recovery from Gateway Failures*, published in Proceedings of the V Workshop in G/MPLS Networks, Girona, Spain, 30-31 March 2006

24. **D. Staessens**, B. Puype, L. Depré, I. Lievens, D. Colle, M. Pickavet, P. Demeester, *Multilayer Recovery Mechanisms in Backbone Networks*, published in Proceedings (on CD-ROM) of the 8e INFORMS Telecommunications Conference, Texas, 30 March - 01 April 2006

25. **D. Staessens**, P. Audenaert, D. Colle, I. Lievens, M. Pickavet, P. Demeester, *Providing Survivable Interdomain Connections over an Optical Backbone Network*, published in Proceedings (on CD-ROM) of the 8e INFORMS Telecommunications Conference, Texas, 30 March - 01 April 2006

26. R. Romeral, **D. Staessens**, D. Larrabeiti, M. Pickavet, P. Demeester, *End-to-end Survivable Connections in Multi-Domain GMPLS Networks*, published in Proceedings of the VI Workshop in G/MPLS Networks, ISBN 978-84-96742-20-8, Girona, Spain, 12-13 April 2007

27. **D. Staessens**, D. Colle, I. Lievens, M. Pickavet, P. Demeester, *Influence of Protection on Cost Savings in Transparent Optical Networks*, published in Proceedings (on CD-ROM) of DRCN2007, the 6th International Workshop on Design and Reliable Communication Networks, La Rochelle, France, 07-10 October 2007

28. **D. Staessens**, D. Colle, I. Lievens, M. Pickavet, P. Demeester, *Path Protection in Transparent Networks*, published in KEIO and Gent University G-COE Joint workshop for future network, Ghent, Belgium, 20-21 March 2008, pp. 17-20

29. **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *Computation of High Availability Connections in Multidomain IP-over-WDM networks*, Proceedings (on CD-ROM) of ICUMT2009, the International Conference on Ultra Modern Telecommunications, St.-Petersburg, Russian Federation, 12-14 October 2009. (best paper award)

30. **D. Staessens**, K. Manousakis, D. Colle, U. Mahlab, M. Pickavet, E. Varvarigos, P. Demeester, *Failure Localization in Transparent Optical Networks*, Proceedings (on CD-ROM) of ICUMT2010, the International Conference on Ultra Modern Telecommunications, Moscow, Russian Federation, 18-20 October 2010

31. S. Sharma, **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *Enabling Fast Failure Recovery in Openflow Networks*, Proceedings of 8th International Workshop on the Design of Reliable Communication Networks (DRCN), October 2011.

32. S. Sharma, **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *A Demonstration of Fast Failure Recovery in Software Defined Networking*, accepted for the 8th Int'l ICST Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM), June 2012.

33. **D. Staessens**, D. Colle, M. Pickavet, P. Demeester, *On the Benefits of Backup Resource Sharing in Transparent and Opaque Networks*, Accepted for publication at RNDM 2012.

1.4.4 Chapters in international publications

34. L. Wosinska, D. Colle, P. Demeester, K. Katrinis, M. Lackovic, O. Lapcevic, I. Lievens, G. Markidis, B. Mikac, M. Pickavet, B. Puype, N. Skorin-Kapov, **D. Staessens**, A. Tzanakaki, *Network resilience in future optical networks*, COST Action 291 Final Report 'Towards Digital Optical Networks, LNCS, 2009, pp. 253-284.

1.4.5 Publications in national conferences

35. **D. Staessens**, K. Manousakis, D. Colle, U. Mahlab, M. Pickavet, E. Varvarigos, P. Demeester, *Failure Localization in Optical Networks*, published in Proceedings of 10th FIrW PhD Symposium, December 2009, pp. 194-195.

36. **D. Staessens**, M. Angelou, M. De Groote, S. Azodolmolky, D. Klonidis, S. Verbrugge, D. Colle, M. Pickavet, I. Tomkos, *Techno-Economic analysis of a Dynamic Impairment-Aware Optical Network*, published in Proceedings of 12th FIrW PhD Symposium, December 2011, p. 88.

2
Future Transport Networks

2.1 Introduction

This chapter provides an overview of the general topic of this thesis: transport networks for telecommunications. First, we give a brief history of the developments which led to the main technology used in backbone networks today: fiber optics. Then we introduce a very high level conceptual view of a transport network. After these two general overviews, we delve deeper into the technologies used for optical transmission (Section 2.4) and switching (Section 2.5). Based on these technologies, optical networks are classified. We explain the difference between transparent and opaque optical networks in Section 2.6. We briefly discuss how an optical network is controlled in Section 2.7. Finally, we present a number of network topologies (Section 2.8) which are used as a reference in this thesis.

2.2 A brief history of telecommunications

We communicate primarily through the sound of our voice. Sound waves can only reach a certain distance, depending on factors such as frequency and wind direction. As the human voice has a limited output range and volume, we quickly found a need for telecommunications technology which enabled us to communicate critical information over longer distances. The earliest solutions were to increase the volume and decrease the frequency by using drums (Africa and South America) extending the reach to several kilometers. Visual cues such as smoke and fire

(North America and China) could reach over 40 km and even up to 100 km depending on the landscape and weather conditions. These beacon type visual cues were further elaborated to convey more complex messages. Around 350 BC Aeneas of Stymphalus invented an optical system similar to a telegraph using identical water clocks containing an indicated volume of water and a stick with predefined messages. A fire sign or heliograph (parabolic mirror reflecting sunlight) was used to tell when to start and stop letting the water flow from the water clock. The final water level indicates the message. In the 2nd century BC the Ancient Greek engineers Kleonexis and Dimoklitos invented the fryctoria, which used two sets of torches and a coding scheme for transmitting letters of the alphabet.

In a submission to the Royal Society in 1684, Robert Hooke outlined many practical details for implementation of a visual telegraph called a semaphore line. The semaphore is an apparatus for conveying information by means of visual signals, with towers equipped with pivoting blades, paddles or shutters, in a matrix. In 1792 the first optical telegraph system, was deployed between Lille and Paris in France by Claude Chappe. This was followed by a line from Strasbourg to Paris and eventually consisted of a network of 556 stations spanning over 4800 km and was used for military communications. The system was imitated in Europe and the U.S. The last commercial semaphore link ceased operation in Sweden in 1880 [1].

In the first decade of the 19th century Samuel Thomas von Sömmering devised an electromechanical telegraph based on an earlier design by Francisco Salvá i Campillo, but the first electromagnetic telegraph, based on studies by Alessandro Volta and Luigi Galvani in Italy, was demonstrated by Baron Pavel Lvovitch Schilling in 1832 [1]. Carl Friedrich Gauss and Wilhelm Weber built their electromagnetic telegraph in 1833 in Göttingen covering a distance of 1 km between the Observatory and the Institute of Physics. The setup consisted of a coil which could be moved up and down over the end of two magnetic steel bars. The resulting induction current was transmitted through two wires to the receiver, consisting of a galvanometer. The direction of the current could be reversed by commuting the two wires in a special switch. Therefore, Gauss and Weber chose to encode the alphabet in a binary code, using positive current and negative as the two states [2].

The first commercial electrical telegraph system was deployed in England by Sir Charles Wheatstone and Sir William Forthergill Cooke in 1839, spanning 21 kilometers of the Great Western Railway. At the same time, Samuel Finley Breeze Morse designed the American telegraph (although it is believed by many to this day to have been the scientific work of Joseph Henry) and its commercial deployment was a fact by 1844 when Morse transmitted the phrase ”What hath God wrought” between Washington DC and Baltimore MD. By 1851 the telegraph net-

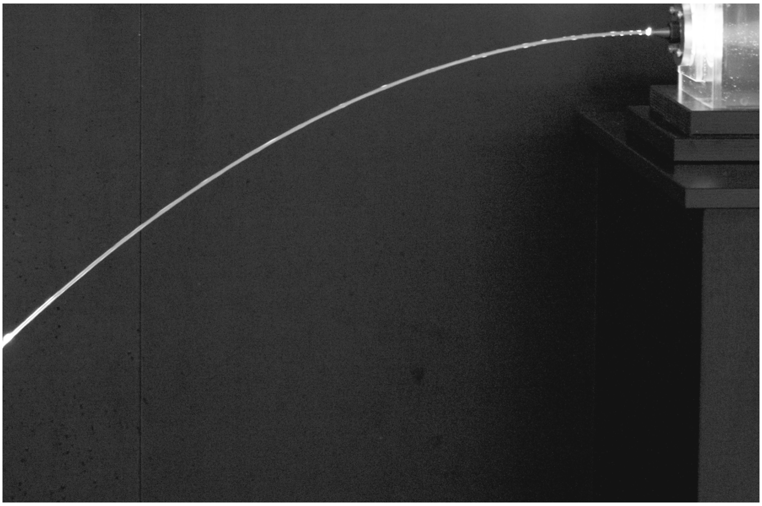


Figure 2.1: Key principle of fiber optics: total internal reflection

work covering the U.S. spanned over 32000 km. The first operational transatlantic telegraph cable was completed on July 27th 1866, after earlier failed attempts in 1857 and 1858.

In the early 1840s Daniel Colladon and Jacques Babinet demonstrated the principle of guiding light through refraction and in 1870 John Tyndall wrote in the chapter ”Total Reflexion” in an introductory book on the nature of light, about the property of total internal reflection [3]. This property is illustrated in Figure 2.1, where a beam of light is reflected inside a jet of water. The water has a higher index of refraction than the air surrounding it, which traps the light waves inside the water if they hit the water-air boundary at a shallow angle.

In August 1870, Antonio Meucci reportedly was able to capture a transmission of articulated human voice at the distance of a mile by using a copper plait as a conductor, insulated by cotton. He called this device, the ”telettrofono”. The conventional telephone was patented by Alexander Graham Bell in 1876. The first commercial telephone services were set up in 1878 and 1879 on both sides of the Atlantic in the cities of New Haven and London and by the mid 1880s every major city of the United States had a telephone exchange.

In 1880, Bell and co-inventor Charles Sumner Tainter conducted the world’s first wireless telephone call via modulated light beams projected by photophones. The scientific principles of their invention would not be utilized until the second half of the 20th century.

At the end of the 19th century, a new revolution came with wireless telegraphy. First demonstrated by Nicola Tesla in 1893, this developed into radio systems and in 1900 Reginald Fessenden was able to transmit the human voice. In December 1909 Guglielmo Marconi established wireless communication between Britain and Newfoundland, earning him a joint Nobel Prize for Physics in 1909 with Karl Braun "in recognition of their contributions to the development of wireless telegraphy."

In the next decades television was developed, depending on the Cathode Ray Tube invented by Karl Braun in 1897 and further improved using a hot cathode by John B. Johnson and Harry Weiner Weinhart of Western Electric in 1922. After mid-century the spread of coaxial cable and microwave radio relay allowed television networks to spread across large countries.

In 1952, Narinder Singh Kapany conducted experiments which led to the invention of optical fiber and in 1956 Lawrence E. Curtiss at the University of Michigan produced the first glass-clad fibers while working on the first fiber-optic semi-flexible gastroscope. In 1963 Jun-ichi Nishizawa proposed the use of optical fiber for telecommunications and later on invented graded-index fiber for transmission. Charles K. Kao and George A. Hockham of the British company Standard Telephones and Cables (STC) were the first to promote the idea that the attenuation in optical fibers could be reduced to below 20 dB/km, allowing fibers to be a practical medium for communication. They correctly and systematically theorized the light-loss properties for optical fiber, and pointed out the right material to manufacture such fibers: silica glass with high purity. This discovery led to Kao being awarded a shared Nobel Prize in Physics in 2009 "for groundbreaking achievements concerning the transmission of light in fibers for optical communication".

Around the same time as the invention of optical fiber came the scientific breakthroughs which led to the invention of the laser, based on theoretic foundations by Albert Einstein [4]. At a conference in 1959, Gordon Gould published the term LASER in the paper "The LASER, Light Amplification by Stimulated Emission of Radiation" [5]. On May 16, 1960, Theodore H. Maiman operated the first functioning (ruby crystal) laser at Hughes Research Laboratories, Malibu, California. In 1962, Robert N. Hall demonstrated the first laser diode device which emitted light at 850 nm the near-infrared band of the spectrum. These first semiconductor lasers could only operate at cryogenic temperatures. In 1970, Zhores Alferov, in the USSR, and Izuo Hayashi and Morton Panish of Bell Telephone Laboratories also independently developed room-temperature, continual-operation diode lasers. In the meantime the crucial attenuation limit of 20 dB/km in optical fiber was first achieved in 1970, by researchers Robert D. Maurer, Donald Keck, Peter C. Schultz, and Frank Zimar working for American glass maker Corning Glass Works, now Corning Incorporated. They demonstrated a fiber with 17 dB/km attenuation by doping silica glass with titanium. In the same year the first concept of Wavelength Division Multiplexing (WDM) was published. In 1973 Gerhard Bernsee of Schott Glass in Germany invented the more robust optical fiber commonly used today, which utilizes glass for both core and sheath and is therefore less prone to aging. The first semiconductor laser operating continuously at room temperature at a wavelength beyond 1 μm was demonstrated at Bell Labs in 1976.

The erbium-doped fiber amplifier (EDFA), which reduced the cost of long-distance fiber systems by reducing or eliminating optical-electrical-optical repeaters, was co-developed by teams led by David N. Payne of the University of Southampton and Emmanuel Desurvire at Bell Labs in 1986. In 1991, the emerging field of photonic crystals led to the development of photonic-crystal fiber which guides light by diffraction from a periodic structure, rather than by total internal reflection. The first photonic crystal fibers became commercially available in 2000. Photonic crystal fibers can carry higher power than conventional fibers and their wavelength-dependent properties can be manipulated to improve performance.

2.3 Networking concepts

A network consists of two basic components: nodes and links (Figure 2.2). A node (Latin nodus, ”knot”) is an active device that is attached to a network, and is capable of sending, receiving, or forwarding information over a communications channel. A link is the physical medium on which the information between 2 nodes is transmitted. A transport network establishes transmission channels (often called connections) between a set of nodes. If the connection is between two nodes, we say the information is sent along a certain path. The end nodes of this path are the terminal points for a connection and perform no other functions than sending and receiving information. The intermediate nodes receive the information on some input and have to decide on which output to forward it so it reaches its intended destination. Based on the technology used, the intermediate nodes are called routers, packet switches or circuit switches (also called cross-connects).

The key functionalities of a network are typically divided into three planes:

- Data plane. The data plane is responsible for the transmission of the raw information (the payload) in the network.
- Control plane. The control plane is responsible for the correct configuration of the data plane. It decides how the information must be forwarded in the

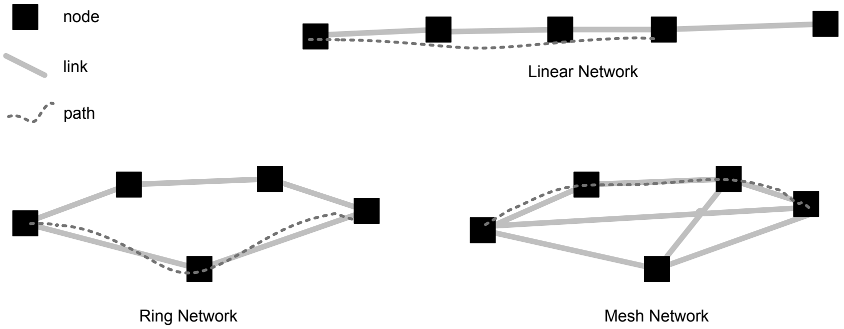


Figure 2.2: Linear, ring and mesh networks

network (routing function), performs the reservation (path setup) and release (path teardown) of required resources (signaling function) and gathers basic information, for instance the network topology (discovery function).

- Management plane. The management plane provides the interface to the network operator and allows further configuration and monitoring of the control plane and data plane equipment.

2.4 Data plane: transmission

In this section, an overview will be presented of the main optical transmission equipment used in current/future optical networks. It will present the equipment which is used for point-to-point communication in current optical networks which encompasses the source and receiver (transponders), the transmission medium (optical fiber), and various components used to cope with signal loss (amplifiers) and signal distortion(dispersion compensation).

2.4.1 Transponders / transceivers / regenerators

Both transponders and transceivers are the elements that send and receive the optical signal from a fiber. They consist of an optical receiver (photodiode) and transmitter (laser and modulator). The transmitter converts from the electrical domain to the optical domain (E/O) and the receiver from the optical domain to the electrical domain (O/E). Most WDM system manufacturers rely on transponders as input interface into the WDM system. They are available for bit rates up to 100Gb/s. Other characteristics of transponders include input/output wavelength (some have tunable wavelength), and the supported modulation format.

A regenerator is an element which restores the transmitted signal after it has become degraded after some distance due to noise, attenuation, etc. Full regen-

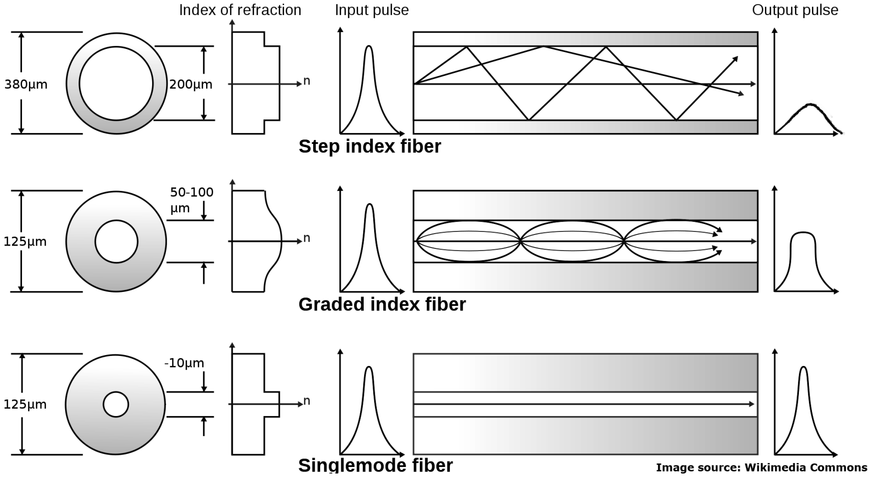


Figure 2.3: Optical fiber types

eration is performed electronically and is also known as 3R: re-amplification, re-shaping, re-timing. Regeneration is either done by connecting two transponders back-to-back or using an integrated regenerator device performing the same functionality. Due to the costly O/E/O conversion required, all-optical regeneration is under intense study.

2.4.2 Optical fibers

2.4.2.1 Transmission fibers

The optical fiber is the transmission medium for optical communications (Figure 2.3). The core has a higher refractive index and the cladding has a lower refractive index. If two materials are used, we speak of a step-index fiber, or if an array of material with slightly different refractive indexes are used, we speak of graded-index fiber. Light can travel through the fiber in various modes. The number of modes a fiber supports is dependent on the core diameter, the wavelength of the guided light and the difference between the refractive indexes of the core and cladding material. Multimode fibers have a larger core (50-200 μm) and suffer from inter modal dispersion which leads to high pulse spreading. If the diameter of the core is not too large (8-10 μm) when compared to the wavelength of the guided light and the difference in refractive index between core and cladding is also small (typically less than 1%), the fiber will only support one mode, which solves the modal dispersion problem. Such fibers are called single mode fibers (SMF) and they are the predominant fiber type used for optical fiber communications.

Fiber characteristics impact the network performance in the following man-

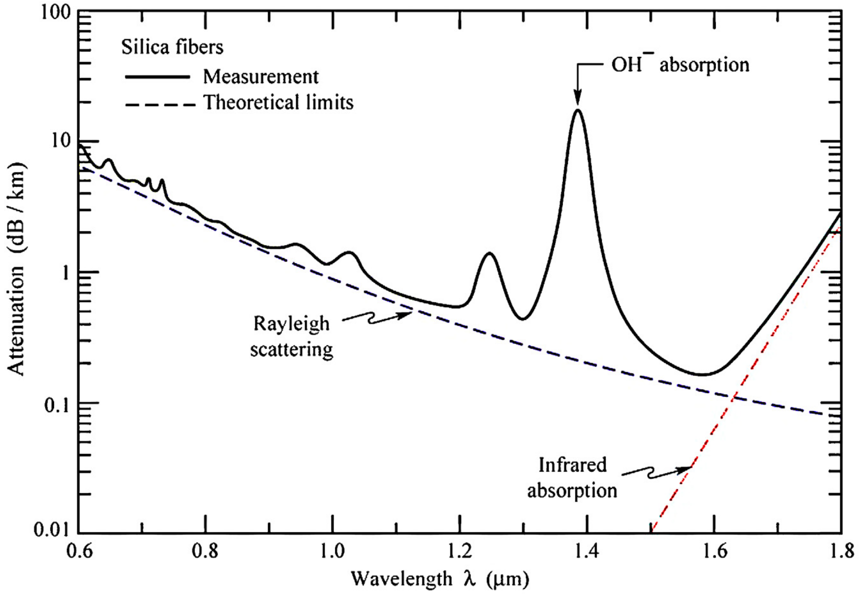


Figure 2.4: Measured attenuation in silica fibers (solid line) and theoretical limits given by Rayleigh scattering and infrared absorption

ners:

Loss characteristics Attenuation is characterized by

$$\frac{\text{Power received at distance L from transmitter}}{\text{Power transmitted}} = e^{-\alpha L} \tag{2.1}$$

where α is the attenuation coefficient in m^{-1}. Practical units of the attenuation coefficient are dB/km or dB/m. The loss versus wavelength of optical fiber is well described in the literature [6]. It determines where optical communication is practical (transmission windows). Current single mode fibers induce the lowest attenuation of roughly 0.2 dB/km at around 1550 nm as shown in Figure 2.4.

Dispersion characteristics Dispersion is the dependency of the velocity of light on the optical frequency (chromatic dispersion) or the mode of propagation (inter modal dispersion) in a wave guide. Chromatic dispersion occurs because the index of the glass varies slightly depending on the wavelength of the light, and light from real optical transmitters necessarily has nonzero spectral width (due to modulation). It causes the duration and shape of an optical pulse to change in the course of propagation. If the pulse duration widens as it travels through a fiber it can interfere with neighboring data bit pulses resulting in an inter-symbol-interference and this limits the bit spacing and the maximum transmission rate on a channel. For different types of fibers, ITU-T Recommendations G.652/3/4/5 [7] give the relevant dispersion coefficients parameters. Conventional fiber has a positive dispersion of $17\frac{\mathrm{ps}}{\mathrm{nm\,km}}$ with 0 dispersion around 1550 nm. Dispersion-shifted fibers exist with 0 dispersion at various wavelengths.

Polarization mode dispersion PMD is another optical effect that can occur in single-mode optical fibers. Single-mode fibers support two perpendicular polarizations of the original transmitted signal. If they were perfectly round and free from all stresses, both polarization modes would propagate at exactly the same speed, resulting in zero PMD. However, practical fibers are not perfect; thus, the two perpendicular polarizations may travel at different speeds and, consequently, arrive at the end of the fiber at slightly different times. The fiber is said to have a fast axis, and a slow axis. The difference in arrival times, normalized with (the square root of) length, is known as PMD and the PMD coefficient of a fiber is given in ($\frac{ps}{\sqrt{km}}$). The square root function is due to the statistical nature (random walk) of the imperfect core. Like chromatic dispersion, PMD causes digital transmitted pulses to spread out as the polarization modes arrive at their destination at different times. At 10 Gb/s, PMD is typically a problem only for long-haul systems, while at 40 Gb/s and definitely 100+ Gb/s, PMD can become a significant issue for metro/regional systems as well.

2.4.2.2 Dispersion compensation fibers

For modern silica glass optical fiber, the maximum transmission distance is limited not by attenuation but by dispersion or spreading of optical pulses as they travel along the fiber. In single-mode fiber performance is primarily limited by chromatic dispersion, which can be removed by a dispersion compensator. This works by using a specially prepared length of fiber that has the opposite dispersion to that induced by the transmission fiber, and this sharpens the pulse so that it can be correctly decoded by the electronics.

2.4.3 Optical amplifiers

Optical amplifiers are used to compensate for the loss (attenuation) in the fiber. Amplifiers are typically used at the input and/or output of switching equipment to compensate for losses, and along longer lengths of transmission fibers to compensate for the attenuation. Typical spacing of amplifiers along a fiber is in the order of 80-300 km. Two important amplifier designs for optical communications are doped fiber amplifiers and Raman amplifiers.

2.4.3.1 Doped fiber amplifiers

Doped fiber amplifiers (DFAs) are optical amplifiers that use a doped optical fiber as a gain medium to amplify an optical signal. They are related to fiber lasers. The signal to be amplified and a pump laser are multiplexed into the doped fiber, and the signal is amplified through interaction with the doping ions. The most common example is the Erbium Doped Fiber Amplifier (EDFA), where the core of a silica fiber is doped with trivalent Erbium ions and can be efficiently pumped with a laser at a wavelength of 980 nm or 1480 nm, and exhibits gain in the 1550 nm region.

2.4.3.2 Raman amplifiers

In a Raman amplifier, the signal is intensified by Raman amplification. Unlike the EDFA the amplification effect is achieved by a nonlinear interaction between the signal and a pump laser within an optical fiber. There are two types of Raman amplifier: distributed and lumped. A distributed Raman amplifier is one in which the transmission fiber is utilized as the gain medium by multiplexing a pump wavelength with signal wavelength, while a lumped Raman amplifier utilizes a dedicated, shorter length of fiber to provide amplification. In the case of a lumped Raman amplifier highly nonlinear fiber with a small core is utilized to increase the interaction between signal and pump wavelengths and thereby reduce the length of fiber required.

2.4.4 Attenuators

An attenuator is a passive optical element that reduces the input power intensity to a value given by its attenuation attribute without appreciably distorting the waveform. Optical attenuators used in fiber optic telecommunication systems may use a variety of principles for their functioning. Those using the gap-loss principle are sensitive to the modal distribution ahead of the attenuator, and should be used at or near the transmitting end, or they may introduce less loss than intended. Optical attenuators using absorptive or reflective techniques avoid this problem. The basic types of optical attenuators are fixed, step-wise variable, and continuously variable. They are for instance used to equalize the individual output levels of different wavelengths after amplification.

2.4.5 Splitters and couplers

An optical coupler is a passive optical component that is able to combine or split transmission data (optical power) from optical fibers. Its main characteristics is the splitting ratio, which is the amount of power that goes to each output. For a two-port splitter/coupler the most common splitting ratio is 50:50, though any ratio can be manufactured.

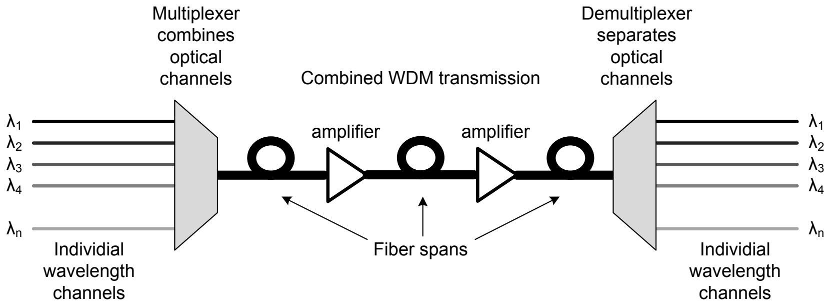

Multiplexer combines optical channels
Combined WDM transmission
Demultiplexer separates optical channels
λ1
λ2
λ3
λ4
λn
amplifier
amplifier
Fiber spans
Individial wavelength channels
Individial wavelength channels
λ1
λ2
λ3
λ4
λn

Figure 2.5: Wavelength Division Multiplexing

2.4.6 Optical mux/demux

Early optical transmission systems were point-to-point an a single carrier wavelength. However, as shown in Figure 2.4, optical fibers support efficient transmission of light in a broad range of wavelengths (called the spectrum). Wavelength-division multiplexing (WDM) is a technology which multiplexes a number of optical carrier signals onto a single optical fiber by using different wavelengths (i.e. colors) of laser light. The main components for WDM transmission are shown in Figure 2.5. A number of carrier wavelengths ($\lambda_1 \ldots \lambda_n$), each on a separate fiber, are multiplexed onto a single fiber using an optical multiplexer. The combined signal is transmitted over this single fiber, and amplified when needed (EDFAs and Raman amplifiers amplify all WDM channels on the fiber simultaneously). At the output a demultiplexer separates the individual wavelength components. Dense Wavelength Division Multiplexing (DWDM) typically supports 80 multiplexed channels (50Ghz spacing), with 160 channels (25 GHz spacing) commercially available. Channel spacing of 12.5 GHz is called Ultra-Dense WDM (UD-WDM). Wavelength channel multiplexers/demultiplexers are typically based on passive components called Arrayed Waveguide Gratings (AWG).

The fixed channel spacing for WDM has some disadvantages. For instance, a 100 Gb/s data rate signal cannot fit within a 50Ghz channel. Recent developments in WDM technology move towards so-called elastic optical networks or flexi-grid. Instead of having a fixed grid of equally spaced channels, in flexi-grid networks, the channel bandwidth is adjustable (typically in blocks of 12.5Ghz), which supports higher data rates and has more efficient spectrum usage in systems which support mixed data rates (e.g. 10Gb/s, 40Gb/s and 100Gb/s in the same equipment). Equipment for (de)multiplexing flexigrid channels is usually based on Liquid Crystal on Silicon (LCoS) technology.

2.5 Data plane: switching

Optical switches are the key devices to build optically switched networks using the transmission systems described above. An optical switch (also called optical cross-connect or OXC) is a device which has multiple input/output ports and can switch the optical signal from any input direction to any output direction. Therefore they allow building flexible networks with pure optical switching. The most popular design for optical cross-connects is based on key components called Wavelength Selective Switches (WSSs).

2.5.1 Wavelength selective switches

Wavelength selective switches (WSSs) are bidirectional devices which have 1 input/output port and a number of output/input ports from which they can demultiplex or multiplex multiple wavelengths while selecting from each input port. They are an important part of today's agile optical networks. Device size has become particularly important: a smaller WSS device physical footprint on a system circuit card could allow for more integration on a single card. Because the choice of switching engine affects the optical design, it also affects the cost and size of the WSS.

Figure 2.6 shows all the functions that are combined in MEMS (Micro Electro-Mechanical System)-based WSS's. When it is used as a reconfigurable optical demultiplexer, the WSS can steer each optical channel present on its input common port toward one of its output ports according to the wavelength of the channel. At the same time it can attenuate the optical power of this channel to a level required by the user. These functions are achieved by a single MEMS that can turn around two orthogonal axis. The first axis is used to address the different ports while the second axis enables the slight shift of the ray as compared to the nominal port position in order to induce attenuation. Of course if the WSS is used in the opposite direction, it acts as a reconfigurable optical multiplexer. The number of MEMS mirrors is equal to the number of channels that the WSS can handle individually. The commercially available WSSs feature up to 10 ports with 100 GHz or 50 GHz channel spacing. Due to the current requirement in terms of spectral efficiency, the 50 GHz channel spacing version suits more the core network applications with up to 96 channels per fiber. The typical insertion loss of such a device is between 5 dB and 7 dB whatever the number of channels handled and the number of ports. Commercially available WSSs are limited in the number of ports, usually to 10 (1x9). A commercial 20 port WSS was launched by Finisar in 2011, a commercial 24 port WSS was introduced by Oclaro at OFC 2012 and the port count of experimental devices increases rapidly, with 1x43 already demonstrated as early as 2009 [8].

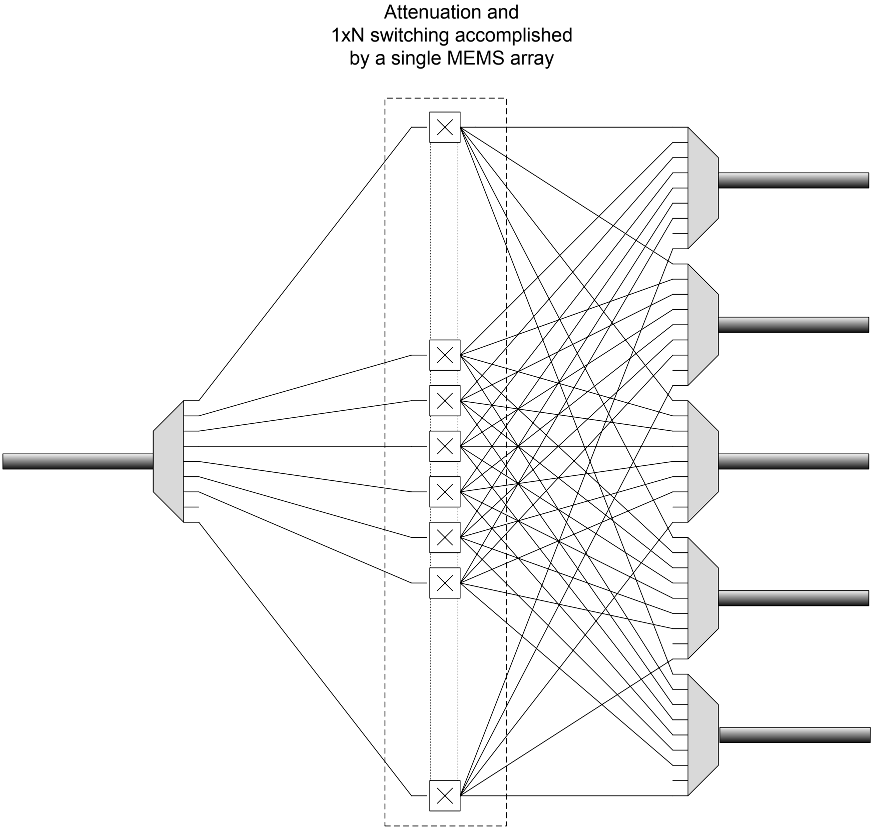


Figure 2.6: Wavelength Selective Switch

2.5.2 Optical add/drop multiplexers

The main function of DWDM was initially to increase capacity for point-to-point SONET/SDH channels and the only function needed was the multiplexing and demultiplexing of the wavelength channels at either end of the fiber. To add flexibility, DWDM networks evolved into multi node linear and ring configurations (Figure 2.2). Every node could choose to add, drop and/or continue each channel [9].

An optical add-drop multiplexer (OADM) is a device used in wavelength-division multiplexing systems for multiplexing and routing different channels of light into or out of a (single mode) fiber. Add and drop refer to the capability of the device to add one or more new wavelength channels to an existing multi-wavelength WDM signal, and/or to drop one or more channels, passing those signals to another network path. An OADM may be considered to be a specific type of optical cross-connect.

A traditional OADM consists of three stages: an optical demultiplexer, an optical multiplexer, and between them a method of reconfiguring the paths between the optical demultiplexer, the optical multiplexer and a set of ports for adding and dropping signals. The optical demultiplexer separates wavelengths in an input fiber onto ports. The reconfiguration can be achieved by a fiber patch panel or by optical switches which direct the wavelengths to the optical multiplexer or to drop ports. The optical multiplexer multiplexes the wavelength channels that are to continue on from demultipexer ports with those from the add ports, onto a single output fiber.

2.5.3 Reconfigurable optical add/drop multiplexers

If an optical channel can be flexibly added or dropped by the network operator under software control we speak of reconfigurable OADMs (ROADMs). Initially reconfigurability was restricted to selecting whether a channel was dropped or continued through the node, but later wavelength switching was added in multi-degree ROADMs (i.e. degree 3 or more). With multi-degree ROADMs, the topology is no longer restricted to a linear or ring configuration, but does now allow mesh networks (Figure 2.2). A wavelength is able to reach any adjacent node in the network through the switching function as long as transmission distance is not an issue.

2.5.4 ROADM limitations and features

ROADM designers are confronted with technical limitations imposed by technological (optics), economical (cost), administrative and regulatory (ITU-T, IETF, IEEE,...) constraints. While fully flexible 3D MEMS optical cross-connects (OXCs) are technologically feasible, they are economically less attractive due to high costs and out-competed by the pay-as-you-grow possibilities of modular ROADM architectures.

These modular ROADM architectures are fully flexible for passthrough traffic, but may exhibit some limitations with respect to the add/drop functionality. These limitations are with respect to color, contention and direction.

A ROADM architecture is defined as colorless if a wavelength can be set up under software control and is not fixed by the physical add/drop port on the ROADM. It is provided by a tunable wavelength source and by implementing an add/drop structure which is not wavelength specific. Colorless add/drop is generally created by replacing a fixed wavelength demultiplexing element (for example an arrayed waveguide grating) with a flexible demultiplexing unit (such as a wavelength selective switch).

When a wavelength can be added from any wavelength source to any output fiber and from any input fiber to any receiver (under software control) we call the

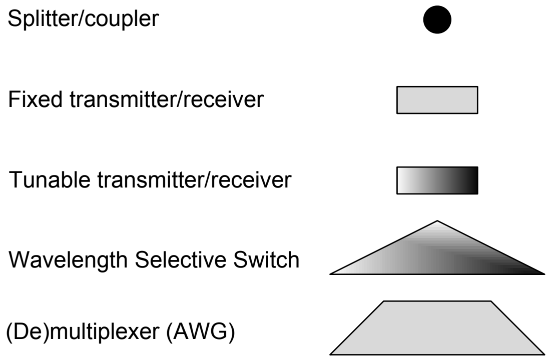

Splitter/coupler
Fixed transmitter/receiver
Tunable transmitter/receiver
Wavelength Selective Switch
(De)multiplexer (AWG)

Figure 2.7: ROADM building blocks

architecture directionless. Typically this directionless port property is realized by dedicating a transmission fiber port to a local port.

It may be that multiple demands are assigned to identical wavelengths but different transmission fiber pairs. Without any special design measures, it is an intrinsic property of ROADMs to provide only a single transmitter/receiver for each wavelength per add/drop structure. In this case - while the wavelength capacity is available on the transmission fibers - wavelength blocking may occur on the lightpath within the ROADM. This is called contention. Typically the contentionless property is realized by a spatial switch matrix or, more recently, a small fully flexible 3D MEMS block. In contrast to a WSS, this type of switch cross-connects input ports irrespective of the specific wavelength.

2.5.5 ROADM designs

In this section we will detail some ROADM designs with respect to the colorless/directionless/contentionless (CDC) features.

Figure 2.7 details the building blocks we consider for the ROADMs in this thesis. The basic components are splitters/couplers, fixed and tunable transponders, multiplexers/demultiplexers (AWG) and WSSs.

2.5.5.1 Basic architecture (directional, colored)

The colored transparent architecture (Figure 2.8) is a WSS-based all-optical cross-connect with a broadcast-and-select architecture. In this node structure fixed transponders are used. As a consequence each transponder is connected via a wavelength multiplexer/de-multiplexer (e.g. AWG) to a fixed port of the node. If a particular wavelength is not equipped in the terminal for a port, it cannot be used for add/drop at that particular port. The advantage is that there is no need for switching

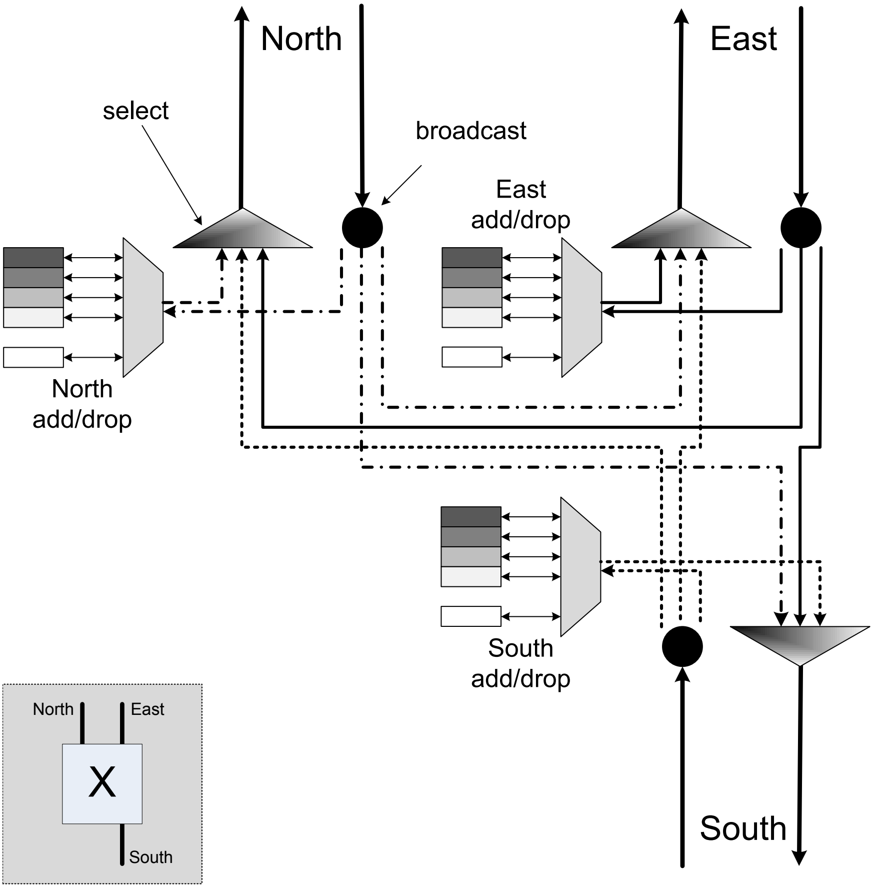


Figure 2.8: Colored, directional ROADM

equipment in the add/drop terminals. If we follow the lightpath on the incoming port for an n-degree ROADM, it is split to n directions (the $n - 1$ outputs and the drop terminal). In the drop terminal it is demultiplexed to the transponders. In the add direction, the transponder output is first aggregated through a multiplexer and then selected by a WSS to the output fiber. The WSS is used to select from which particular input port to allow passthrough of a specific wavelength on the output port. It could be replaced by a wavelength blocker/filter, which may further reduce costs, however, it seems that commercially it makes little sense as most ROADMs on the market are based on WSS.

Regeneration in ROADMs is either implemented by back-to-back interconnection of the transponders, or by replacing these back-to-back transponders by a single regenerator (as shown in Figure 2.8.

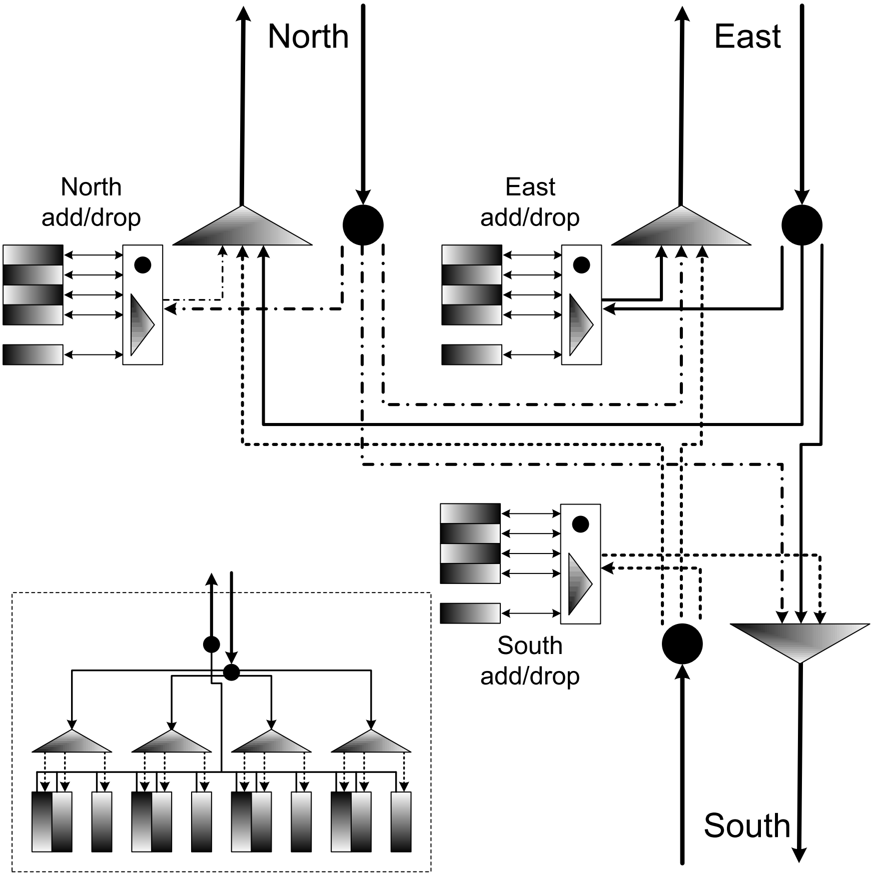


Figure 2.9: Colorless, directional ROADM

2.5.5.2 Colorless directional architecture

One of the first limitations which can be addressed is the colored add drop. Using fixed transponders can lead to higher operational costs if traffic patterns in the networks are dynamic and some lightpaths need to be rerouted or assigned to another wavelength to improve resource efficiency in the network. A colorless architecture is shown in Figure 2.9, and is basically accomplished by exchanging the fixed demultiplexer (AWGs) in an add/drop terminal with a block of splitters and WSSs. These WSSs allow the switching of any wavelength to any transponder. If the number of transponders on a terminal exceeds the number of ports on the WSSs, a configuration using additional splitters can be used (Figure 2.9 inset).

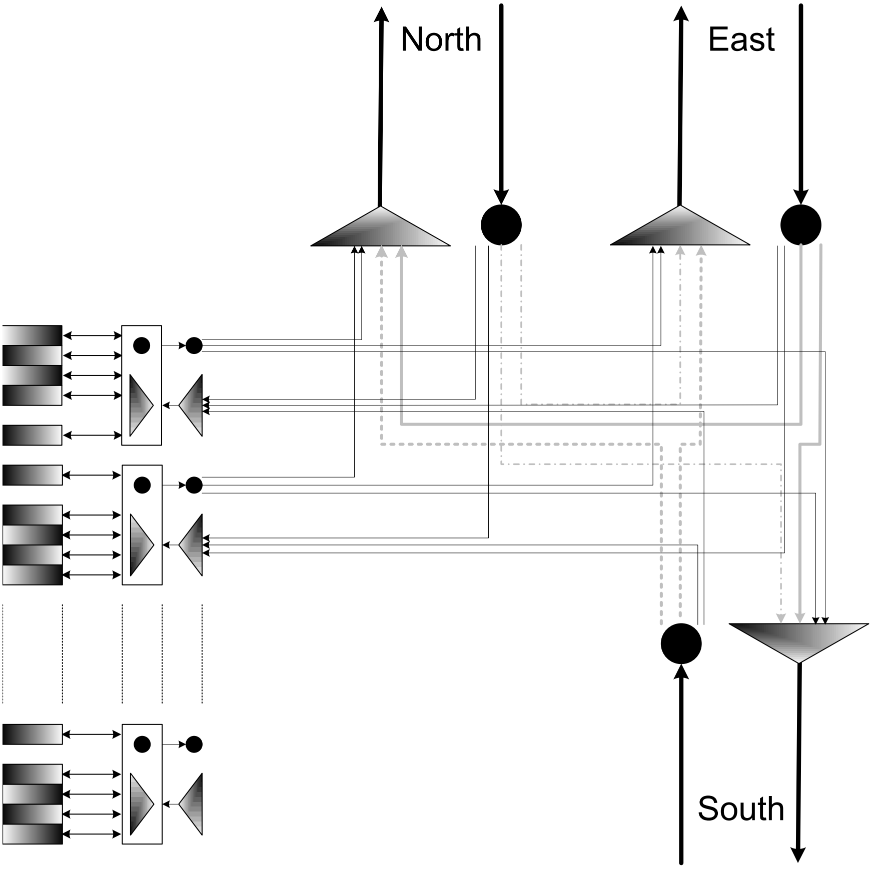


Figure 2.10: Colorless, directionless ROADM

2.5.5.3 Colorless directionless partly contentionless architecture

In order to alleviate directionality, we can use extra splitting hardware to use broadcast-and-select on the output terminals, in effect increasing the degree of the ROADM. The main advantage is increased flexibility. Using such an architecture, it is for instance possible to perform restoration, which, as expected and quantified in this thesis, has little practical use using a directional architecture (see Section 5.2).

A possible architecture is shown in Figure 2.10. The incoming traffic is split to the other directions and the add/drop terminal(s). This can be a single terminal, but in practice this means that each wavelength can only be add/dropped once at each node. This level of contention is usually unacceptable, requiring some wavelengths to be used in multiple directions. To (partly) alleviate this contention, extra add/drop terminals can be added, putting the restriction that a wavelength can be added/dropped per terminal. So a node with two terminals can terminate a specific wavelength twice, on two arbitrary interfaces.

In the terminals, WSSs are needed to allow switching from the drop side. In a directional architecture, it is impossible to have the same wavelength incoming into the terminal because it is coming from a single input fiber, however, in this directional architecture, multiple input fibers are connected to the terminal, requiring a WSS to block coalescing wavelengths.

In the add direction, the architecture is much simpler. Per terminal all transponder outputs can be coupled into a single fiber and immediately split to the output ports where a WSS performs selection of the wavelengths.

This colorless directionless and partly contentionless architecture requires a larger splitting factor on the incoming ports than the directional architectures, and also requires some extra equipment (WSSs and splitters). This means that loss/attenuation in the equipment will be higher.

2.5.6 Electronic cross-connects

In the previous section we described technologies for switching the connections in the optical domain using ROADMs. This section gives a brief description of electronic cross-connects, which switch optical signals by first converting them to electronics and then re-converting to optics. SONET/SDH and OTN are the main formats current electronic cross-connects support.

2.5.6.1 SONET/SDH

SONET (Telcordia/ANSI standard T1.105) defines optical signals and a synchronous frame structure for multiplexed digital traffic. It is a set of standards that define the rates and formats for optical networks. A similar standard, Synchronous Digital Hierarchy (SDH), is used in Europe by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T), defined in standards G.707, G.783, G.784 and G.803. SONET equipment is generally used in North America, and SDH equipment is generally accepted everywhere else in the world. Both SONET and SDH are based on a structure that has a basic frame format and speed. The frame format used by SONET is the Synchronous Transport Signal (STS), with STS-1 as the base-level signal at 51.84 Mb/s. An STS-1 frame can be carried in an OC-1 signal. The frame format used by SDH is the Synchronous Transport Module (STM), with STM-1 as the base-level signal at 155.52Mb/s. An STM-1 frame can be carried in an OC-3 signal. Both SONET and SDH have a hierarchy of signaling speeds. Multiple lower-level signals can be multiplexed to form higher-level signals. For example, three STS-1 signals can be multiplexed together to form an STS-3 signal, and four STM-1 signals multiplexed together to form an STM-4 signal. SONET and SDH are technically comparable standards. The term SONET is often used to refer to either.

2.5.6.2 OTN

The optical transport network (OTN) was created with the intention of combining the benefits of SONET/SDH technology with the bandwidth expansion capabilities offered by dense wavelength-division multiplexing (DWDM) technology and the support for optical switching. In addition to further enhancing the support for operations, administration, maintenance and provisioning (OAM&P) functions of SONET/SDH in DWDM networks, the purpose of the ITU G.709 standard (based on ITU G.872) is threefold. First, it defines the optical transport hierarchy of the OTN; second, it defines the functionality of its overhead in support of multi wavelength optical networks; and third, it defines its frame structures, bit rates and formats for mapping client signals.

At a basic level, G.709 OTN defines a frame format that encapsulates data packets, in a format quite similar to that of a SONET frame. There are six distinct layers to this format:

- OPU: Optical Channel Payload Unit. This contains the encapsulated client data, and a header describing the type of that data.

- ODU: Optical Data Unit. This level adds optical path-level monitoring, alarm indication signals and automatic protection switching.

- OTU: Optical Transport Unit. This represents a physical optical port (such as OTU2, 10Gb/s), and adds performance monitoring (for the optical layer) and the FEC (Forward Error Correction).

- OCh: Optical Channel. This represents an end-to-end optical path.

- OMS: Optical Multiplex Section. This deals with fixed wavelength DWDM (Dense Wavelength Division Multiplexing) between OADMs (Optical Add Drop Multiplexer).

- OTS: Optical Transmission Section. This deals with fixed wavelength DWDM between line amplifiers.

The OPUk, ODUk, and OTUk are in the electrical domain. The OCh is in the Optical domain. The other two layers in the Optical domain (OMS and OTS) are not currently being used.

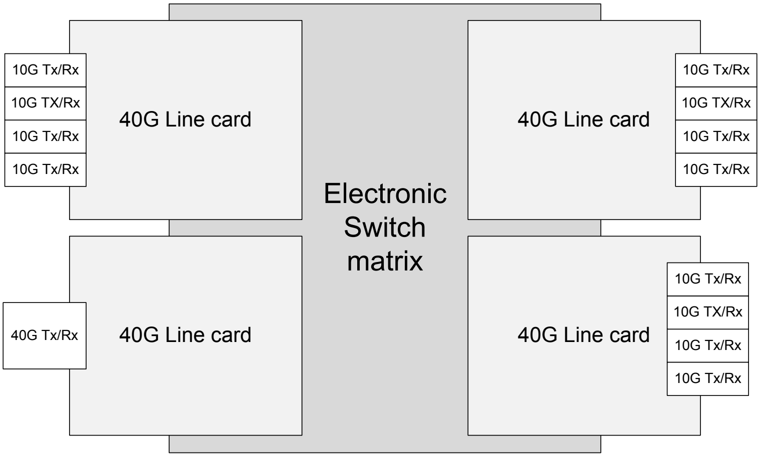


Figure 2.11: Electronic cross-connect

2.5.6.3 Electronic cross-connects

An electronic cross-connect is used to switch traffic in SONET/SDH/OTN networks and consists of three major functional components: The switch matrix (or basic node), the line cards and the transceivers (Figure 2.11). The switch matrix performs all switching functions and has a certain number of available line card slots. The line cards perform a conversion function from the transceivers to the switch fabric. This allows the switch fabric to operate independently of the protocol and support for instance 10GE (Ethernet), OTU2 (OTN) and STM-64 line cards. It is also possible for line cards to support multiple transceivers at lower data rates (for instance 4x10G transceivers in a 40G line card).

2.6 Transparent vs opaque vs. translucent networks

According to the utilization of opto-electronic conversion, three types of networks are identified: opaque, transparent, and translucent networks. An opaque network is characterized by Optical/Electronical/Optical (O/E/O) conversions for regeneration at every node. Full regeneration of a signal means re-amplification (compensating for power loss), re-shaping (compensating for distortion), re-timing (re-aligning synchronization). For instance, a network using only OTN cross-connects in each node is an opaque network. In a transparent network the signal bypasses the O/E/O devices during its transmission using for instance ROADMs. An optical signal that passes through a number of ROADMs is called a lightpath. Translucent networks are situated somewhere in between, where some paths require intermediate OEO regeneration.

One of the key issues in transparent networks is due to the increased length the signal travels without regeneration. Every amplifier adds some noise to the signal. In addition to this, longer lightpaths are sensitive to various nonlinear optical impairments, especially when considering high data rates (>10 Gb/s). This means that the signal will have to be regenerated at some point. The maximum transparent length (MTL) of a system puts a limit on the size of a completely transparent network. Another issue in transparent networks is the wavelength continuity constraint: lightpaths travel end-to-end on the same wavelength, unless there is specific equipment (called wavelength converters) present to change the wavelength from one link to another. Although all-optical wavelength converters are heavily investigated, current solutions use an electronic regenerator for wavelength conversion. Therefore transparent networks are currently all subject to the wavelength continuity constraint.

One way of dealing with the impairments in transparent networks is to introduce Islands of Transparency [10]. This is a part of the network where all possible transparent lightpaths are feasible end-to-end. Connections exiting a transparent island are regenerated.

2.7 A control plane for optical networking

Traditional optical networks were very static: optical connections were either point-to-point or at best set up manually by connecting the correct fiber to the correct ports of a static fiber patch panel. The introduction of the technologies just described allows the network to be reconfigured remotely and automatically, which led to the definition of the Automatically Switched Optical Network (ASON). The software which performs this automated network configuration is called the control plane. The control plane of choice for optical networks is Generalized Multi-Protocol Label Switching, GMPLS [11]. GMPLS evolved from MPLS, which was designed as a data-plane protocol to alleviate the high load on early IP routers but quickly evolved into a Traffic Engineering (TE) tool, bringing ATM (Asynchronous Transfer Mode)-like functionality such as recovery to IP networks. MPLS adds a label to each packet and performs switching on these labels, meaning it uses an exact match on a label to determine the output on which to forward a packet. This is much simpler to implement than routing, which performs longest prefix matching on a routing table to determine on which output to transmit a packet. Where the label in MPLS is a 20-bit header of a packet, GMPLS supports routing extensions [12] which extend this label to any datalink interface (Layer 2 Switch Capable, L2SC), electronic cross-connect interfaces (Time-Division-Multiplex Capable, TDM), optical wavelength (Lambda Switch Capable, LSC) or even an entire fiber (Fiber Switch Capable, FSC). The control plane functionality of GMPLS supports routing through the OSPF-TE [13] or ISIS-

Network	Nodes	Links	max degree
NSFNet	14	21	4
DTAG	14	23	6
Geant2	34	54	5
e1net	67	120	6

Table 2.1: Network topologies

TE [14] protocols, signaling through the RSVP-TE [15] (and CR-LDP [16]) protocols and link discovery and fault isolation through the Link Management Protocol LMP [17].

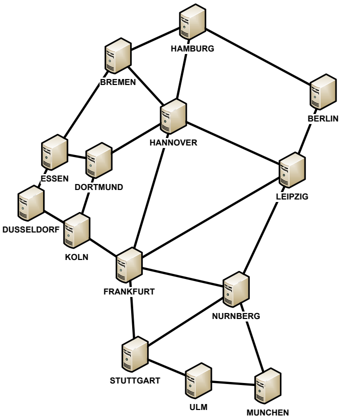
Another important development in the control of optical networks is the Path Computation Element (PCE) [18]. The PCE is an entity (node or process) that computes paths on request within its domain. Typically the routing in a GMPLS-controlled network is performed by each network node, which builds a topology database using the functionality of its (OSPF-TE) routing protocol to determine how to set up a new path in the network. In order to reduce costs and add flexibility for implementing sophisticated routing mechanisms, the PCE centralizes the routing functionality into a single entity in the network. The network switches are Path Computation Clients (PCCs) and request routing information from the PCE whenever they have to set up a new connection in the network. The PCE responds with a message containing the correct information needed to establish (signal) the path.

2.8 Reference networks

In this section we describe the reference network topologies used in this thesis. The NSFnet is a well-known American backbone topology, DTAG is based on a German nation-wide network, Geant2 is minor modification of the GEANT2 topology and e1net is a large European topology developed in the European NoE E-Photon/ONe project. These networks are summarized in Table 2.1.

2.8.1 National Topology based on the Deutsche Telekom (DTAG) Network

This network consists of 14 nodes and 23 links with an average node degree of 3.29. The topology of this network is depicted in Figure2.12 and a summary of its characteristics is given in Table 2.2. We have computed all-pair shortest path (all-pair Dijkstra) for deriving the path statistics and the metric for shortest is the total length of the path (expressed in km). The demand matrix for 2009 is projected



Node	Name	
1	Berlin	B
2	Bremen	HB
3	Dortmund	Do
4	Düsseldorf	D
5	Essen	E
6	Frankfurt/Main	F
7	Hamburg	HH
8	Hannover	H
9	Köln	K
10	Leipzig	L
11	München	M
12	Nürnberg	N
13	Stuttgart	S
14	Ulm	Ul

Figure 2.12: DTAG reference network

Parameter	Value
Number of Nodes	14
Number of links	23
Node degree	3.29 (min. 2, Max. 6)
Link length (km)	186 km (min. 37, Max:353 km)
Path length (km)	410 km (min.:37, Max.:874)
Hop count	2.35 (min:1, Max:5)

Table 2.2: DTAG topology : characteristics

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.00	8.98	12.35	13.64	9.74	32.70	19.34	21.04	14.59	33.68	15.40	12.32	23.74	11.07
2	8.98	0.00	5.76	6.23	4.51	14.19	9.92	10.56	6.59	12.59	6.43	5.13	10.15	4.69
3	12.35	5.76	0.00	12.27	10.90	21.94	11.38	13.34	12.17	17.73	9.33	7.50	15.10	6.88
4	13.64	6.23	12.27	0.00	12.52	24.58	12.48	14.31	18.02	19.54	10.42	8.33	16.96	7.68
5	9.74	4.51	10.90	12.52	0.00	17.29	8.95	10.25	10.46	13.96	7.39	5.92	11.98	5.45
6	32.70	14.19	21.94	24.58	17.29	0.00	28.99	33.09	27.13	47.75	26.20	21.64	27.56	19.88
7	19.34	9.92	11.38	12.48	8.95	28.99	0.00	20.87	13.26	26.42	13.30	10.60	20.84	9.65
8	21.04	10.56	13.34	14.31	10.25	33.09	20.87	0.00	15.16	30.04	14.81	11.94	23.42	10.79
9	14.59	6.59	12.17	18.02	10.46	27.13	13.26	15.16	0.00	20.96	11.22	8.99	18.44	8.30
10	33.68	12.59	17.73	19.54	13.96	47.75	26.42	30.04	20.96	0.00	22.38	18.38	34.50	16.09
11	15.40	6.43	9.33	10.42	7.39	26.20	13.30	14.81	11.22	22.38	0.00	10.82	20.38	10.49
12	12.32	5.13	7.50	8.33	5.92	21.64	10.60	11.94	8.99	18.38	10.82	0.00	16.32	7.82
13	23.74	10.15	15.10	16.96	11.98	27.56	20.84	23.42	18.44	34.50	20.38	16.32	0.00	17.52
14	11.07	4.69	6.88	7.68	5.45	19.88	9.65	10.79	8.30	16.09	10.49	7.82	17.52	0.00

Table 2.3: DTAG topology : traffic matrix (Gb/s)

and calculated based on the population density, considering the population growth. The overall traffic demand is summed to 2.8 Tb/s. This demand matrix is shown in Table 2.3. In this table the Node ID are the corresponding assigned number to each node name and can simply be mapped to the actual node names using the table in Figure 2.12.

2.8.2 GEANT2

This network consists of 34 nodes and 54 links with an average node degree of 3.18. It is based on the GEANT2 research network and used as a reference network for a pan-European operator. The topology of this network is depicted in Figure 2.14 and a summary of its characteristics is given in Table 2.4. We have computed all-pair shortest path (all-pair Dijkstra) for deriving the path statistics and the metric for shortest is the total length of the path (expressed in km).

2.8.3 NSFNET

The NSFNet is a pan-American network topology commonly used in literature and reflects the actual National Science Foundation Network T1 infrastructure from 1991. It has 14 nodes and 21 links and is shown in Figure 2.13.

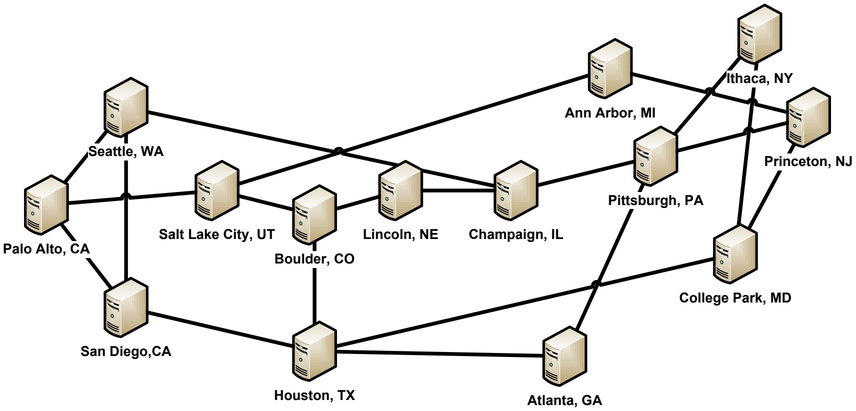


Figure 2.13: NSFNet reference network

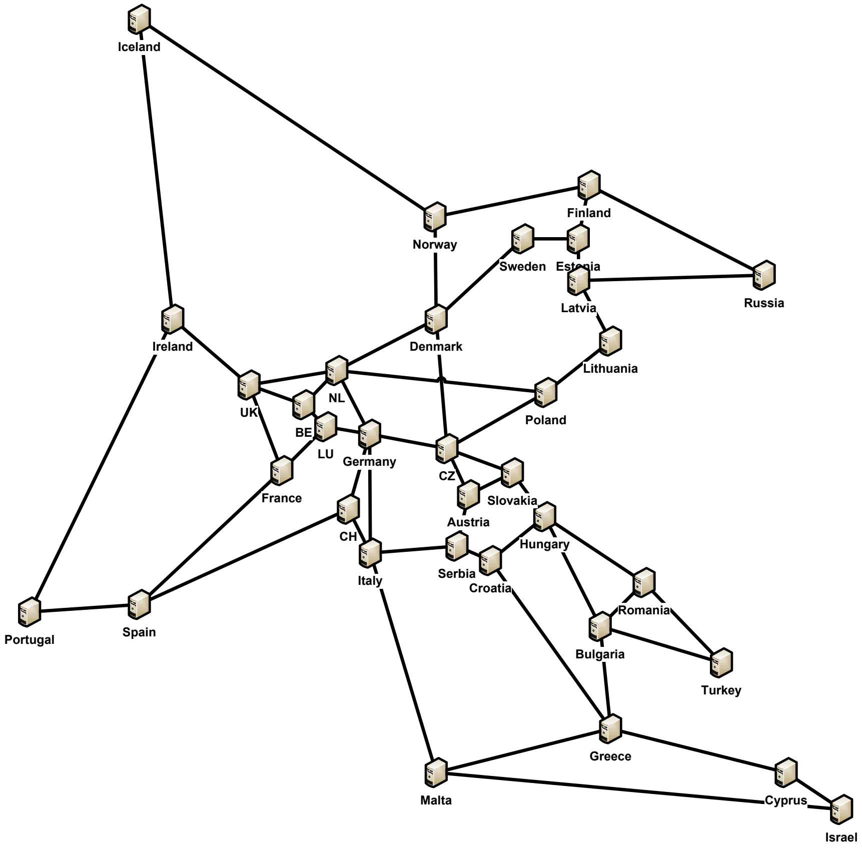


Figure 2.14: GEANT2 reference network

Parameter	Value
Number of Nodes	34
Number of links	54
Node degree	3.18 (min. 2, Max. 5)
Link length (km)	752 km (min. 67, Max:2361 km)
Path length (km)	2393 km (min.:67, Max.:7550 km)
Hop count	4.12 (min:1, Max: 11)

Table 2.4: GEANT2 topology : characteristics

References

[1] International Telecommunications Union. *50 Years of Excellence*. http://www.itu.int/itudoc/gs/promo/tsb/88192.pdf, 2006.

[2] Wikipedia. *Wikipedia: Telegraphy*. http://en.wikipedia.org/wiki/Telegraphy, retrieved November 2010.

[3] J. Tyndall et al. *Notes of a course of nine lectures on light delivered at the Royal institution of Great Britain April 8-June 3, 1869*. http://www.archive.org/download/notesofcourseofn00tyndrich, 1870.

[4] A. Einstein. *Zur Quantentheorie der Strahlung (On quantum theory of radiation)*. physikalische Zeitschrift, 1917.

[5] R. Gordon Gould. *The LASER, Light Amplification by Stimulated Emission of Radiation*. In The Ann Arbor Conference on Optical Pumping, the University of Michigan, 15 June through 18 June, 1959.

[6] E. Fred Schubert. *Light-Emitting Diodes, 2nd Ed.* 2006.

[7] ITU-T Standardization Organization. *ITU-T recommendation G.652/3/4/5*. 2009.

[8] Y. Ishii et al. *MEMS-based 143 wavelength-selective switch with flat passband*. In ECOC '09. 35th European Conference on Optical Communication, 2009.

[9] S. Gringeri et al. *Flexible Architecture for flexible transport nodes and networks*. IEEE Communications Magazine, 2010.

[10] R. D. Doverspike et al. *Future Transport Network Architectures*. IEEE Communications Magazine, Special Issue on Reliable Communication Networks, August 1999.

[11] A. Farrel and Y. Bryskin. *GMPLS: Architecture and Applications*. Elsevier, 2005.

[12] Ed. K. Kompella and Ed. Y. Rekhter. *Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS), RFC 4202*. IETF standards track, 2005.

[13] D. Katz et al. *Traffic Engineering (TE) Extensions to OSPF Version 2, RFC 3630*. IETF standards track, 2003.

[14] T. Li and H.Smit. *IS-IS Extensions for Traffic Engineering, RFC 3784*. IETF standards track, 2008.

[15] D. Awduche et al. *RSVP-TE: Extensions to RSVP for LSP Tunnels, RFC 3209*. IETF standards track, 2001.

[16] Ed. B. Jamoussi. *Constraint-Based LSP Setup using LDP, RFC 3212*. IETF standards track, 2002.

[17] Ed. J. lang. *Link Management Protocol (LMP), RFC 4204*. IETF standards track, 2005.

[18] A. Farrel et al. *A Path Computation Element (PCE)-Based Architecture, RFC 4655*. IETF standards track, 2006.

3 Recovery in multilayer networks

3.1 Introduction

This chapter gives an overview of the state-of-the art in path recovery in multilayer networks. In the previous Chapter, we focused on transport networks, and viewed them as a single layer switching circuits (light). The predominant network technology used today (commonly known as the Internet) is based on the Internet Protocol. The Internet is a packet switched network, where information is divided into small chunks and sent between routers towards its destination. Each chunk of information (think of a letter) is encapsulated into a well-defined packet (an envelope) which has a header which contains a source address and a destination address (for instance, the IP address of the Google web server is 74.125.132.94). In an IP network, each router (think of a post office) forwards the packets based on the destination address to the next router (called the next hop) until that router directly knows the destination. Transport networks are the networks (backbone) which interconnect different large routers in the Internet. We can view the IP network as a layer above the optical transport network, with a client-server relationship between them. We call such networks multilayer networks.

A small example is given in Figure 3.1. It shows a number of routers (labeled lowercase $a - e$) in the client (packet) layer and a number of cross-connects (labeled uppercase $A - E$) in the server (optical) layer. Suppose a message is being sent from router a to router d. First a will encapsulate the message into an IP packet, and labels it with destination address d (1). Based on this destination ad-

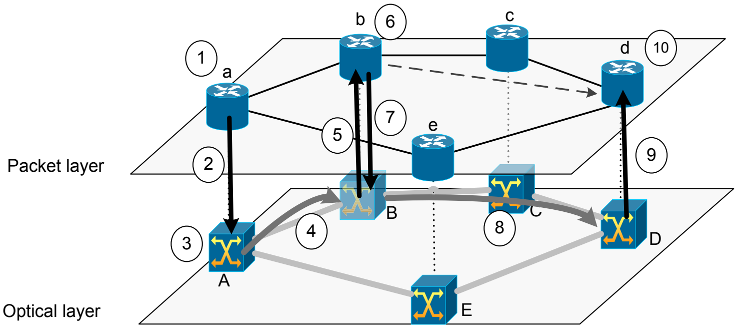


Figure 3.1: Multilayer network: IP-over-Optical transport

dress, it decides the next hop, b, and will send it out the interface which connects to b. While, from the perspective of the router a, it is directly connected to b, in reality its outgoing interface is connected to the optical cross-connect A(2). A will receive the packet and will add it to the wavelength channel (3) destined to B(4). B will know the information in this channel should go to b and thus forwards the packet to b (5). b inspects the destination and decides the next hop is d, the destination. Its output ports is again connected to B (7) which puts the packet on the wavelength channel to D. This wavelength channel is switched in the optical layer in C (8), where the channel between $B - C - D$ implements a logical link $b - d$ in the packet layer. While there is no direct connection in the optical layer, b and d are unaware of this. Finally, cross-connect D receives the information on the wavelength channel for d, and thus forwards it to d (9) where d decapsulates the message (10).

The fibers in the optical layer can carry a huge amount of data. Current data rates for a single wavelength channel are 10Gb/s, with 40Gb/s and 100Gb/s being deployed. Taken into account that a single fiber can support up to 160 channels at 10Gb/s, we are talking 1.6 Tb of data sent along this fiber every second. This is roughly the equivalent of 25 million simultaneous telephone calls. Fibers are typically packed into cables of hundreds of fibers, so the amount of traffic that will be lost should a cable be cut ensures fast failure recovery in backbone networks is paramount.

In Section 3.2, we present some formal terminology regarding recovery. We differentiate recovery techniques based on the locality: link recovery, segment recovery and end-to-end recovery. We also classify recovery techniques based on the methodology for assignment and reservation of recovery resources: restoration vs. protection. Then we turn our attention to the multilayer aspects of recovery in Section 3.3. Here we define what constitutes a multilayer network and first present strategies where only one network layer is responsible for recovery (Section 3.3.1).

We show how single-layer recovery in lower layers cannot restore traffic if the failure occurs in a higher layer, and also how a failure in a lower layer (called the root failure) can result in multiple failures in a higher layer (called secondary failures) which makes single-layer recovery in the upper layers more complex. How to resolve race conditions when deploying single-layer recovery techniques in multiple network layers at the same time is detailed in Section 3.3.2. True integrated solutions to recover traffic in multilayer networks are presented in Section 3.3.3, where we show that having a holistic view of all layers and a single multilayer recovery mechanism can drastically improve resource efficiency. The common pool strategy (Section 3.3.3.3) for multilayer recovery is the starting point for our contributions presented in Chapter 4, where we extend that solution to multi-domain networks. We conclude the current chapter with a short discussion of dynamic recovery, where the lower layer reconfigures higher layer topologies to dynamically cope with failures (Section 3.4). In the final section, we present a cost comparison for the different approaches [1]. For a full discussion of multilayer recovery, also see [2].

The concepts in this chapter also form the basis for the work in Chapter 5, where we evaluate the different protection mechanisms discussed here with respect to their resource consumption and their influence on the node cost of the network.

3.2 Network recovery

Recovery is a general term for any actions which enable a network to return to an operational state after a network failure. There is a lot of mixed use of terminology in literature. In this thesis, we will follow the terminology for GMPLS based recovery schemes as defined in RFC4427 [3].

Recovery can be applied at various levels throughout the network. A path may be subject to local (span), segment, and/or end-to-end recovery. Local (span, link) recovery refers to the recovery of a path over a link between two nodes. End-to-end recovery refers to the recovery of an entire LSP from its source (ingress node end-point) to its destination (egress node end-point). Segment recovery refers to the recovery over a portion of the network of a segment of the path. Such recovery protects against span and/or node failures over a particular portion of the network that is traversed by an end-to-end path. In general, a *recovery scheme*

- specifies resources which will carry the traffic under normal, failure-free conditions, these resources define the *working path*

- specifies alternate resources designated to carry the traffic when any of the resources along the working path fail. These resources define the *backup path(s)*.

- specifies procedures when and how to establish these paths and when and how to move the traffic between these paths.

3.2.1 Restoration

Restoration is the term that is used when the establishment of the backup path(s) is finalized after the failure has occurred. Restoration is therefore a *reactive* strategy.

Pre-planned restoration: Before failure detection and/or notification, one or more restoration paths are instantiated between the same ingress-egress node pair as the working path. Note that the restoration resources must be pre-computed, must be signaled, and may be selected a priori, but may not be reserved (cross-connected). Thus, the restoration path is not able to carry any extra-traffic. The complete establishment of the restoration path occurs only after detection and/or notification of the working path failure, and requires some additional restoration signaling.

Shared-mesh restoration: is defined as a particular case of pre-planned restoration which reduces the restoration resource requirements by allowing multiple restoration paths (initiated from distinct ingress nodes) to share common resources (including links and nodes).

Path restoration: The ingress node switches the normal traffic to an alternate path which is signaled and fully established (i.e., cross-connected) after failure detection and/or notification. The alternate path may be computed after failure detection and/or notification. The alternate path is signaled from the ingress node and may reuse the intermediate node's resources of the working path under failure condition (and may also include additional intermediate nodes.) There are two approaches for the implementation of path restoration: *hard path restoration*, which follows a break-before-make strategy, and soft path restoration which follows a make-before-break strategy.

3.2.2 Protection

Protection is the term that is used when the establishment of the backup path(s) is done before the failure has occurred. Protection is therefore a *proactive* strategy. The different protection schemes can be classified depending on the number of recovery paths that are protecting a given number of working paths. The definitions given hereafter are from the point of view of a working path that needs to be protected by a recovery scheme.

1+1 dedicated protection

One dedicated protection path protects exactly one working path, and the normal traffic is permanently duplicated at the ingress node on both the working and protection path. No extra traffic can be carried over the protection path.

1:1 dedicated protection

One specific recovery path protects exactly one specific working path, but the normal traffic is transmitted over only one path (working or recovery) at a time. Extra traffic can be transported using the recovery path resources.

1:N (N > 1) shared protection

A specific recovery is dedicated to the protection of up to N working paths. The set of working paths is explicitly identified. Extra traffic can be transported over the recovery path. All these paths must start and end at the same nodes. If all the working paths that are protected by a shared backup path are resource-disjoint, they do not share any failure probability and all of them are protected for a single failure. If more than one working path in the set of N are affected by some failure(s) at the same time, the traffic on only one of these failed LSPs/spans may be recovered over the recovery path.

M:N shared protection

This is an extension of 1:N protection where A set of M specific recovery paths protects a set of up to N specific working paths. The two sets are explicitly identified. If several working paths in the set of N are concurrently affected by some failure(s), the traffic on only M of these can be recovered.

Shared mesh protection This is a network-wide protection scheme where resources between backup paths for a pool of disjoint working paths are shared. After the failure of any of the working paths, the backup paths of the other working paths in the pool are pre-empted and therefore these working paths are left unprotected after the failure.

3.3 Multilayer aspects

A multilayer transport network can be viewed of as consisting of a stack of single-layer networks. Typically, there is a client-server relationship between the adjacent layers of this stack. Each of these network layers may have its own (single-layer) recovery schemes. As will be shown in the following sections, it is important to be able to combine recovery schemes in several layers in order to cope with the variety of possible failures in an efficient way and to benefit from the advantages of the schemes in each layer. It is worth mentioning that implementing a multilayer recovery strategy does not mean that all the recovery mechanisms will be used at every layer. As Internet traffic is continuously shifting and changing in volume over time, for instance due to diurnal traffic fluctuation and overall traffic growth, there is ongoing research towards creating optical networks with the flexibility to reconfigure transmission according to traffic demands. This requires the possibility to set up and tear down OTN layer connections that implement logical links in the higher network layer in real-time, which has led to the concept of automatically switched optical networks (ASONs [4]). In addition to allowing the network to adapt to changing traffic demands, this flexibility in setting up lightpaths on

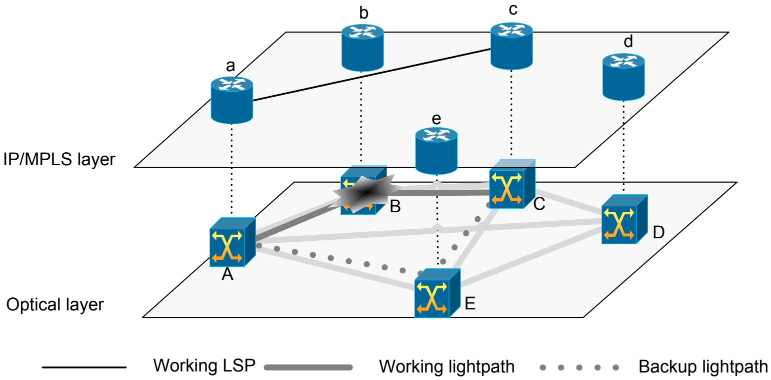


Figure 3.2: Survivability at the bottom layer

demand turns restoration into a viable recovery option.

3.3.1 Single layer recovery in multilayer networks

This section discusses the provisioning of recovery functionality in multilayer networks by starting from single-layer recovery schemes. The concepts and discussions are focused on a two-layer network, but are generic and therefore applicable to any multilayer network.

3.3.1.1 Survivability at the bottom layer

In this recovery approach, recovery of a failure is always done at the bottom layer of the multilayer network. In an IP/MPLS-over-Optical network for example, this implies that the 1+1 optical protection scheme, or any other recovery scheme which is deployed at the optical layer, attempts to restore the affected traffic in case of a failure. By recovering a failure at the bottom layer, this strategy has the benefit that only a simple root failure has to be treated, and that the number of required recovery actions is minimal (the recovery actions are performed on the coarsest granularity). In addition, failures do not need to propagate through multiple layers before triggering any recovery action.

This is illustrated on an example in Figure 3.2. The considered network carries traffic between client layer nodes a and c. The traffic flow $a-c$ uses a direct logical link from a to c, and only transits the server-layer node B. Now let's assume that a failure occurs in the bottom layer, affecting node B. In this case the optical layer will detect the failure and can recover the traffic using path $A - E - C$. In the ideal case the IP/MPLS layer will not even notice there was a failure.

However, this recovery strategy cannot handle problems that occur due to fail-

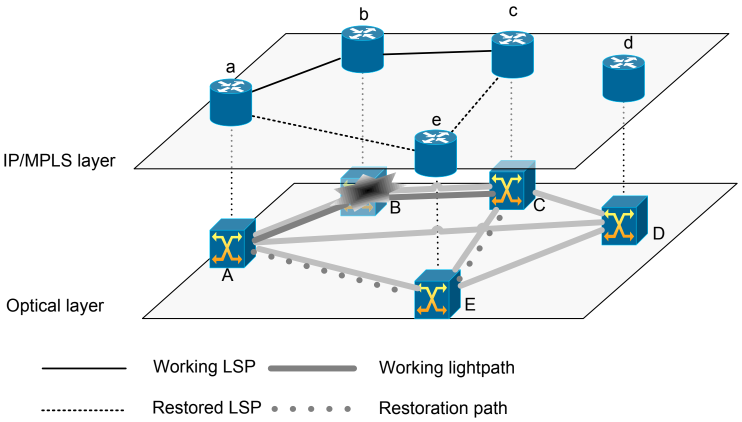


Figure 3.3: Survivability at the bottom layer, unrestorable LSP

ures in a higher network layer. If a node failure occurs in the OTN layer, the OTN layer recovery mechanism will only be able to restore the affected traffic that transits the failed bottom-layer node. The co-located higher-layer IP router will become isolated due to the failure of the OXC underneath, and thus all traffic that transits this IP router cannot be restored in the lower (optical) layer. Figure 3.3 illustrates that the server layer cannot recover the first traffic flow $a-b-c$. This is due to the fact that the client-layer node b is isolated due to the failure of B, which is terminating both logical links $a - b$ and $b - c$. This failure must be resolved at the higher layer (for instance by establishing a new connection $a - e - c$).

3.3.1.2 Survivability at the top layer

Another strategy for providing survivability in a multilayered network is to provide the survivability at the top layer of the network. The main advantage of this strategy is that it can cope with higher layer failures as well. A major drawback of this strategy, however, is that it typically requires a lot of recovery actions, due to the finer granularity of the flow entities in the top layer.

As a consequence of a single root failure in the lower layer, a complex scenario of secondary failures is typically induced in the higher network layer. This is illustrated in Figure 3.4, where the failure of the optical link $E - D$ in the bottom layer corresponds with the simultaneous failure of three logical IP links ($a - d$, $c - e$ and $d - e$) in the top layer. These three logical IP links are part of a Shared Risk Link Group (SRLG) [5]. This implies that the recovery scheme in the top layer will have to recover from three simultaneous link failures, which is quite complex. This is in clear contrast with a recovery scheme at the bottom layer, that would only have to cope with the simpler scenario of a single link failure. Another

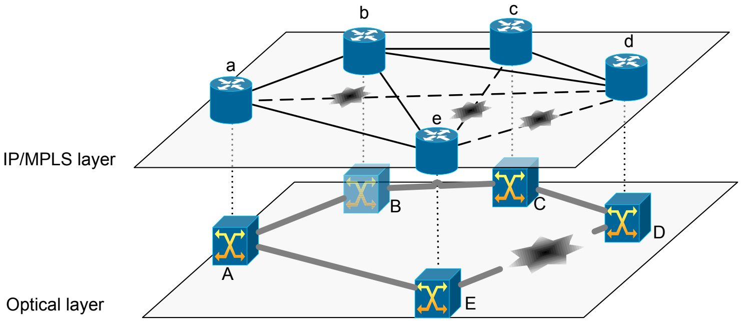


Figure 3.4: Survivability at the top layer, secondary failures

disadvantage of recovery at the top layer only is that traffic injected directly in the lower layer (e.g. wavelength channels directly leased by a customer) can not be recovered by the optical network operator, even if the failure happens in the optical layer itself.

3.3.1.3 Variants

A slightly different variant on the strategy that applies survivability at the bottom layer is the survivability at the lowest detecting layer strategy. The lowest detecting layer is the lowest layer in the layered network hierarchy that is able to detect the failure. This implies that multiple layers in the network will deploy a recovery scheme, but that the (single) layer that detects the root failure is still the only layer that takes any recovery actions. With this kind of strategy, the problem that the bottom layer recovery scheme does not detect a higher layer failure is avoided because the higher layer that detects the failure will recover the affected traffic. However, it still suffers from the fact that it cannot restore any traffic transiting higher layer equipment isolated by a node failure in the detecting layer. With this strategy the client layer in the example (Figure 3.3) could deploy a recovery scheme, but the considered traffic flow $a - b - c$ is still lost, since this client layer recovery scheme is not triggered by the occurrence of the node failure in the server layer. So, although this strategy considers the deployment of recovery schemes in multiple layers, it is still considered as a single layer survivability strategy in a multilayer network, since for each failure scenario the responsibility to recover all traffic is situated in one and only one layer (being the lowest one detecting the failure).

A slightly different variant of the strategy that provides survivability at the top layer is the survivability at the highest possible layer strategy. Since not all traffic has to be injected (by the customer) at the top layer, with this strategy a traffic flow is recovered in the layer in which it is injected, or in other words the highest possible layer for this traffic flow. This means that this highest possible layer is to be determined on a per traffic flow basis. This survivability at the highest possible layer strategy is also considered as a single layer survivability strategy for providing survivability in a multilayer network, even though it considers a recovery scheme in multiple layers. Indeed, survivability at the highest possible layer may lead to recovery schemes in multiple layers, but these will never recover the same traffic flow. Actually, this strategy deploys the survivability at the top layer strategy for each traffic flow individually. This strategy will successfully restore the traffic flow in Figure 3.3.

3.3.2 Interworking between layers

In the previous section some strategies are discussed that apply a single-layer recovery mechanism in order to provide survivability in the multilayer network. The advantages of these approaches can be combined, which implies that recovery mechanisms will run in different layers of the network as a reaction to the occurrence of one single network failure. More generally speaking, the choice in which layer(s) to recover the affected traffic due to a failure will depend on the circumstances, like, for example, which failure scenario occurred. This interworking between layers requires some rules or coordination actions in order to ensure an efficient recovery process. These rules strictly define how layers and the recovery mechanisms within those layers react to different failure scenarios, and form a so-called escalation strategy. Several escalation strategies are discussed: uncoordinated, sequential, and integrated escalation.

3.3.2.1 Uncoordinated

The easiest way of providing an escalation strategy, is to simply deploy recovery schemes in the multiple layers without any coordination at all. This will result in parallel recovery actions at distinct layers. Consider again the two-layered network (Figure 3.5), with, for instance, the failure of the physical link $A - D$ in the server layer. This failure of the physical link will also affect the corresponding logical link $a - d$ in the client layer, and hence affects the considered traffic flow $a - c$. Since the recovery actions in both layers are not coordinated, both the recovery strategies in the client and the server layer will attempt recovery of the affected traffic. This implies that in the client layer the traffic flow $a - c$ is rerouted by the recovery mechanism of the client layer, resulting in a replacement of the failed path $a - d - c$ by for instance a new path $a - b - c$. At the same time, the server layer recovers the logical link $a - d$ of the client layer topology by rerouting all traffic on the failing link $A - D$ through node E. It is clear that in this example recovery actions in a single layer would have been sufficient to restore the affected traffic.

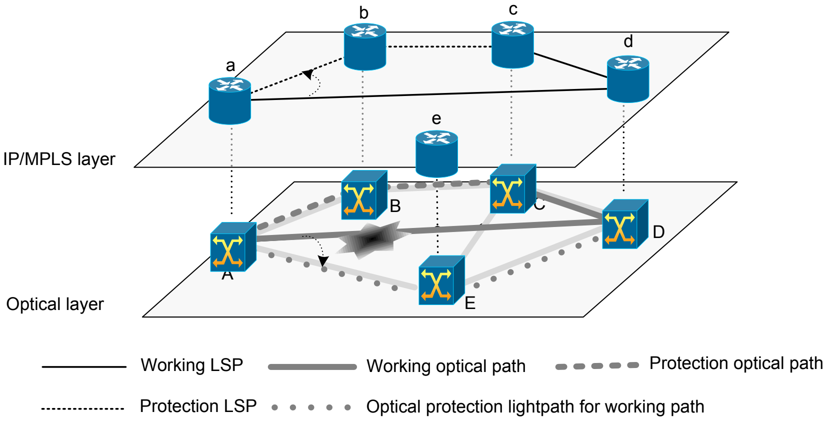


Figure 3.5: Uncoordinated approach

The main advantage of the uncoordinated approach is that this solution is simple and straightforward from an implementation and operational point of view. However, Figure 3.5 shows the drawbacks of this strategy. Both recovery mechanisms occupy spare resources during the failure, although one recovery scheme occupying spare resources would have been sufficient. This implies that more extra traffic than necessary is potentially disrupted. The situation can even be worse, consider for example that the server layer reroutes the logical link $a - d$ over the path $A - B - C - D$ instead of $A - E - D$, then both recovery mechanisms need spare capacity on the links $A - B$ and $B - C$. If these higher layer spare resources are supported as extra traffic in the lower layer, then there is a risk that these client layer spare resources are pre-empted by the recovery action in the server layer, resulting in ”destructive interference”. Or in other words, none of the two recovery actions were able to restore the traffic, since the client layer reroutes the considered flow over the path $a - b - c$, which was disrupted by the server layer recovery over $A - B - C - D$. The research done in [6] illustrates that these risks may exist in real networks: the authors prove that a switchover in the optical domain may trigger traditional client layer protection. Moreover, such a multilayer recovery strategy can have a significant impact on the overall network stability. In [7], the authors show a real life example of network convergence problems that follow the impetuous use of the uncoordinated approach in an IP-over-OTN network, where the OTN layer features 1+1 link protection. They observe IP network convergence times after the occurrence of a link failure in the OTN layer. Although protection in the optical layer recovers a link within 20 ms, the recovery of the IP traffic that was transiting the link takes over 60 s in some cases. These slow recovery times are a result of the IP layer topology discovery algorithms trying to rediscover the new IP-network topology, while the OTN layer is recovering by switching over to the backup fiber. More specific, the authors show that IS-IS adjacency recovery may take up to 13 s, IS-IS route recovery up to 18 s and, depending on the BGP scanning timing, BGP routes recovery may take up 80 s if relevant IGP topology information is lost. Note that this problem can be solved easily with a sequential approach using a hold-off timer (see next section). In summary, although simple and straightforward, just letting the recovery mechanisms in each layer run without a coordinating escalation strategy has its consequences on efficiency, capacity requirements and even ability to restore the traffic.

3.3.2.2 Sequential approach

A more efficient escalation strategy, in comparison with the uncoordinated approach, is the sequential approach. Here the responsibility for the recovery is handed over to the next layer when it is clear that the current network layer is not able to do the recovery task. For this escalation strategy two questions must be answered: in which layer to start the recovery process, and when to escalate to the next layer. Two approaches exist, the bottom-up escalation strategy and the top-down escalation approach, each having different variants.

Bottom-up escalation With this strategy, the recovery starts in the lowest detecting layer and escalates upwards. The higher layer recovery scheme will only try to recover affected traffic that could not be recovered by the lower layer. The advantage of this approach is that recovery actions are taken at the appropriate granularity: first the coarse granularities are handled, recovering as much traffic as soon as possible, and recovery actions on a finer granularity (implying in a higher layer) only have to recover a small fraction of the affected traffic. This also implies that complex secondary failures are handled only when needed. In the client-server example of Figure 3.2 for instance, there is the failure of OXC D as the root failure. This corresponds with the simultaneous failure of three IP links ($a-d$, $a-c$, and $d-c$) in the client layer. If the server layer recovery mechanism copes with the failure of OXC D, then the client layer recovery mechanism will only have to handle the recovery of the traffic over the links $a-d$ and $d-c$, being less complex than the simultaneous failure of 3 links.

This is illustrated in Figure 3.6 and Figure 3.7. The server layer starts with the recovery process, attempting to restore the logical link $a-d$. The server layer fails in this recovery since this logical link terminates on the failing node D. As such, the client layer recovery scheme is triggered (the implementation of this trigger mechanism is discussed at the end of this section) to restore the corresponding affected traffic flow $a-c$ (originally following the route $a-d-c$), by rerouting it over node b instead of node d.

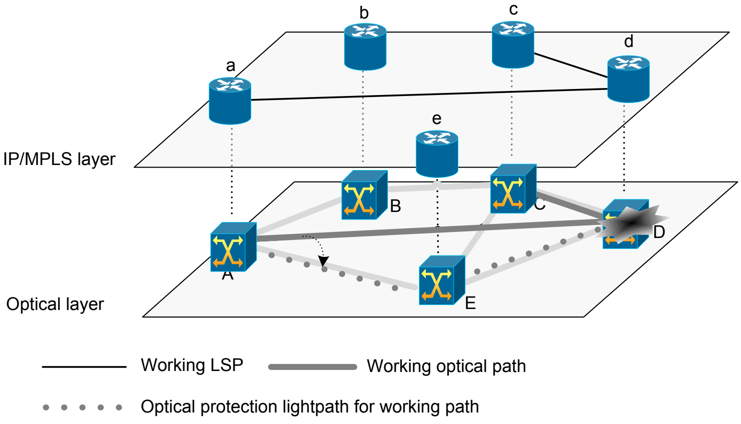


Figure 3.6: Phase 1 - Recovery action in server layer

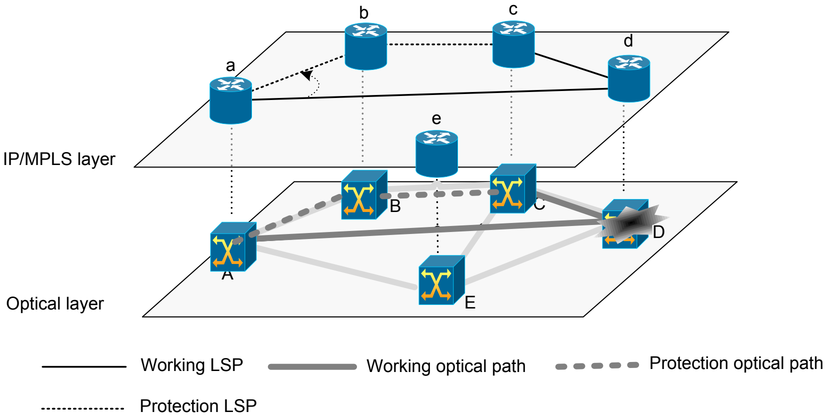


Figure 3.7: Phase 2 - recovery action in client layer

An issue that must be handled in the bottom-up escalation strategy is how a higher network layer knows whether it is the lowest layer that detects the failure (so it can start with the recovery) or has to wait for a lower layer. Typically the fault signals that are exchanged to indicate a failure will carry sufficient information, so it can be derived in which layer the failure occurred. Suppose however that this is not the case. Assume that we have a 4-layer network, where a failure occurs in the bottom layer. Assume that the failure is detected in all 4 layers at the same time, and that it cannot be derived from those signals in which layer the failure has occurred. This means that each of the higher layers can think to be the lowest-detecting layer, and start with the recovery. This can be overcome by appropriately using the mechanism of hold-off timers (see below), which are set progressively higher as we move upwards in the stack of layers. In this way, the recovery mechanisms in the higher layers will give their server layers an opportunity to do the recovery.

Top-down escalation With top-down escalation it is the other way around. Recovery actions are now initiated in the highest-possible layer, and the escalation goes downwards in the layered network. Only if the higher layer cannot restore all traffic, actions in the lower network layer are triggered. An advantage of this approach is that a higher layer can more easily differentiate traffic with respect to service types and so it can try to restore high priority traffic first. A drawback of this approach however is that a lower layer has no easy way to detect on its own, whether a higher layer was able to restore traffic (an explicit signal is needed for this purpose). So here the implementation is somewhat more complex and not currently implemented. There is also a problem of efficiency, since it is very well possible that for example 50% of the traffic carried by a wavelength channel in an optical network is already restored by a higher network layer recovery mechanism, hence protecting this wavelength in the optical layer as well is only useful for the other 50% of the carried traffic.

3.3.2.3 Implementation of an escalation strategy

The actual implementation of these escalation strategies is another issue. Two possible solutions are described here (for the ease of explanation, the bottom-up escalation strategy is assumed in what follows). A first implementation solution is based on a hold-off timer T_w. Upon detection of a failure, the server layer starts the recovery, while the recovery mechanism in the client layer has a built-in hold-off timer that must expire before initiating its recovery process. In this way, no client recovery action will be taken if the failure is resolved by the server-layer recovery mechanism before the hold-off timer expires. The main drawback of a hold-off timer is that recovery actions in a higher layer are always delayed, independent of the failure scenario. The challenge of determining the optimal value for T_w is driven by a trade-off between recovery time versus network stability and recovery performance. The second escalation implementation overcomes this delay by using a recovery token signal between layers. This means that the server layer sends the recovery token (by means of an explicit signal) to the client layer from the moment that it knows that it cannot recover (all or part of) the traffic. Upon reception of this token, the client layer recovery mechanism is initiated. This allows limiting the traffic disruption time in case the server layer is unable to do the recovery. A disadvantage, compared to the hold-off timer interworking, is that a recovery token signal needs to be included in the standardization of the interface

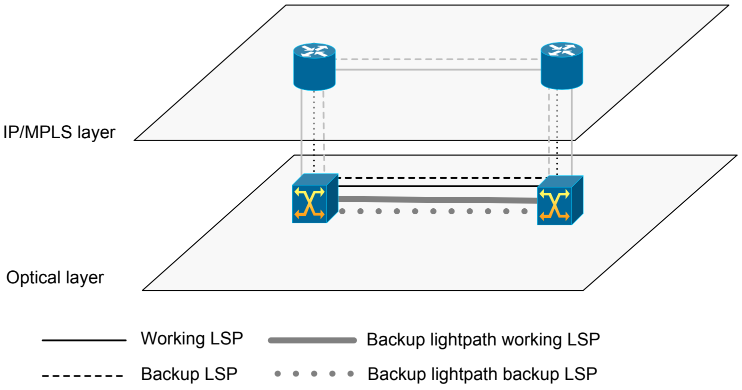


Figure 3.8: Double protection

between network layers. Note that, at the time of writing, only the timer-based approach is available in commercial networking products.

3.3.3 Multilayer survivability strategies

Multilayer survivability involves more than just coordinating the recovery actions in multiple layers. There is also the issue of the spare resources, and how they have to be provided and used in an efficient way in the different layers of the network. One way or another the logical (spare) capacity assigned to the recovery mechanisms that are deployed at higher network layers, must be transported at the lower layer. There are several ways to do this.

3.3.3.1 Double protection

The most straightforward option is called double protection, and is depicted in Figure 3.8 for one point-to-point example. Each IP link is protected both in the IP layer and in the optical layer. The spare capacity that is provisioned in the logical IP network is simply protected again in the underlying optical layer. Despite the reduced complexity, this double protection is a rather expensive solution, so investing in double protection is very debatable and probably only meaningful in a few exceptional network scenarios.

3.3.3.2 Logical spare unprotected

A first possibility to save investment in physical capacity is carrying the spare capacity in the logical higher-layer network allocated to the higher-layer network

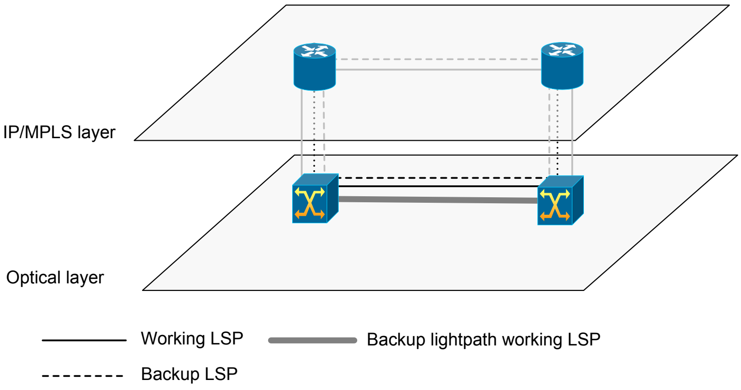


Figure 3.9: Logical spare unprotected

recovery techniques, as unprotected traffic in the underlying network layer(s) (see Figure 3.9 for the IP-over-OTN example).

This strategy, called logical spare unprotected, still allows protecting against any single failure: a cut of the bottom fiber (carrying the lightpath of the working IP link) would trigger the optical network recovery, while a failure of one of the outer router line cards would trigger the IP layer network recovery. A prerequisite for such a scenario is that the optical network supports both protected and unprotected lightpaths. It is crucial to guarantee that the unprotected spare lightpath (which carries the spare capacity of the logical higher network layer) is not affected by the failure that triggers the IP layer network recovery (that actually uses this unprotected spare lightpath). Otherwise, the spare IP capacity would also become unavailable for recovery of this failure, and the recovery process would fail.

3.3.3.3 Common Pool

One step beyond simply carrying the spare capacity of the logical higher network layers as unprotected traffic in the underlying layer is to allow pre-empting this unprotected traffic by the network recovery technique of the underlying network layer. This is the common pool strategy [8], and an example is given in Figure 3.10 for an IP-over-OTN network. The lightpath implementing the working logical IP link is optically protected. The lightpath implementing the spare logical IP link is then routed in the (optical) spare capacity which is needed to protect the aforementioned lightpath (the one that implements the working logical IP link). In case of a failure of the fiber carrying the working logical IP link, the optical protection will be triggered, pre-empting the lightpath implementing the spare logical IP

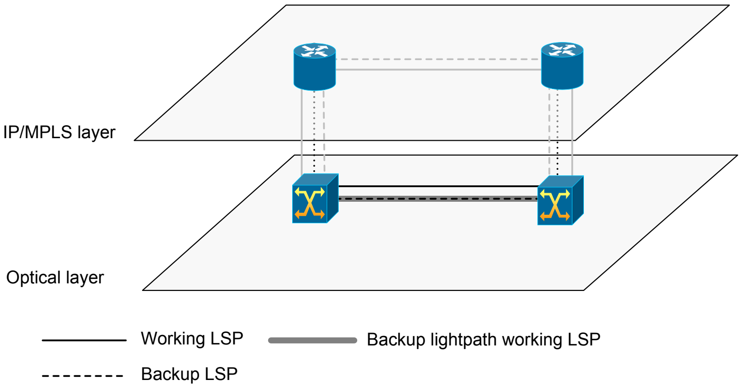


Figure 3.10: Common pool strategy

link. In that case, there is no problem in pre-empting this lightpath since it is not needed in the failure scenario. However, the pre-emption of lightpaths carrying logical spare capacity requires additional complexity. In summary, the common pool strategy provides a pool of physical spare capacity that can be used by the recovery technique in either the IP or the optical layer (but not simultaneously).

3.4 Dynamic recovery techniques

In the previous section, static multilayer recovery strategies have been discussed. They are called static, because at the time of a failure the logical network topology (in an IP-over-OTN network, this is the IP layer topology) is left unchanged (static). As such, the logical network must be provided with a recovery technique and the required spare resources in order to be able to survive failures. Dynamic multilayer survivability strategies differ from such static strategies in the sense that they actually use logical topology modification for recovery purposes. This requires the possibility to set up and tear down lower layer network connections that implement logical links in the higher network layer in real-time. Optical networks will therefore be enhanced with a control plane, which gives the client networks the possibility to initiate the set-up and tear-down of lightpaths in the optical layer. This is used to reconfigure the logical IP network in case of a network failure. This approach has the advantage that the logical network spare resources should not be established in advance in the logical IP network (at least no spare line capacities) and thus the underlying optical network should not care about how to treat these client layer spare resources. In the optical layer, however, spare capacity still has to be provided to deal with lower layer failures such as cable cuts or OXC failures.

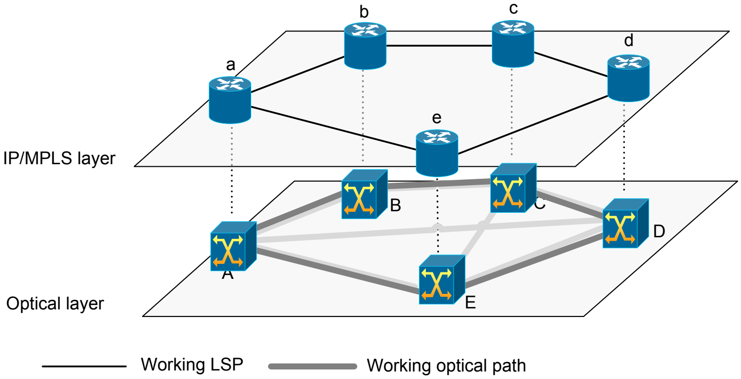


Figure 3.11: Dynamic recovery - scenario before failure

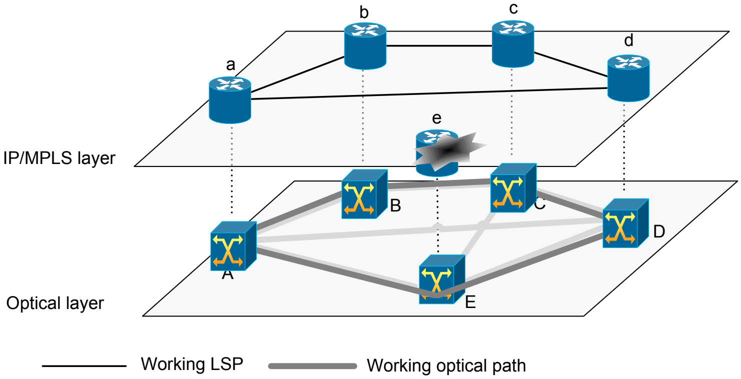


Figure 3.12: Dynamic recovery - scenario after failure

Enough capacity is also needed in the optical layer to support the reconfiguration of the logical IP network topology and the traffic routed on that topology.

An illustration of a dynamic reconfiguration of the logical higher-layer topology in case of failures is given in Figures 3.11 and 3.12 for an IP-over-OTN network. Initially, the traffic flow from router a to router c is forwarded via the intermediate router b. To this end the logical IP network contains the IP links $a - b$ and $b - c$, implemented by the lightpaths $A - B$ and $B - C$ in the OTN network. When router b fails, routers a and c will detect this failure, and use the User-Network Interface (UNI) to request the optical layer for a tear-down of the links $a - b$ and $b - c$. The resulting free capacity in the optical layer can be used to set up a direct logical IP link from router a to router c. This is requested to the underlying optical network by requesting the set-up of the lightpath between OXCs A and C. So, at the time of the failure, the logical IP network topology is reconfigured. As mentioned before, a special feature of the underlying optical network is needed for this: it must be able to provide a switched connection service to the client network quickly. Automatic Switched Optical Networks (ASONs) [4], or more generally Intelligent Optical Networks (IONs) [1], have this particular feature.

3.5 Cost comparison

Here we present a cost comparison for these schemes as a reference. This section is taken from [1]. Figure 3.13 shows the results in terms of cost (relative to the nominal failure-free situation) for the static resilience options using MPLS rerouting to protect against IP router failures and for the dynamic options using ION flexibility. In all options resilience against single optical node or link failures is provided using path protection in the optical layer. The total network cost is split in three parts: a line cost proportional to the length of the links, a node cost proportional to the number of wavelengths entering or leaving an OXC via an aggregate port, and a tributary cost for each IP router line card connected to an OXC. Figure 3.13 first of all confirms that for all strategies the optical network needs to install more capacity than for the support of the nominal logical IP network. Second, ION local reconfiguration is clearly the most cost-efficient multilayer resilience scheme. The decreasing cost trend from double protection to IP spare not protected to common pool was expected as the IP spare resources are supported more and more efficiently by the OTN resources. The higher flexibility needed to optimize the logical IP topology in each particular fault scenario in ION global reconfiguration requires a higher amount of installed capacity and equipment in the optical layer than ION local reconfiguration, making this global strategy more expensive (even as expensive as the quite inefficient static double protection strategy). The ION local reconfiguration solution is thus less expensive than the common pool one. The main cost difference lays in the tributary cost. ION local rerouting needs fewer IP router line cards, and since this equipment is relatively expensive, this equipment saving results in quite large cost savings.

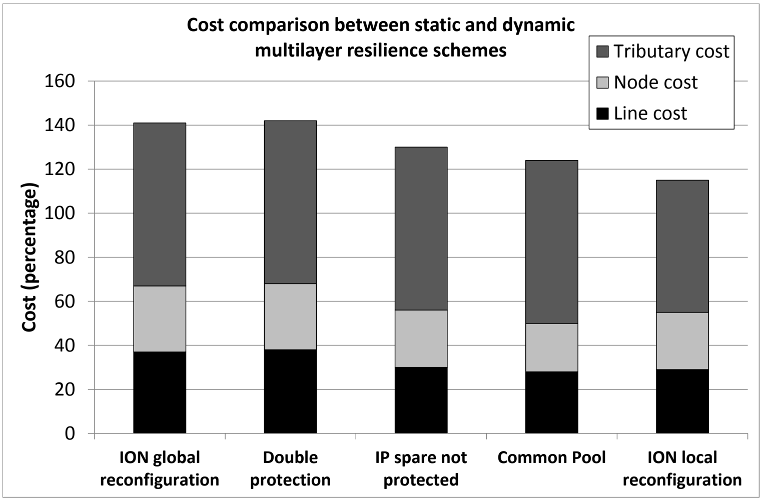


Figure 3.13: Cost comparison between static and dynamic multilayer resilience schemes. Adapted from [1]

References

[1] S. De Maesschalck et al. *Intelligent Optical Networking for Multilayer Survivability*. IEEE Communications Magazine, pp. 42-49, June 2002.

[2] J. P. Vasseur et al. *Network recovery, protection and restoration of optical, SONET-SDH, IP and MPLS*. Elsevier, 2004.

[3] E. Mannie and D. Papadimitriou. *Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS), RFC 4427*. IETF standards track, 2006.

[4] ITU-T Standardization Organization. *ITU-T Recommendation G.807/Y.1302, "Requirements for automatic switched transport networks (ASTN)*. July 2001.

[5] B. T. Doshi et al. *Optical network design and restoration. Bell Labs Tech. J. 4, pp. 58-84*. 1999.

[6] N. Wouters et al. *Survivability in a New Pan-European Carriers' Carrier Network Based on WDM and SDH Technology: Current Implementation and Future Requirements*. IEEE Communications Magazine, Aug 1999.

[7] C. Guillemot et al. *VTHD French NGI Initiative: IP and WDM Interworking with WDM Channel Protection*. In Proceedings IP-over-WDM conference, 2000.

[8] P. Demeester et al. *Resilience in multi-layer networks*. IEEE Communications Magazine, August 1998.

4 Recovery in G/MPLS controlled multidomain networks

4.1 Introduction

Up to now, most resilience mechanisms are developed for single-domain environments, which can be used more or less effectively in the current hierarchical network structure. However, in the near future peer-to-peer type network connections are expected to increase significantly, causing a flattened network structure with many networks on the same level. This means that end-to-end traffic will traverse through different networks and the end-to-end resilience can only be provided if interworking between different networks is considered in the resilience mechanism. In the field of path computation in multiple domains, most research was previously done on the computation of single traffic engineered paths [1]. Protection is a more recent topic. This chapter presents the work performed with respect to multi-domain survivability in multi-layer networks. First, in Section 4.2 we give some background on the state-of-the art in multidomain recovery and a high-level classification of different multidomain survivability options. In Section 4.3 we give a definition for what a domain is, based on the terminology defined within the Internet Engineering Task Force (IETF) for IP networks.

Section 4.4 compares a number of different approaches for multi-domain recovery with respect to resource efficiency. It shows that the common pool principle is beneficial for multidomain networks. In the following sections, we optimize this common pool solution. In Section 4.5 we clearly identify where the resource sharing is performed in the optical layer. We show, using some examples, how this solution handles intradomain failures and gateway failures. We further motivate our design choices and show that our solution has very high availability. Further, we provide mathematical proof that the solution can always be setup in any 2-connected network. Since the proof is constructive, we immediately find a heuristic solution. In the following subsection, we use Integer Linear Programming techniques to formulate an optimum solution. Simulation results show that in most cases, the heuristic performs optimally, but in the worst cases, the difference can be substantial (up to 30%). We also provide a protocol extension which allows us to set up the connection in GMPLS networks with a Path Computation Element (PCE). The PCE is an entity (node or process) that computes paths on request within its domain.

4.2 State-of-the-art

There has been some research effort towards routing [2] and protection [3] of connections spanning multiple domains. Interconnection of domains on the network layer level, using different network layer protocols and MPLS has been studied in e.g. [4] [5]. In [3], three dedicated schemes for optical protection are evaluated with regard to blocking probability of lightpath requests through interconnected optical networks. These schemes are: Basic end-to-end, Disjoint Segment and Concatenated Segment. The first scheme simply assumes total knowledge of the complete network topology. The disjoint segment scheme simply routes the primary and backup paths through disjoint domains, which requires certification that two domain-disjoint networks are also physically disjoint. The concatenated segment scheme has the backup path follow exactly the same domains as the primary path and uses a segment-like protection scheme. This approach is only link-disjoint and therefore does not cover gateway failures.

Excellent up-to-date overviews of current research focused on multidomain protection are given in [6] and [7]. The protection schemes are categorized in two large classes, being Multiple Intradomain Protection (MIDP), Hierarchical Routing with Topology Aggregation (HiTA), and a third class of more specific approaches.

Protection mechanisms in the MIDP class use intradomain methods (like PP) in each domain, and then ”stitch” them together to create the end-to-end connection, much resembling a (non-overlapping) segment protected path where each domain forms a segment.

HiTA approaches create an overlay network where each domain is represented by an aggregated topology (e.g. a star topology connecting all visible border nodes) which contains some metrics derived from the original network. On this overlay network some path computation scheme is computed (e.g. p-cycles [8]).

In HiTA schemes the path segments in each domain are usually not specifically protected.

4.3 Multidomain networks

A multidomain network consists of different independently operated subnetworks, called domains or autonomous systems. In IP networks, an Autonomous System (AS) is defined as a set of routers under a single technical administration, using some interior gateway protocol(s) (IGPs) and common metrics to route packets within the AS, and using an exterior gateway protocol (EGP) to route packets to other ASes. The use of the term Autonomous System here stresses the fact that, even when multiple IGPs and metrics are used, the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it [9]. An OSPF domain is divided into areas which are labeled with 32-bit area identifiers. Areas are logical groupings of hosts and networks, including their routers having interfaces connected to any of the included networks. Each area maintains a separate link state database whose information may be summarized towards the rest of the network by the connecting router. Thus, the topology of an area is unknown outside of the area. This reduces the amount of routing traffic between parts of an AS. Routers are classified based on their functionality. An area border router (ABR) is a router that connects one or more areas to the main backbone network. It is considered a member of all areas it is connected to. An autonomous system boundary router (ASBR) is a router that is connected to more than one routing protocol and that exchanges routing information with routers in other protocols. ASBRs typically also run an exterior routing protocol (e.g., BGP), or use static routes, or both. An internal router (IR) is a router that has neighbor relationships with interfaces in the same area. Backbone routers (BR) are all routers that are connected to the OSPF backbone. We will use the generic terms ”domain” and ”gateway” in this work.

In this work, each domain considered is a multilayer network consisting of an optical WDM transport network with an IP/MPLS layer on top. For IP networks, a domain is an AS and a gateway is an ASBR. Every node in the transport network consists of an optical cross connect (OXC) and can be equipped with a label switched router (LSR). The boundary nodes must have an LSR to forward traffic. All traffic enters and exits each domain as IP or MPLS traffic. We do not consider direct interconnection of transparent optical signals (lightpaths) between domains. Domains can have multiple entry/exit points. We require at least 2 entry and 2 exit points for each domain.

Figure 4.1 shows the terminology for recovery within multidomain networks. We label IP/MPLS nodes in lowercase and optical nodes in uppercase. To provide end-to-end recovery between a source node s and destination d within a multi-

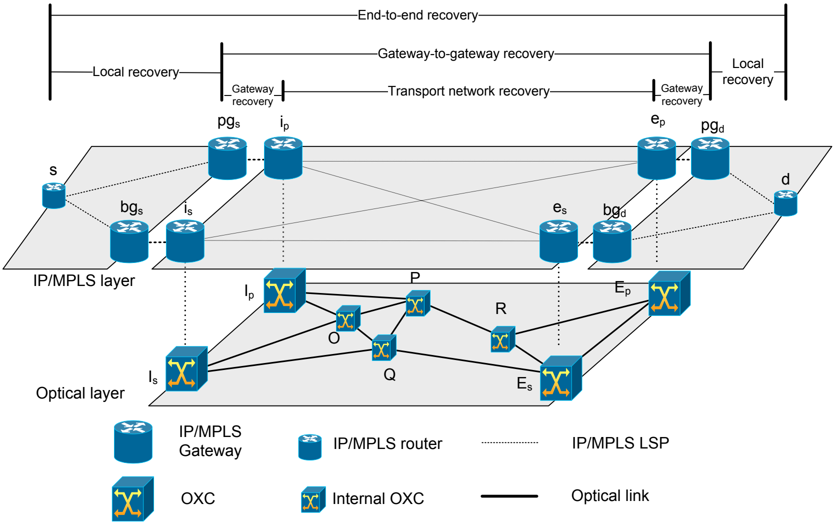


Figure 4.1: Multilayer multidomain recovery

domain multilayer environment, we can use existing recovery techniques within the different sections and layers. We need to provide proper coordination between them in order to achieve end-to-end protection and ensure interdomain connectivity in case of single network node or link failures. We can break down end-to-end recovery in three different types of sections. The first section is local network recovery in the source and destination domains, where we recover failures between the source node s and the primary gateway $pg_s$and also between the gateway pg_d and the destination d. The second section is gateway recovery where we try to recover from failures of the link connecting two gateways (e.g., between pg_s and i_p). The third section is transport network recovery, where we recovery from failures between the optical transport nodes of the gateways, e.g. I_p and E_p. Between the gateways of the source (pg_s) and destination (pg_d) domain we perform gateway-to-gateway recovery.

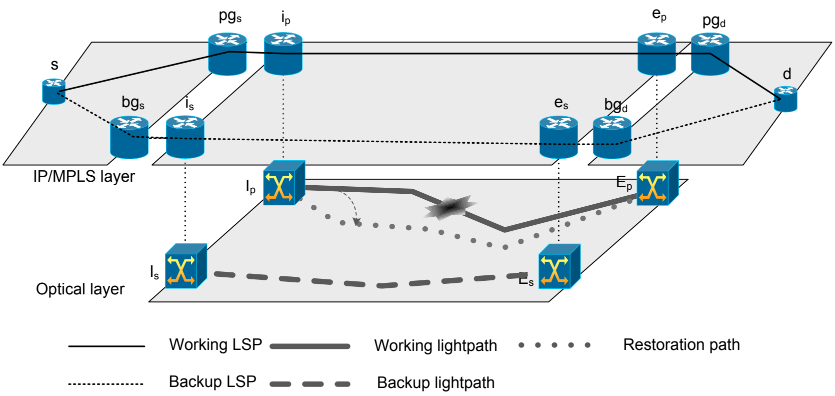


Figure 4.2: Optical restoration

4.4 Generic multidomain recovery approaches

This section shows some restoration and protection schemes for multidomain networks. It makes a comparison with regard to the capacity requirements for each solution. This work was first published in [10].

4.4.1 Restoration in multidomain networks

In this subsection, we consider dynamic restoration on an Automatically Switched Optical Network (ASON [11]). In this case, the optical layer tries to reroute all traffic among the available links and routers after a failure. These lightpaths are not protected optically. In the IP/MPLS layer there is still protection.

4.4.1.1 Restoration in the optical layer

As a first scenario, we set up two paths between each domain-pair (protection in the IP/MPLS layer), and let the ION resolve optical layer failures in any IP/MPLS connections. This will try to restore both the working lighpath and the backup lightpath in case of failures (Figure 4.2).

4.4.1.2 Restoration in the IP/MPLS layer

We can also choose to set up only one path, and, in case of a gateway failure, route the traffic over another gateway. This scenario is highly specific for multidomain survivability, since the traffic originating (from the interconnection provider point of view) in the failing gateway can in fact be recovered, which is not possible in conventional single-domain networks (Figure 4.3).

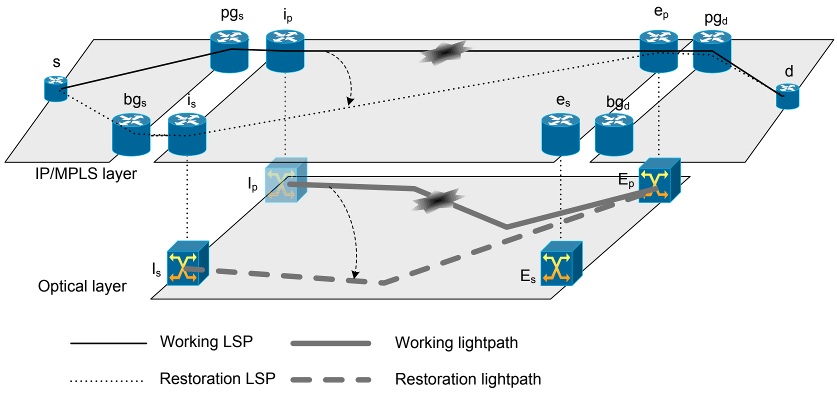


Figure 4.3: IP/MPLS Restoration

In some sense this is what currently happens with basic IP restoration. However, in order to provide this type of recovery in subsecond timescales, a reliable protocol should be developed with extensive signaling between IP domain and backbone network. In this thesis, we focus on protection.

4.4.2 Protection in multidomain networks

This section discusses the provisioning of end-to-end recovery functionality in multilayer multidomain networks by starting from an IP/MPLS-layer-only protection scheme and then introducing optical protection. Note that, in order to ensure connectivity, each domain needs at least two gateways to serve as entry-points into the intermediate networks. In case of a failure of one of these gateways, another is available to take over. After the IP/MPLS-only recovery, improvements towards recovery times and capacity requirements are proposed.

4.4.2.1 No optical protection

We interconnect the different domains using two disjoint IP/MPLS LSPs between two distinct gateways in each domain-pair, which will be referred to as the working and backup LSP. In this scenario, the IP connections are implemented as unprotected node-disjoint lightpaths in the optical domain (Figure 4.4). This scheme provides full protection against single node or link failures in the optical backbone network, but requires an MPLS capability in each IP domain in order to switch over from the working connection to the backup connection in case of a failure in the working lightpath. In absence of MPLS, the IP layer will have to converge to the new topology without the working connection, which will have some down-time as a result. From the point of view of the backbone operator, any single

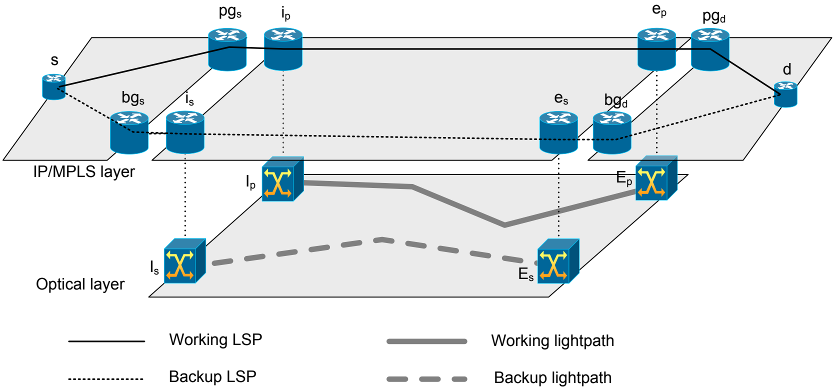


Figure 4.4: No optical protection

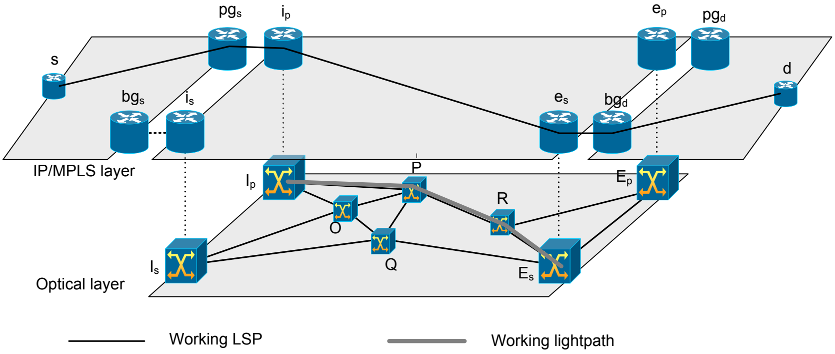


Figure 4.5: Trap topology

node failure will affect 50% of the working connections between different domains (clients). So in absence of MPLS in the IP domains, this scenario will not be the best option. In case of the setup of two IP connections between a domain-pair, the choice of which gateway to connect to another gateway is not arbitrary when a node disjoint path is required for protection. This is illustrated in Figure 4.5: there is no node-disjoint implementation of IP connections $i_p - e_s$ and $i_s - e_p$, however, if we connect $i_p - e_p$ and $i_s - e_s$, node disjoint lightpaths can be implemented in the optical domain. Topologies where such problems can occur are often called *trap topologies*.

This restriction has a serious impact on how multidomain protection should be realized. Since network operators are not keen on disclosing details about their networks, there are two options: a mechanism which precomputes the correct gateways (for instance using PCE-PCE communication between different domains) or some backtracking mechanism.

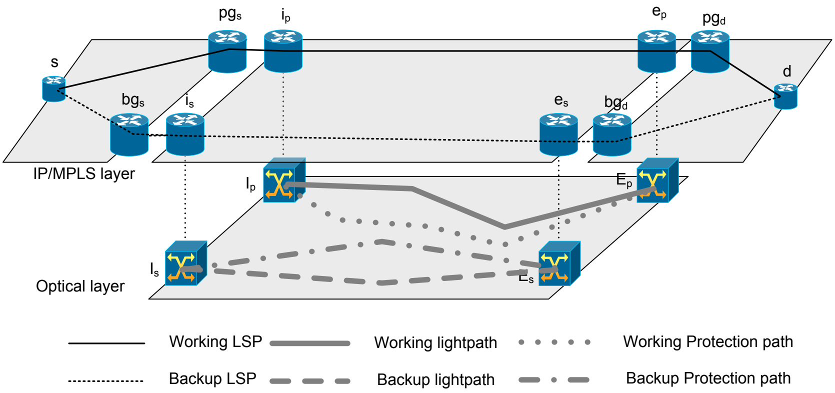


Figure 4.6: Double optical protection

4.4.2.2 Protecting both connections optically

In order to provide a more robust domain interconnection, and resolve the issue of down-time during IP-connection switchovers, the backbone operator can choose to protect each lightpath optically. The optical protection lightpaths for the working and backup lightpath will be called primary and secondary backup lightpath respectively. It should be kept in mind that in all of the following scenarios, the working and backup lightpath are not necessarily link - or node-disjoint.

Dedicated double protection The working and backup lightpath can be implemented as two shortest paths between two distinct gateways, and protected with optical 1+1 path protection. In this case, the connections between two domains are protected twice, in the IP/MPLS domain from the multidomain point of view, and in the optical layer of the interconnection domain(s). This scheme has the upside that, in case of a non-gateway failure, the IP/MPLS links are not disturbed, so only a critical gateway failure (or a failure of the interface between the gateway OXC and IP gateway) will lead to potential downtime in the absence of an MPLS control plane. As a downside we may expect a lot of capacity overhead required in the backbone network (Figure 4.6).

Shared double protection In order to reduce the required network capacity, the operator can apply capacity sharing [12] between both backup lightpaths (Figure 4.7) and between backup lightpaths of different domain-pairs (not shown). A downside is that it's not possible to do 1+1 protection of the data by simultaneously sending it over both the primary and secondary path.

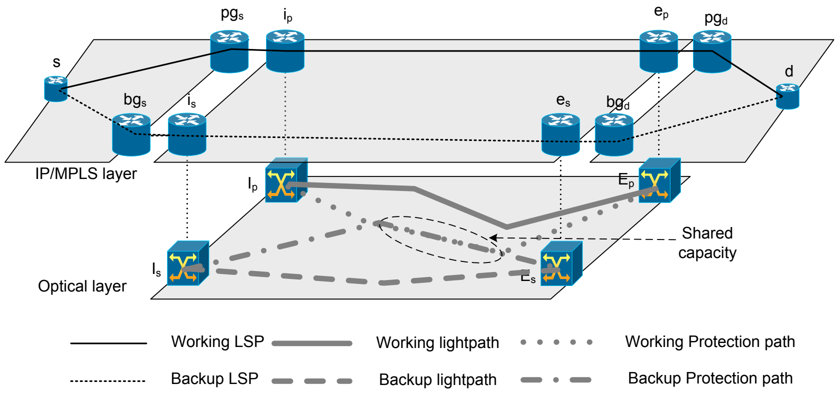


Figure 4.7: Shared double optical protection

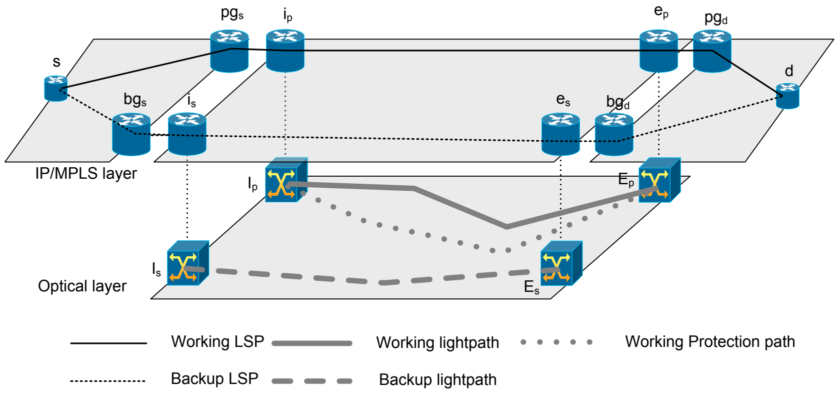


Figure 4.8: Protecting only the working connection

4.4.2.3 Protecting only the working connection

In the previous options, the choice between working connection and backup connection is somewhat arbitrary. In fact, it is possible to divide the traffic over both connections, and have the same level of protection. If we make a more formal choice between working and backup connection, then there is no need to protect the backup lightpath optically, since a failure along its path will not disturb any working connections. If only the working path is protected optically the operators should make concrete decisions on which gateways are primary and which routers are used as backup for every inter-domain connection. The reason why routers should be declared for every inter-domain connection is that a gateway must be able to serve as primary gateway for some connection while serving as backup for another.

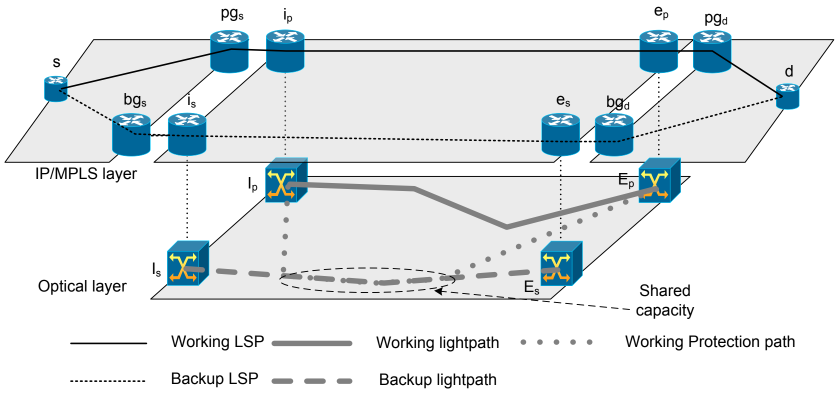


Figure 4.9: Common pool multidomain protection

4.4.2.4 Common pool multidomain protection

It is possible to further reduce capacity requirements by allowing the primary backup lightpath to preempt the backup lightpath (Figure 4.9). This is because the backup lightpath will only be used when an unrecoverable failure in the working lightpath occurs, i.e. a gateway failure. In case of a failure in the working lightpath, the backup IP/MPLS connection is torn down, but the working IP connection remains intact. This is referred to as common pool capacity sharing (See also section 3.3.3) [13].

4.4.3 Capacity requirements

When we compare the total capacity requirements against each other, we see that optical protection will always require extra capacity in the backbone network (Figure 4.10). We have normalized the results versus the optically unprotected scheme. The higher we share backup capacity between backup paths and the more intelligently we provide protection (e.g. by not protecting backup connections optically), the less capacity we need in order to ensure connectivity between different domains. We have differentiated, where possible, the required optical capacity in 5 categories, namely Working Connection Working Capacity (WCWC), Backup Connection Working Capacity (BCWC), Working Connection Backup Capacity (WCBC), Backup Connection Backup Capacity (BCBC) and Shared Capacity (SHDC). Taking a closer look at Figure 4.10, we can clearly see the gain of sharing capacity. In the unprotected scheme, we only have two dedicated lightpaths. When we apply dedicated optical protection to these, we gain a little capacity, because both paths do not need to be disjoint. However, protecting both paths is very costly. When we apply sharing between the backup lightpaths, we again gain a lot

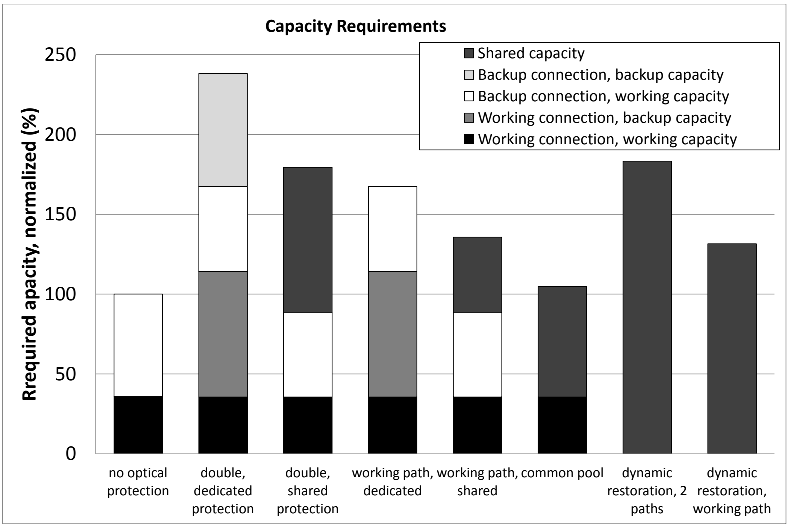


Figure 4.10: Capacity requirement for different multidomain solutions

of capacity. The SHDC part of the third option can route the same paths as WCBC and BCBC from the second option, however, not both at the same time. This is our most efficient symmetrical scheme. Note that in schemes 2-5, the WCWP and BCWC bars are exactly the same (for no optical protection, the working paths are slightly longer on average). When we abandon symmetry and explicitly specify which connection is to be used as working connection, we can refrain from protecting the backup path, again resulting in a capacity gain. Note that both bars in the second and fourth option are exactly the same, except, of course, for the BCBC part. Now, sharing between the WCBCs of different domain-pairs reduces the total capacity requirement again, and letting the WCBC use the BCWC, by preempting the backup path, leaves us with the most efficient scheme regarding capacity-requirement, requiring a mere 7% extra capacity when compared to providing no optical protection.

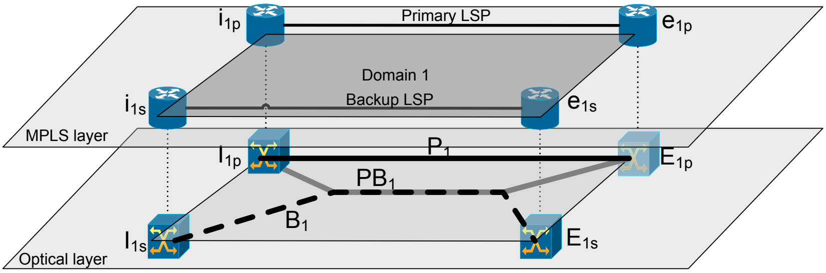


Figure 4.11: Domain structure

4.5 Common pool multidomain protection

In this section, we show how the connection is implemented in the MPLS layer using two LSPs, and then we focus on the implementation of the LSP-links in the optical layer using lightpaths.

4.5.1 The MPLS layer

The end-to-end connection is protected: we have two disjoint LSPs which we call the primary LSP and the backup LSP respectively. Both the primary LSP and the backup LSP run through the same domains in the same order. This is why our solution falls in the MIDP category. This immediately sets a requirement for at least 2 ingress nodes and 2 egress nodes in each domain. These nodes are collectively called the gateways. The structure of a domain in the network is shown in Figure 4.11.

In each Domain δ, the primary LSP runs through the primary ingress gateway $i_{\delta p}$ and the primary egress gateway $e_{\delta p}$, the backup LSP uses the secondary ingress/egress gateways $i_{\delta s}$ and $e_{\delta s}$. Moreover, both LSPs bridge each domain in a single hop.

4.5.2 The optical layer

The primary LSP is protected in the optical layer of each domain. Each Domain δ implements the $I_{\delta p} - E_{\delta p}$ LSP using optical path protection: there are two disjoint lightpaths, the primary lightpath P_δ and it's backup PB_δ, between the OXCs of the primary gateways.

In each domain, the lightpath B_δ implementing the backup LSP between $I_{\delta s}$ and $E_{\delta s}$ is left unprotected, and tries to share as much resources as possible with the lightpaths P_δ and PB_δ (not shown in Figure 4.11). B cannot run through the OXCs $I_{\delta p}$ and $E_{\delta p}$.

In a global view of the optical layer, the implementation of the end-to-end primary LSP looks like a segment protected lightpath (each domain being a segment), the implementation of the backup LSP is a long, unprotected lightpath. The backup LSP effectively protects against failures of the primary gateway nodes (due to the common pool sharing) [14].

4.5.3 Failure scenarios

In this section we will show the behavior of the paths in different failure scenarios.

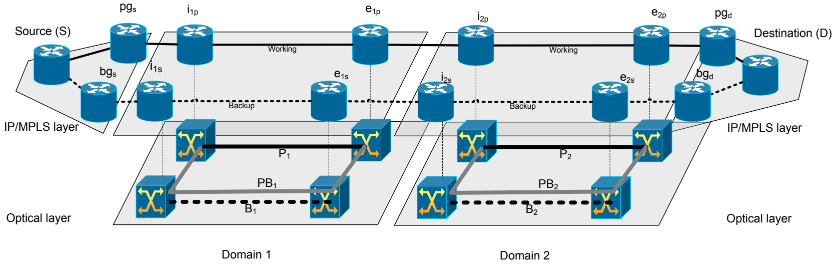


Figure 4.12: Failure free scenario

In failure-free operation (Figure 4.12), the head-end has a working LSP and a backup LSP available for the connection. Although both are disjoint in the IP/MPLS layer, it is possible that the backup LSPs optical segment B_δ shares optical resources with the working LSP P_δ or its protection lightpath PB_δ in some domains. This means that both paths cannot be considered equal. The sharing of resources for protection between multiple layers in a network is known as common pool sharing [13]. This is the reason why we call this solution a common pool multi-domain multi-layer protection solution.

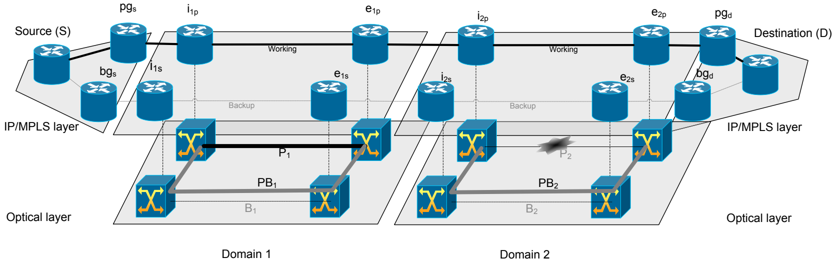


Figure 4.13: Failure in the optical layer affecting the primary path P_2.

In Figure 4.13, we show what happens when a failure affects the primary path

P_2 in Domain 2. In this case, the traffic along P will be redirected over its backup path PB_2 and the working LSP survives. First, the reserved resources along PB_2 are activated for PB_2, which means the backup path B_2 is preempted. All resources in this domain along P_2 are released. Now the preemption of the backup segment B_2 renders the backup LSP invalid. The source (S) sees a failure of its backup LSP.

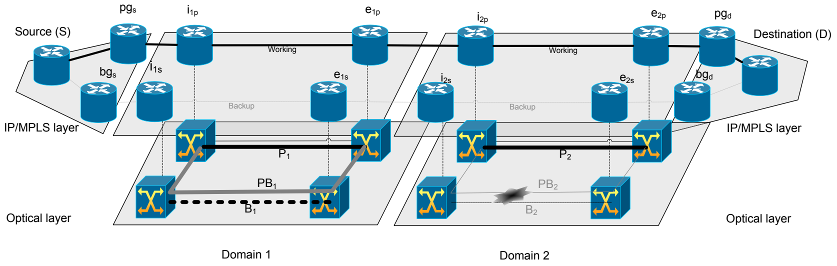


Figure 4.14: Failure in the optical layer affecting the backup paths B_2 and PB_2.

In Figure 4.14, the failure occurs along a link affecting both PB_2 and B_2. This does not affect the working LSP directly, but the head-end again sees a failure of its backup LSP. Both these failures have more or less the same effect from a point of view of the LSP end nodes, being the teardown of the backup LSP without interruption of the working LSP.

It is however also possible that a failure inside a domain δ affects only B_δ or PB_δ. The first failure will affect the LSPs in the same way, rendering the backup LSP invalid but leaving P_δ protected in the considered domain. Failure of only PB_δ will leave both LSPs active, only leaving P_δ unprotected in the considered domain.

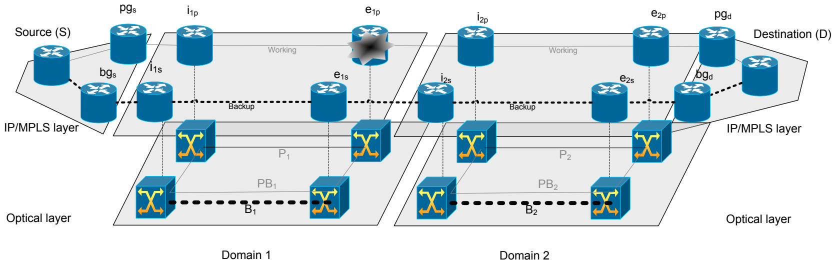


Figure 4.15: Gateway failure in the IP/MPLS layer affecting the working LSP.

In Figure 4.15, an LSR failure along the working LSP is considered. In our approach, this means a failure of the working LSP and all traffic is rerouted over the backup LSP. This is a head-end (S) operated recovery action. The head-end activates the backup LSP and associated lightpaths B_δ in every domain δ, overriding the PB_δ paths and thus PB_δ resources are released. After that, the working path P_δ resources are released in every domain, which must be done through control plane signaling. If there are enough resources, each domain could try to further protect the B_δ path after this failure.

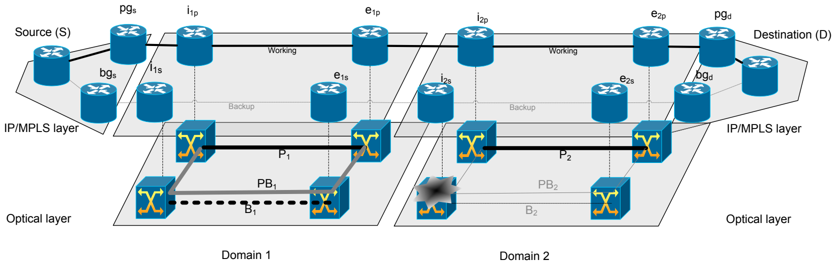


Figure 4.16: Gateway failure in the optical layer affecting the backup paths B_2 and PB_2.

In the example in Figure 4.16 is an OXC failure of a backup gateway. In this case the backup LSR fails just like an ordinary OXC failure inside a domain would be the cause of the interruption. All resources along the B paths are released.

4.5.4 Motivation

In this section we motivate our design choices for the multidomain protection. First, there are the technical advantages. Setting up both the primary connection and backup connection over the same domains gives certainty that they can be computed physically disjoint. If you run your backup path through a different network, it's for instance possible that both paths cross a river over the same bridge. Also, in every domain you can effectively share resources. The most important features are the high availability (§.4.5.5) due to protection in each domain and the fact that the existence of the routing solution can be easily guaranteed (§.4.5.6). Guaranteed existence of the solution is crucial for the path reservation mechanism specified in Section 4.5.9.

There are also obvious administrative benefits, because other domains are usually operated by competitors. If both connections run through the same sequence of domains, there are less domains involved for the overall connection.

4.5.5 Availability considerations

A widely used approach for detailed analysis of specific networks is the use of Markov models [15] [16]. We will, however, use a more general approach using

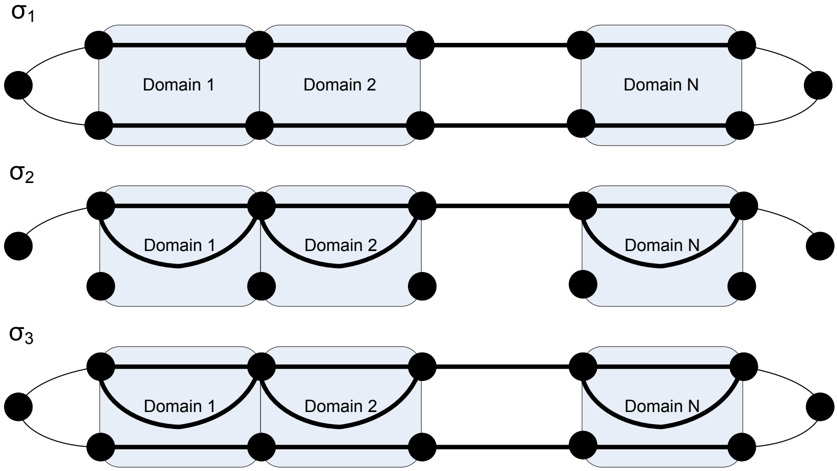


Figure 4.17: Conceptual view of multidomain solutions: σ_1 (HiTA), σ_2 (MIDP), σ_3 highest availability.

the well-known formulae for the availability of systems consisting of serial and parallel elements with statistically independent availability [17].

If a system consists of a number η of elements $\epsilon_1, \epsilon_2, \ldots, \epsilon_\eta$, with given availabilities $0 \leq \alpha(\epsilon_i) \leq 1$, which are in *serial*, then the total availability of this system given by

$$\prod_{i=1}^{\eta} \alpha(\epsilon_i) \tag{4.1}$$

This result is commonly known as Lusser's Law.

If the elements are in parallel, then the total availability is given by

$$1 - \prod_{i=1}^{\eta} (1 - \alpha(\epsilon_i)) \tag{4.2}$$

Suppose the estimated link availability is given by λ and the node availability is given by ν. We define the (dimensionless) parameter ϕ as the ratio of the lengths of the backup path over the primary path. Usually, a backup path is longer than a primary path so $\phi \geq 1$. Let N be the number of traversed domains, and n the number of hops for the primary path in each domain. We will not include the source and destination domain in these calculations, because they are (obviously) not included in SLAs. Figure 4.17 shows three conceptual solutions. The first solution σ_1 corresponds to a HiTA solution. Note that the two LSPs do not necessarily run through the same domains (not shown in picture). The second solution σ_2 is a MIDP solution, which is logically reduced to a segment protected path. Solution σ_3 is a 'common pool' solution where there is no sharing, which will give the maximum achievable availability.

First we derive the availability for a multidomain solution σ_1 which has 2 disjoint (optically unprotected) end-to-end LSPs π_1 and π_2. We easily find

$$\alpha(\sigma_1) = 1 - (1 - \alpha(\pi_1))\,(1 - \alpha(\pi_2)) \tag{4.3}$$

$$\alpha(\pi_1) = \lambda^{Nn}\nu^{Nn+1} \tag{4.4}$$

$$\alpha(\pi_2) = \lambda^{\phi Nn}\nu^{\phi(Nn+1)} \tag{4.5}$$

A multidomain solution σ_2 which uses only a single LSP, and implements it using optical path protection in each domain (e.g. the primary LSP in our solution) has availability

$$\alpha(\sigma_2) = \nu\left(\nu\alpha_\Delta\right)^N \tag{4.6}$$

$$\alpha_\Delta = 1 - (1 - \lambda^n\nu^n)(1 - \lambda^{\phi n}\nu^{\phi n}) \tag{4.7}$$

This is in effect equivalent to a segment protected path with equal segments. Now we arrive at the availability of our proposed solution. We cannot use the straightforward analysis we used for σ_1 and σ_2 because our backup LSP is possibly sharing optical resources. We can, however, expect that the availability will lie between $\alpha(\sigma_2)$ and the best case scenario in which the backup LSP is not sharing any resources ($\alpha(\sigma_3)$). The availability of σ_3 is easily derivable:

$$\alpha(\sigma_3) = 1 - (1 - \alpha(\sigma_2))(1 - \alpha(\pi_2)) \tag{4.8}$$

These three simple equations actually give us some powerful insights into the general availability of different multidomain protection solutions.

In Fig. 4.18 we show the relation between the availability of the connection and the number of domains. The number of nodes per domain is 6, $\lambda = 0.999, \nu = 0.99999, \phi = 1.1$ We immediately notice that MIDP solutions σ_2 achieves higher availability than HiTA solutions σ_1, while both effectively provide protection for a single failure. For a single domain σ_1 outperforms σ_2 because it protects against the gateway failure. If $\nu = 1$ both solutions have the same availability for a single domain.

In Fig. 4.19 we show the relation between the availability of the connection and the number of nodes per domain. The number of domains is 3, again $\lambda = 0.999, \nu = 0.99999, \phi = 1.1$. If we compare Fig. 4.18 for $N = 5$ and Fig. 4.19 for $n = 10$, we see that the primary path length for the end-to-end connection ($N.n$) is the same. σ_1 has the same availability for both scenarios. The availability for the other solutions will drop slightly, because the number of segments will be less (and the segments therefore a bit longer).

In short, we see that σ_2 improves an order of magnitude (an extra 9) over σ_1 and σ_3 improves an order of magnitude over σ_2. The σ_3 solution achieves 'five nines' over 5 domains.

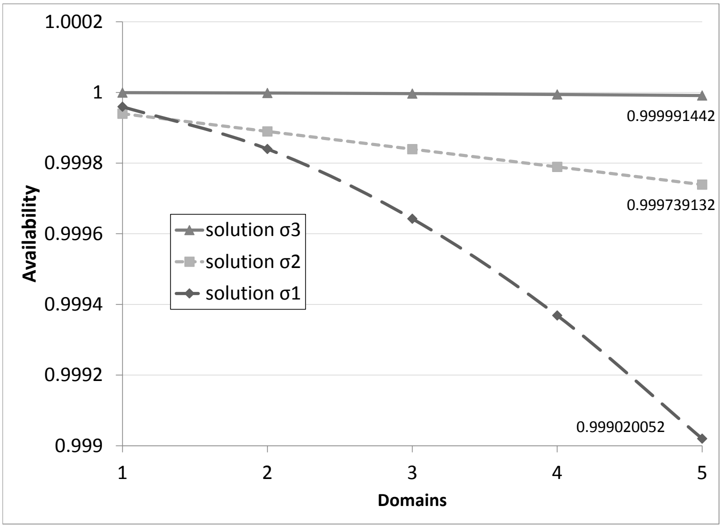


Figure 4.18: Availability in function of N, $n = 6$

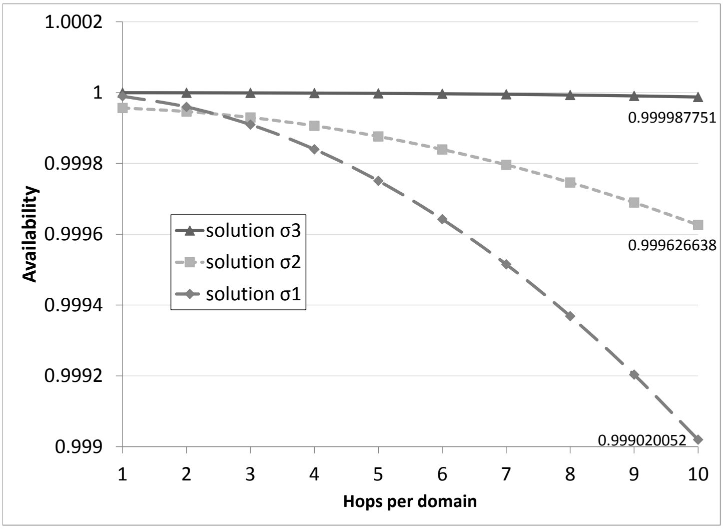


Figure 4.19: Availability in function of n, $N = 3$

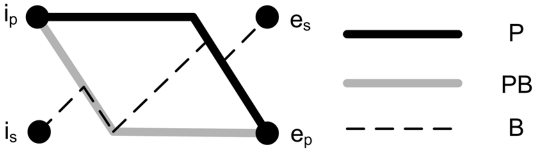


Figure 4.20: Survivable path structure

4.5.6 Applicability

To facilitate the goal of multidomain protection, we can impose some light requirements on each domain. One observation is that if you want a survivable connection over multiple domains, it's only reasonable to require that each domain in itself can setup survivable connections between two internal nodes. This means that there must be the topological constraint of a 2-connected graph, allowing for at least 2 disjoint paths between any node-pair in the network. Another requirement on the network is that it must be possible to route your paths in a flexible way (as opposed to the shortest path routing common to the OSPF mechanism in classic IP networks). So, some basic traffic engineering is needed in each network. We will therefore continue with the *formal* requirement that each domain is a *GMPLS-capable optical network with a* 2*-connected physical topology.*

An abstraction of the three optical paths in a single domain is shown in Figure 4.20. We have omitted the domain superscript δ.

4.5.6.1 Proof

We will now prove that this structure is always possible in any 2-connected network, if all the gateway nodes (i_p, i_s, e_1, e_2) are chosen in the network and only the primary ingress (i_p) and secondary ingress (i_s) are fixed. In other words, we can choose which of the egress gateways (e_1,e_2) will be primary (e_p) and the other egress will then be secondary (e_s).

Definition 4.5.1. *A* graph *G is a pair $G = (V, E)$ consisting of a set $V \neq \emptyset$ and a set E of two-element subsets of V. The elements of V are called* vertices. *An element $e = ab \in E$ is called an* edge *with end vertices $a \in V$ and $b \in V$. If V is finite, we call G a finite graph. Note that this definition specifies a simple (no parallel edges) undirected graph. A graph $S = (V_1, E_1)$ is called a* subgraph *of $G = (V, E) \Leftrightarrow V_1 \subseteq V \wedge E_1 \subseteq E$. A path is a non-empty graph $P = (V_2, E_2)$ of the form $V_2 = \{v_0, v_1, \ldots, v_k\}, E_2 = \{v_0v_1, v_1v_2, \ldots, v_{k-1}v_k\}$. We will specify paths with a vertex list: $P\,(v_0v_1 \ldots v_z)$. A* cycle *is a closed path, i.e. the end nodes v_0 and v_z in the list representation are the same. We define the* intersection *of two paths, $P_1 \cap P_2$ as the set of vertices V_i that are in both paths.*

If $P_1 \cap P_2 = \emptyset$ we call P_1 and P_2 disjoint. If for two paths having the same end nodes $P_1\left(a \ldots b\right) \cap P_2\left(a \ldots b\right) = \left(a, b\right)$ then we also call these paths disjoint. If for any pair of vertices $\left(v_1, v_2\right) \in V^2$ there exist k mutually node-disjoint paths between them, we call the graph k-connected and the connectivity $\kappa(G) = k$. In a k-connected graph, it will require the removal of at least k nodes to disconnect the graph into different components. A set of k nodes which disconnects the graph is called a set of cut–vertices.

The k-connectedness of a graph does not imply it is always possible to construct disjoint paths between 2 given nodes successively. If it is not possible to find a disjoint path between two nodes after construction of a first path P_β between them, we call P_β a *blocking path*. It is easy to see that a blocking path must contain a set of cut–vertices.

Theorem 4.5.1. *Let $G(V, E)$ be a graph on the set of vertices V with $|V| \geq 4$ and connectivity $\kappa(G) \geq 2$. If we choose four different nodes $\left(i_p, i_s, e_1, e_2\right) \in V^4$, it is always possible to find three paths P, PB and B in G, that suffice one of the following two conditions:*

$$\begin{aligned} &P\left(i_p \ldots e_1\right) \cap PB\left(i_p \ldots e_1\right) = \left\{i_p, e_1\right\} \wedge \\ &B\left(i_s \ldots e_2\right) \cap \left\{i_p, e_1\right\} = \emptyset \end{aligned} \tag{4.9}$$

$$\begin{aligned} &P\left(i_p \ldots e_2\right) \cap PB\left(i_p \ldots e_2\right) = \left\{i_p, e_2\right\} \wedge \\ &B\left(i_s \ldots e_1\right\} \cap \left\{i_p, e_2\right\} = \emptyset \end{aligned} \tag{4.10}$$

Theorem 4.5.1 states the requirements for common pool interdomain connection in a formal way. The two conditions (4.9) and (4.10) are identical but the roles of e_1 and e_2 are reversed. Saying that one of the two conditions must be true is to state that e_p and e_s can be chosen and i_p and i_s are fixed.[1]

The conditions state that it is possible (a) to find two disjoint paths between i_p and e_p and (b) a path from i_s to e_s that does not cross i_p or e_p. The first condition is trivial, but in a 2–connected network it's not trivially possible to specify two nodes which must be excluded from a path and find a solution. They can form a pair of cut–vertices. This is why both conditions (4.9, 4.10) are included in the theorem, which will become clear in Lemma 4.5.2.

[1] Note that due to symmetry, Theorem 4.5.1 also proves that e_p and e_s can be fixed and i_p and i_s interchangeable. This could be useful for using backward propagation in signaling and setting up interdomain connections.

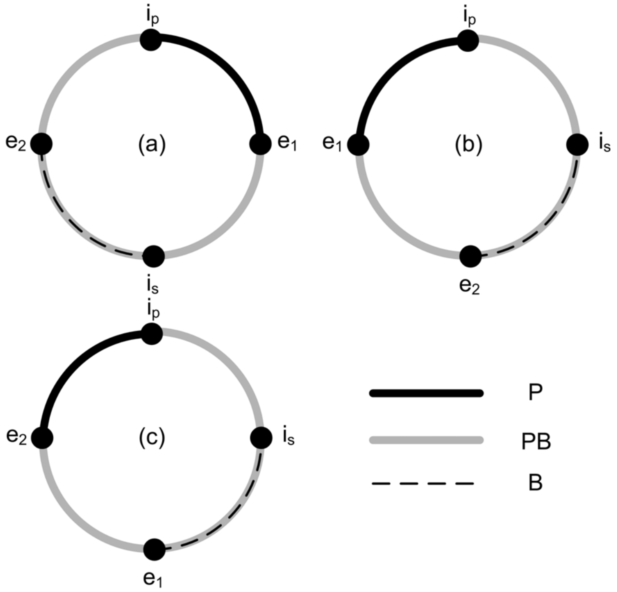


Figure 4.21: Survivable structure on a cycle (Lemma 4.5.2)

Lemma 4.5.2. *Theorem 4.5.1 holds if there is a cycle containing i_p, e_1, i_s, e_2.*

Proof: *If all nodes i_p, e_1, i_s, e_2 are on a cycle, it is immediately clear that the path P will follow one side of this cycle and that the path PB must span the other side of the cycle. Now, it is also easy to see that there are three distinct permutations of the gateway nodes which we must consider, since all others are rotations or mirror images of these.*

1. *$i_p - e_1 - i_s - e_2$. (Figure 4.21.a)*

   In this configuration, we can choose the path $P\,(i_p \ldots e_1)$ not to contain any other gateways and the path $PB\,(i_p \ldots e_2 \ldots i_s \ldots e_1)$ must then contain i_s and e_2. The path $B\,(i_s \ldots e_2)$ can be chosen as the $i_s - e_2$ subpath fully contained in PB. In this way, condition (4.9) is met.

2. *$i_p - i_s - e_2 - e_1$. (Figure 4.21.b)*

   We can setup the same paths as in the first option: we can choose the path $P\,(i_p \ldots e_1)$ not to contain any other gateways and again the path $PB\,(i_p \ldots i_s \ldots e_2 \ldots e_1)$ (along the other side of the cycle) will contain i_s and e_2. The path $B\,(i_s \ldots e_2)$ is the $i_s - e_2$ subpath fully contained in PB, meeting condition (4.9).

3. *$i_p - i_s - e_1 - e_2$. (Figure 4.21.c)*

   It is impossible to satisfy condition (4.9): should we consider the paths $P\left(i_p \ldots i_s \ldots e_1\right)$ and $PB\left(i_p \ldots e_2 \ldots e_1\right)$, then $B\left(i_s \ldots e_2\right)$ will either contain i_p or e_1, which is a violation of the subcondition $B\left(i_s \ldots e_2\right) \cap \{i_p, e_1\} = \emptyset$. However, it is immediately clear that this configuration is the same as the second one, where e_1 and e_2 switched places. Following the same argument as above, we can therefore conclude that, the paths $P\left(i_p \ldots e_2\right)$ (not containing the other gateways), the path $PB\left(i_p \ldots i_s \ldots e_1 \ldots e_2\right)$ and the PB–subpath $B\left(i_s \ldots e_1\right)$ will satisfy condition (4.10).

□

With the result from Lemma 4.5.2, it is easy to prove Theorem 4.5.1. Given the four nodes i_p, i_s, e_1, e_2 in G, we can find 2 disjoint paths P and PB between i_p and e_1, because of the 2–connectedness of G. This fulfills the first clause of condition (4.9) in Theorem 4.5.1. Similarly, we can find 2 disjoint paths P_q and P_r from i_s to e_2. The second clause of the requirement (4.9) states that there must be a path B from i_s to e_2 not containing i_p or e_1. This condition is met when one of the paths P_q and P_r does not contain i_p or e_1 and we choose $B\left(i_s \ldots e_2\right)$ equal to this path. The only situation when $B = P_q$ or $B = P_r$ does not satisfy condition (4.9) is when both P_q and P_r run through either i_p or e_1, but then all four nodes i_p, i_s, e_1 and e_2 are on a cycle formed by P_q and P_r and, following Lemma 4.5.2, there are paths $P\left(i_p \ldots e_2\right)$, $PB\left(i_p \ldots e_2\right)$ and $B\left\{i_s \ldots e_1\right)$ that satisfy condition (4.10).

□

Note that 2–connectedness of the graph G is a sufficient condition, but not a neccesary condition. The necessary condition is that there is a 2–connected subgraph of G containing i_p and e_1 and a 2–connected subgraph of G containing i_s and e_2. However, survivable networks typically have a 2–connected topology.

4.5.6.2 Heuristic solution

Our proof of Theorem 4.5.1 immediately gives rise to one algorithm for finding a suitable solution: first, we remove the nodes i_p and $e_p = e_1$ from the graph G and compute B as the shortest path from i_s to $e_s = e_2$. If no such path can be found, remove i_p and $e_p = e_2$ from the graph G and compute B as the shortest path from i_s to $e_s = e_1$, which then must exist. Then compute a pair of disjoint paths from i_p to e_p using the Suurballe-Tarjan [18] algorithm to form P and PB.

4.5.7 Optimum solution

Now that we know that a structure can always be computed in any 2–connected network, the question remains how to compute the optimal solution requiring the least resources. In this section, we provide a flow-based *Integer Linear Programming* (ILP) model that specifies this optimal solution. The input is the graph G and four nodes i_p, i_s, e_p, e_s, and the output are the links that make up P, PB and B. Note that the choice of the egress gateways must be input into the model. If the model finds no solution, the egress gateways must be switched.

The ingress gateways i_s and i_p are *sources* for the flow, the egress gateways e_p and e_s are *sinks*. The boolean variable $x^P_{v_iv_j}$ denotes that the directed link v_iv_j is used by path P. Since the paths are directed (but symmetrical), the reverse link is also used and denoted by the variable $xr^P_{v_iv_j}$. All these variables are initialized 0 (zero). We try to minimize the number of links (resources) used.

First, we set the restriction on the sources. The difference of the outgoing flow and the incoming flow must be 1 for the sources. We can specify this in different ways. We chose to specify that the total incoming flow per path for the sources is 0 and the total outgoing flow is 1. i_p is the source for P and PB, i_s is similarly the source for the path B. The equations for the outgoing flows are (E is the edge-set as defined above):

$$\sum_{l=1}^{n} x^P_{i_pv_l} = 1, i_pv_l \in E \tag{4.11}$$

$$\sum_{l=1}^{n} x^{PB}_{i_pv_l} = 1, i_pv_l \in E \tag{4.12}$$

$$\sum_{l=1}^{n} x^B_{i_sv_l} = 1, i_sv_l \in E \tag{4.13}$$

and for the incoming flows:

$$\sum_{k=1}^{n} x^P_{v_ki_p} = 0, v_ki_p \in E \tag{4.14}$$

$$\sum_{k=1}^{n} x^{PB}_{v_ki_p} = 0, v_ki_p \in E \tag{4.15}$$

$$\sum_{k=1}^{n} x^B_{v_ki_s} = 0, v_ki_s \in E \tag{4.16}$$

We use the same approach for the restrictions on the sinks. The total incoming flow is 1, the total outgoing flow is 0. The equations for the incoming flows:

$$\sum_{k=1}^{n} x^{P}_{v_k e_p} = 1, v_k e_p \in E \tag{4.17}$$

$$\sum_{k=1}^{n} x^{PB}_{v_k e_p} = 1, v_k e_p \in E \tag{4.18}$$

$$\sum_{k=1}^{n} x^{B}_{v_k e_s} = 1, v_k e_s \in E \tag{4.19}$$

and for the outgoing flows:

$$\sum_{l=1}^{n} x^{P}_{e_p v_l} = 0, e_p v_l \in E \tag{4.20}$$

$$\sum_{l=1}^{n} x^{PB}_{e_p v_l} = 0, e_p v_l \in E \tag{4.21}$$

$$\sum_{l=1}^{n} x^{B}_{e_s v_l} = 0, e_s v_l \in E \tag{4.22}$$

The paths must be continuous, so for all nodes, except the sources and sinks, the total incoming flow must equal the total outgoing flow:

$$\forall v_k \notin \{i_p, e_p\} \sum_{z=1}^{n} x^{P}_{v_z v_k} = \sum_{y=1}^{n} x^{P}_{v_k v_y}, v_z v_k \in E, v_k v_y \in E \tag{4.23}$$

$$\forall v_k \notin \{i_p, e_p\} \sum_{z=1}^{n} x^{PB}_{v_z v_k} = \sum_{y=1}^{n} x^{PB}_{v_k v_y}, v_z v_k \in E, v_k v_y \in E \tag{4.24}$$

$$\forall v_k \notin \{i_s, e_s\} \sum_{z=1}^{n} x^{B}_{v_z v_k} = \sum_{y=1}^{n} x^{B}_{v_k v_y}, v_z v_k \in E, v_k v_y \in E \tag{4.25}$$

We now specify that P and PB must be disjoint. The solution requires node-disjointness, but we will add the link-disjoint constraints for completeness, so a choice can be made. Note that we have to make sure that PB does not use a link in the reverse direction of P and vice versa. The link-disjointness constraints are given by:

$$x^{P}_{v_k v_l} + x^{PB}_{v_k v_l} + x^{P}_{v_l v_k} + x^{PB}_{v_l v_k} \leq 1, \forall v_k v_l \in E \tag{4.26}$$

Node disjointness is achieved by adding the following constraints, stating that only one of P and PB can enter or exit any given node, except for the source i_p and the sink e_p:

$$\forall v_l \neq i_p \sum_{k=1}^{n} \left(x^{P}_{v_k v_l} + x^{PB}_{v_k v_l} \right) \leq 1, v_k v_l \in E \tag{4.27}$$

$$\forall v_k \neq e_p \sum_{k=1}^{n} \left(x^{P}_{v_k v_l} + x^{PB}_{v_k v_l} \right) \leq 1, v_k v_l \in E \tag{4.28}$$

The following restrictions ensure that the path B cannot run through i_p or e_p:

$$\sum_{k=1}^{n} \left(x^{B}_{v_k i_p} \right) = 0, v_k i_p \in E \tag{4.29}$$

$$\sum_{k=1}^{n} \left(x^{B}_{v_k e_p} \right) = 0, v_k e_p \in E \tag{4.30}$$

$$\sum_{l=1}^{n} \left(x^{B}_{i_p v_l} \right) = 0, i_p v_l \in E \tag{4.31}$$

$$\sum_{l=1}^{n} \left(x^{B}_{e_p v_l} \right) = 0, e_p v_l \in E \tag{4.32}$$

The reverse paths are given as:

$$x^{P}_{v_k v_l} = xr^{P}_{v_l v_k}, \forall v_k v_l \in E \tag{4.33}$$

$$x^{PB}_{v_k v_l} = xr^{PB}_{v_l v_k}, \forall v_k v_l \in E \tag{4.34}$$

$$x^{B}_{v_k v_l} = xr^{B}_{v_l v_k}, \forall v_k v_l \in E \tag{4.35}$$

Now that all paths are specified, we can state our optimization goal. The goal is to minimize the capacity requirement of the network, therefore, minimizing the total number of links used. P and PB are node-disjoint, therefore, we can simply add up any x_P and x_{PB}. Now, B can share with both those paths, which implies a maximum operation. To ensure our sharing goal, we add a variable $y_{v_k v_l}$ which must always be greater than or equal to any of the other variables on each link. This ensures the sharing between P and B, and between PB and B.

$$y_{v_k v_l} \geq x^{P}_{v_k v_l}, \forall v_k v_l \in E \tag{4.36}$$

$$y_{v_k v_l} \geq xr^{P}_{v_k v_l}, \forall v_k v_l \in E \tag{4.37}$$

$$y_{v_k v_l} \geq x^{PB}_{v_k v_l}, \forall v_k v_l \in E \tag{4.38}$$

$$y_{v_k v_l} \geq xr^{PB}_{v_k v_l}, \forall v_k v_l \in E \tag{4.39}$$

$$y_{v_k v_l} \geq x^{B}_{v_k v_l}, \forall v_k v_l \in E \tag{4.40}$$

$$y_{v_k v_l} \geq xr^{B}_{v_k v_l}, \forall v_k v_l \in E \tag{4.41}$$

The optimization goal is:

$$minimize : \sum_{k=1}^{n} \sum_{l=1}^{n} y_{v_k v_l}, v_k v_l \in E \tag{4.42}$$

This specifies the paths, but there is still one loophole, being paths which back up on themselves. It is not directly obvious from these equations, but we noticed when examining our solutions that paths of the form $(i_s...e_s)$ with a disconnected cycle of zero net cost are also solutions to these equations. Indeed, the flow is conserved on all nodes of this disconnected cycle. If we would optimize for minimal path lengths, these solutions would immediately disappear, but since we optimize for sharing, some cycles can have zero net cost for the optimization. Since P and PB are disjoint, these paths always add to the optimization cost, but the path B does not. Generally these cycles for B are filtered out because of the disjointness requirement of the paths P and PB. If we were working in an undirected graph, the only plausible cycle to add for B would be the cycle formed by P and PB, but since B cannot go through i_p or e_p this is not allowed. But cycles in the form of a path looping back onto itself are still plausible solutions in the directed graph. To remove these solutions, we add the restriction that, if the directed link v_kv_l is on B, its reverse link v_lv_k cannot be on B. The extra restrictions are thus:

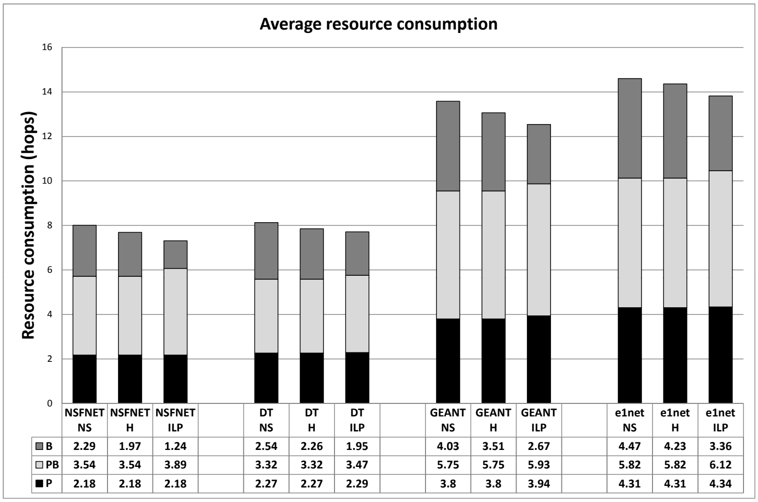
$$x^B_{v_kv_l} + x^B_{v_lv_k} \leq 1, \forall v_kv_l \in E \tag{4.43}$$

4.5.8 Simulation results

We will now compare our simple heuristic to an ILP solution computed using CPLEX 11 [19]. We have computed results for all four reference networks.

In Figure 4.22 we summarize the results for 100 simulations on each network. In each simulation we choose four nodes at random, representing i_p, i_s, e_p and e_s, and we compare 3 different solution options for the survivable structure. In the first option, we do not share resources (NS). The paths are computed, like the heuristic, as shortest cycle for P and PB and the shortest path for B. Without sharing this is the optimum solution regarding resource consumption. The second option is the heuristic solution (H) described above, and the third option is the CPLEX solution to our ILP. The figure shows the average resource consumption for each path seperately, and these are stacked to show the total resource consumption. In the first two solutions P and PB use the same (minimum) amount of resources, as they are computed along the shortest cycle. The difference between these two solutions lies in sharing of B. Taking the Geant2 network as an example, we see that for the heuristic, P uses on average 3.8 hops, and PB on average 5.75 hops, meaning the shortest cycle is on average 9.55 hops. The average total resource consumption for the H solution is 13.06, the average total for the ILP solution is 12.54. The difference is achieved by choosing a longer cycle for $P + PB$ and (9.87 hops vs. 9.55) hops and improving the sharing with B. The capacity gain for the ILP over the heuristic is 4.94% for NSFNet, 1.78% for the DT network, 3.98% for Geant2 and 3.76% for e1net.

In Figure 4.23 we show the resource gain (fractional from 0-1) of the ILP over



	NSFNET NS	NSFNET H	NSFNET ILP		DT NS	DT H	DT ILP		GEANT NS	GEANT H	GEANT ILP		e1net NS	e1net H	e1net ILP
▨ B	2.29	1.97	1.24		2.54	2.26	1.95		4.03	3.51	2.67		4.47	4.23	3.36
□ PB	3.54	3.54	3.89		3.32	3.32	3.47		5.75	5.75	5.93		5.82	5.82	6.12
■ P	2.18	2.18	2.18		2.27	2.27	2.29		3.8	3.8	3.94		4.31	4.31	4.34

Figure 4.22: Average resource consumption

the heuristic for each individual simulation. We sorted all 100 results by gain and show the 41 best solutions (60-100). The graph immediately shows two things: on the one hand, in 64 (Geant2) to 87 (DT) percent of the cases, the solution found by the heuristic is optimal. On the other hand, individual gain for a single solution can go to 18% (DT) and even 30% (NSFNet). So, while the average gain of the ILP may be rather small (around 5%), for individual connections it may be as much as 30 percent.

We also assess the blocking performance of these algorithms in an operational network scenario. We take the e1net topology which interconnects 17 domains, where each domain is assigned at least 2 gateways. We assume full wavelength conversion and 80 available channels on every link. We offer random sets of 1000 connections between domain pairs to the network, calculating for each of these sets the solution using each of the algorithms above, taking into account the remaining resources in the network. If a connection cannot be set up in its entirety (i.e., all paths P,PB and B) it is blocked and we try the next connection. Note that, because we offer the connections in a sequential way, there is no global optimization in resource consumption. We do not try to reroute existing connections if the network has insufficient resources to accommodate the currently requested connection. This would be an area for future study.

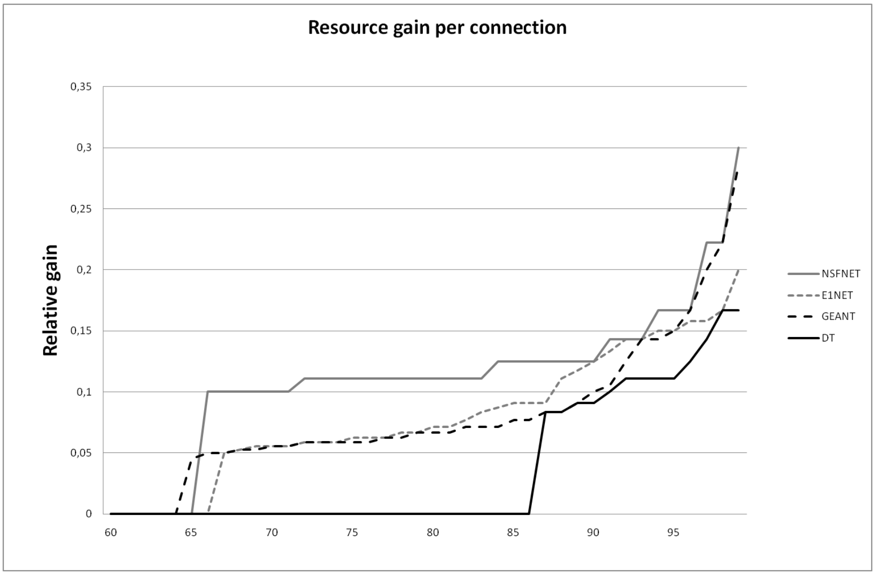


Figure 4.23: Resource gain per connection

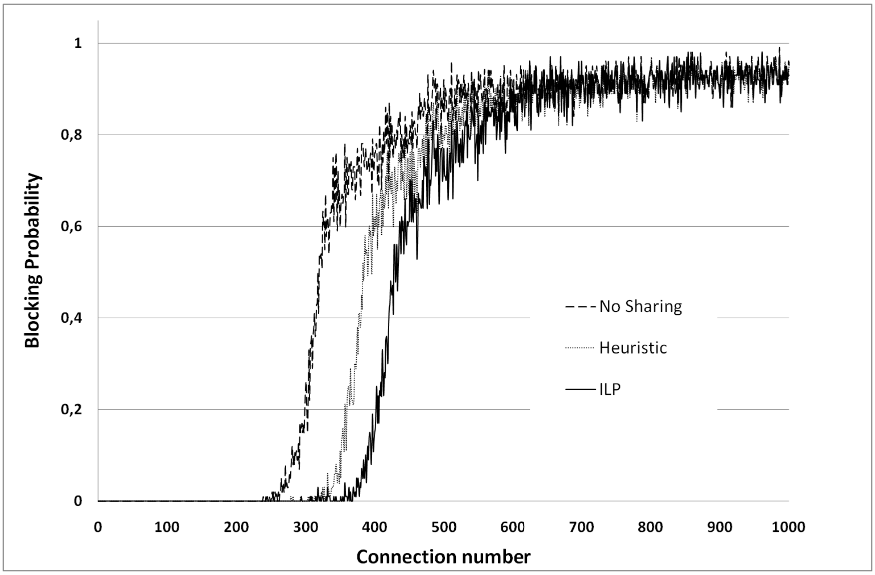


Figure 4.24: Blocking probability

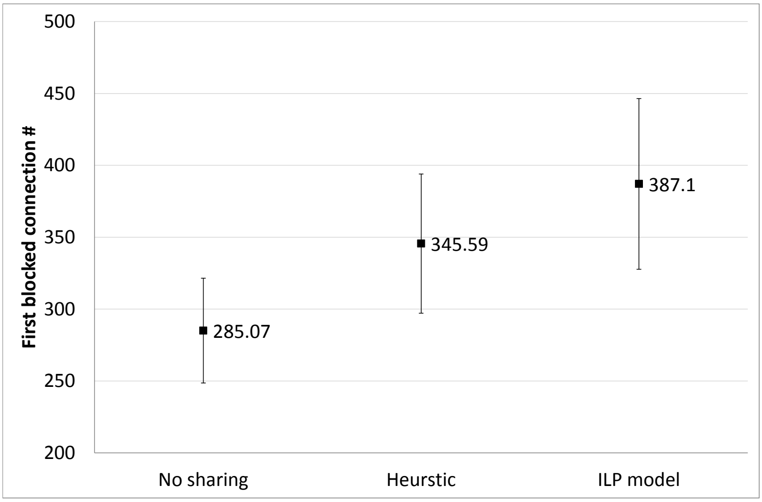


Figure 4.25: Blocking performance of multidomain protection

Results are summarized in Figure 4.24, which shows the estimated blocking probability (average taken from 100 experiments) versus the connection number. The graph clearly shows the trend that, whatever the solution strategy used, whenever the network starts blocking connections, the blocking probability for subsequent connections rises very steeply. It also indicates that the ILP has better blocking performance. Since a blocking probability of more than 1% is definitely unacceptable, we can choose the number of connections succesfully set up in the network before the first blocking occurs as a figure of merit for the three solutions.

Figure 4.25 shows the average connection number for the first blocked connection, with a 95% (2σ) confidence interval. Without resource sharing between the paths, we can offer 285 connections on average to the network without blocking. Enabling sharing raises this number to 345, or a 21% increase. This is within 12% of the optimum ILP solution, which accepts 387 random connections on average. Note that these are averages over multiple offered sets of connections.

Due to the different load distributions on the network, it is perfectly possible that, for instance, the ”No Sharing”' solution outperforms the heuristic. This is also confirmed in our experiments. Figure 4.26 shows the relative blocking performance compared to the ”No Sharing” solution. For two connection sets, the ”No Sharing” solution could accommodate 1 connection more than the heuristic (ex-

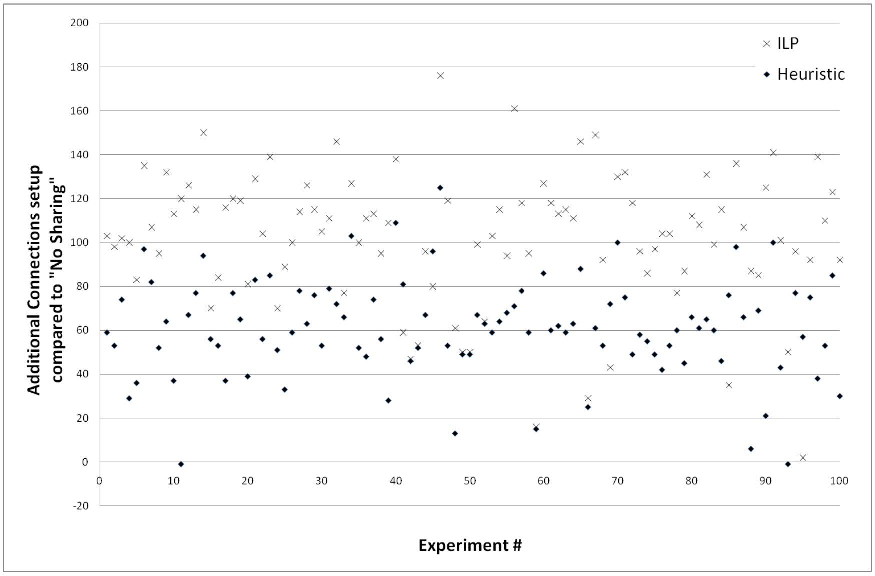


Figure 4.26: Relative blocking performance

periments 11 and 93). On one occasion, the ILP only accommodated 2 connections more, while the heuristic could set up 57 extra connections after the ”no sharing”’ started blocking (experiment 95). On average however, the heuristic allows 60.52, and the ILP 102.03 more connections (see Figure 4.25). We calculated the standard error (σ) of the values in Figure 4.26 to be 22.71 for the heuristic and 30.61 for the ILP. This means that these specific results are atypical and well outside of the 2σ confidence interval.

4.5.9 Multidomain connection signaling

We now detail how the survivable connection is set up regarding the full multilayer view of the network, how different failure scenarios are explicitly handled, and how they affect the overall network state. Note that in the upper (network) layer, the LSP segments depicted in each domain are single-hop connections between gateways. In the lower (optical) layer, paths P_δ, PB_δ, and B_δ are actual multi hop paths implementing single-hop upper layer LSPs. Two failure scenarios are considered, being OXC failures inside a domain and OXC or LSR failure of a gateway. We now discuss our PCE/Resource Reservation Protocol with Traffic Engineering Extensions (RSVP-TE)-based approach here. RFC 5151 [20] details restoration for multidomain MPLS. There is also a short section (5.2) in RFC5151 which details some aspects for protection using segment recovery [21], route ex-

<table>
<tr><td>S</td><td>Type=32
(AS NUMBER)</td><td>Length=4</td><td colspan="2">AS identifier</td></tr>
<tr><td>S</td><td>Type=11
(IPv4 I-LIST)</td><td>Length=4+4*n</td><td colspan="2">IPv4address 1</td></tr>
<tr><td colspan="3">IPv4address 1 (continued)</td><td colspan="2">IPv4address 2</td></tr>
<tr><td colspan="3">IPv4address 2 (continued)</td><td colspan="2"></td></tr>
<tr><td colspan="3"></td><td colspan="2">IPv4address n</td></tr>
<tr><td colspan="3">IPv4address n (continued)</td><td>Prefix = 32</td><td>Resvd</td></tr>
<tr><td>L</td><td>Type=21
(IPv4 E-LIST)</td><td>Length=4+4*m</td><td colspan="2">IPv4address 1</td></tr>
<tr><td colspan="3">IPv4address 1 (continued)</td><td colspan="2">IPv4address 2</td></tr>
<tr><td colspan="3">IPv4address 2 (continued)</td><td colspan="2"></td></tr>
<tr><td colspan="3"></td><td colspan="2">IPv4address m</td></tr>
<tr><td colspan="3">IPv4address m (continued)</td><td>Prefix = 32</td><td>Resvd</td></tr>
</table>

Figure 4.27: Gateway Specification Routing Object

clusion [22] and cooperation between PCE's.

We propose a new RSVP-TE object to specify gateways and their priorities, the gateway specification routing object (GSRO), and its most important contents are a domain identifier and two lists of gateways (an ingress list and an egress list) in priority order. This enables more flexible solutions.

4.5.9.1 Gateway Specification Routing Object (GSRO)

The Gateway Specification Routing Object is used to signal the multidomain connection. Design goals are compatibility and genericity: the GSRO should not interfere with the current RSVP-TE operation and it should be useful for all multidomain applications. We based its design on the Explicit Route (ERO) Object, which is a collections of subobjects specifying abstract nodes (currently, IPv4 prefixes (Type 1), IPv6 (Type 2) prefixes or AS numbers (Type 32) are defined) [23]. These abstract nodes can specify a loose hop (L bit set to 1) or a strict hop (L bit set to 0). When a loose hop is configured, it identifies one or more transit nodes through which the path must be routed. The network IGP determines the exact route from the inbound gateway to the first loose hop, or from one loose hop to the next. The loose hop specifies only that a particular node be included in the path. When a strict hop is configured, it identifies an exact path through which the LSP must be routed. Strict-hop EROs specify the exact order of routers through which the RSVP messages are sent. Loose-hop and strict-hop EROs can be configured simultaneously. In this case, the IGP determines the route between loose hops, and the strict-hop configuration specifies the exact path for particular LSP path segments.

A GSRO can contain the following subobjects: AS numbers (proposed Type 32, as in ERO), ingress lists (I-LIST, Type 11 (IPv4), Type 12 (IPv6)) and egress lists (E-LIST, Type 21 (IPv4) and Type 22 (IPv6)). The I-LISTs and E-LISTs can contain any number of gateways. The lists may overlap, indicating a gateway which is used for both ingress as egress (transit through a single node). The ingress lists must be in fixed priority order, the egress lists can be in loose priority order. The loose priority may be necessary for swapping the egress gateways in case the topology requires it. The destination node should receive a path message which contains one GSRO per domain. A well-formed GSRO object must contain at least an I-LIST or an E-LIST containing at least one node. So it is possible to specify only the egress gateways in a GSRO for the source domain and only the ingress gateways in the GSRO for the destination domain. The specified nodes must be gateways and cannot be generic prefixes, so the number of prefix bits is set as 32.

The format of a standard GSRO object is shown in Figure 4.27. The first subobject is the AS number (a 16 bit unique identifier). This is followed by an I-LIST and an E-LIST. For an IPv4 addressed domain with $n > 0$ ingress gateways and $m > 0$ egress gateways, the size of the GSRO is $12 + 4n + 4m$ bytes. In an IPv6 addressed domain this would be $12 + 16n + 16m$.

4.5.9.2 GSRO PATH-RESV signaling

The scenario is shown in Figure 4.28. The first step is that the head-end sends an RSVP-TE PATH message to the tail-end, saying it wants to set up a survivable connection through the intermediate domains (Domain 1 and Domain 2). To do this, the head-end, for example, performs a lookup in its IP table or OSPF-TE database to determine its preferred gateway toward the destination. The source sends the PATH message (with as destination the IP address of node D) toward its preferred gateway e_{1p}. This gateway determines the next hop (for instance, using its BGP table or using the explicit route object (ERO) in the PATH message), being

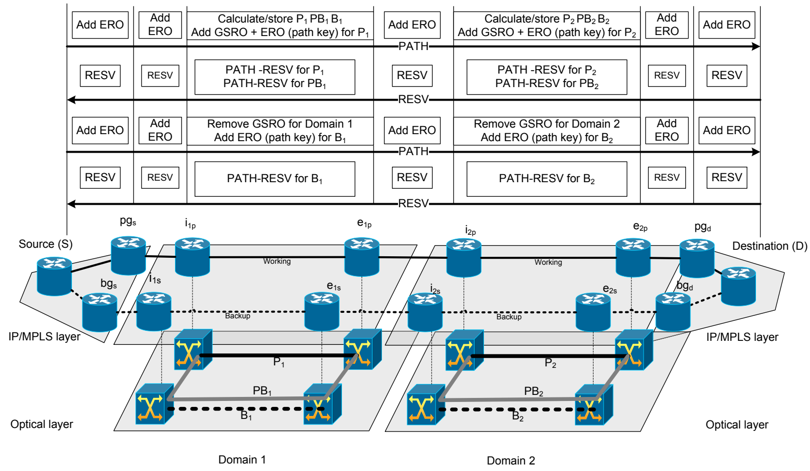


Figure 4.28: Signaling messages for multidomain connection setup

the primary ingress gateway for Domain 1, i_{1p}, and forwards the message. The primary gateway i_{1p} then consults the PCE for Domain 1. This PCE determines the secondary ingress gateway, i_{1s}, and computes the optical paths P_1, PB_1, and B_1. This also determines the primary egress gateway e_{1p} and back-up egress e_{1s}. The path information for P_1, PB_1, and B_1 is stored in a temporary database, and the routing information is returned to i_{1p}. The information returned by the PCE to i_{1p} contains the four gateway nodes (i_{1p},e_{1p}, i_{1s}, e_{1s}) and the path information for P_1 and PB_1. A GSRO with the gateway information and an ERO with a path key for P_1 are attached to the PATH message. Path keys are an alias for an explicit path and provide a means for keeping paths confidential. The PATH message is sent to e_{1p}. Then e_{1p} determines the next domain (Domain 2) using its BGP table or by consulting the PCE and forwards the PATH message to i_{2p}. Note that any internal topology information that may be stored in a record route object (RRO) should be filtered out by the egress gateways. The gateway i_{2p} consults the PCE of Domain 2 and also sends the information contained in the GSRO to the PCE. The PCE first determines the gateway i_{2s} with this information. The rest of the path computation (e_{2p}, e_{2s}, P_2, PB_2, and B_2) is the same as in the first domain, and a second GSRO and ERO are attached. The PATH message is finally forwarded by e_{2p} to the destination node via pg_d.

Upon reception of the PATH message containing the two GSROs and two EROs, the destination replies with a reservation (RESV) message. This RESV message contains the ERO's and GSRO's from the PATH message and is sent to e_{2p} via bg_d and forwarded to i_{2p}, which requests the path for PB_2 from the PCE (the path key for P_2 is in the RESV message). Then the paths P_2 and PB_2 are reserved in the optical layer using a standard PATH-RESV according to the standard procedure detailed in RFC 4872 [24]. If this reservation is successful, the RESV message is forwarded to Domain 1, and the paths P_1 and PB_1 are reserved in the same way. Finally, the source receives the RESV message and establishes the final part of the connection. Then, the working LSP is reserved and can be used.

Next, the back-up LSP is signaled using the same identifier in the session object as used for the primary LSP. The source sends a PATH message containing the two GSROs to i_{1s} via bg_s. This gateway i_{1s} then consults its PCE. The PCE will recognize that the request is coming from the back-up gateway for this connection by inspection of the RSVP-TE session object, and replies with the route information for B_1. The PATH message with an ERO is sent to Domain 2 (after possible filtering of the RRO by the egress node); an ERO with path key for B_2 is added, and the message is forwarded toward the destination. The destination replies with a final RESV message (the GSROs can be discarded), and the back-up path is set up using standard ERO messages as above. When the head-end receives the RESV message and establishes the final segment, the setup is complete.

The reason to perform the path reservation in two phases is that RSVP-TE does not allow path set up initiated by any node other than the path head-end (meaning that node gateway $i_{\delta p}$ cannot set up a connection between nodes $i_{\delta s}$ and $e_{\delta s}$) [23].

There are some issues worth mentioning. First, it must be possible that the head-end specifies all (abstract) nodes or gateways it wants to use in loose EROs and/or GSROs (in case back-up gateways are specified). Because these objects implicitly or explicitly specify the egress router priorities, the PCE should be able to choose their priority if the preferred connection is not possible (and notify the head-end). Second, for resource sharing, RSVP-TE uses an association object. The path B must be associated with both PB and P for resource sharing. Third, for a failure of the primary path, there are two possible restoration solutions: the path PB initiated by the gateway and the path B initiated by the head-end. Because there are two solutions, there is a race condition between these two back-ups. In an ideal scenario, the intradomain restoration performed by the gateway should be under 50 ms, and the head-end should not notice the interruption. However, any escalation strategy developed for multilayer race conditions, such as hold-off timers (Section 3.3.2) can be readily adopted for this multidomain race condition. Fourth, the path structure is computed as a single entity; only the signaling is performed in two phases to comply as much as possible with current standards. There is some possibility that after the working LSP has been set up, the reservation of the path B would fail in some domain. This is the reason why the PCE stores the information for all three paths during the set up of the working LSP, and why the back-up LSP should be signaled with the same session identifier (null padded string) as the working LSP. When (in Domain δ) the RESVmessage is successfully returned for P_δ and PB_δ, the resources for B_δ must be marked as used by the PCE to avoid blocking of B_δ during its PATH-RESV phase. Should the path B_δ fail (for instance, due to a failure in the short time frame between the RESV of P_δ and PB_δ and the PATH-RESV of B_δ), the PCE for that domain should try to find an alternate path for B_δ. If this fails, the entire session should be rolled back, and the connection blocked. Note that this is a relatively rare scenario. Fifth, in failure-free operation, the head-end has a working LSP and a back-up LSP available for the connection. Note that, although both are disjoint in the IP/MPLS layer, it is possible that the back-up LSP optical segment (B_δ) shares optical resources with the working LSP (P_δ) and/or its protection lightpath (PB_δ) in some domain δ. This means that these LSPs cannot be considered equal and load balancing may not be possible.

4.6 Conclusion

In this chapter we have presented some properties of common pool multidomain resilience. We compared a number of different approaches for multi-domain recovery with respect to resource efficiency which show that the common pool principle is beneficial for multidomain networks. We have noted that our solution lies in the MIDP class of multidomain protection mechanisms, and shown through analytical analysis that this class has very high availability. We have shown, using some examples, how this solution handles intradomain failures and gateway failures. We have given a proof that the solution can be computed in any 2-connected network, which is important for signaling such connections in real networks. This proof leads to an easily computable heuristic solution. We have also presented an ILP model for computing an optimal solution, and compared the heuristic to the ILP, and the optimum solution without sharing, using simulations on four different networks. The results show that the heuristic often finds the optimum solution, but that for individual connections, the gain of the ILP may be considerable. We also compared the solutions on an operational scenario, where we offered a random set of 1000 connections to the e1net network. We show that, while it is possible that for certain sets either of the solutions may be the best one, on average the heuristic and ILP allow 21 resp. 36 percent more connections compared to the solutions without resource sharing. The ILP achieves better results by increasing some paths in length to allow for more sharing. We also provided a protocol extension which allows us to set up the connection in GMPLS networks with a Path Computation Element (PCE).

References

[1] F. Aslam et al. *Interdomain Path Computation: Challenges and Solutions for Label Switched Networks*. IEEE-Communications Magazine Volume 45 (10), pp. 94 - 101, Oct. 2007.

[2] Admela Jukan et al. *End-to-End Service Provisioning in Multi-granularity Multi-domain Optical Networks*. In Proceedings of ICC 2004, IEEE International Conference on Communication, June 2004.

[3] Srinivasan Seetharaman et al. *End-to-End Dedicated Protection in Multi-Segment Optical Networks*. Technical report, http://www.stanford.edu/s̃eethara/papers/e2eprot.pdf, 2003.

[4] Thomas Engel et al. *Increasing End-to-End Availability over Multiple Autonomous Systems*. In Proceedings of PDPTA'05, Las Vegas, USA, 2005.

[5] Thomas Schwabe et al. *Resilient Routing Using ECMP and MPLS*. In Proceedings of HPSR 2004, Phoenix, AZ, USA, 2004.

[6] D-L Truong et al. *Recent Progress in Dynamic Routing for Shared Protection in Multidomain Networks*. IEEE Communications Magazine, Vol 46 (6), pp. 112-119, June 2008.

[7] H. Drid et al. *A survey of survivability in multi-domain optical networks*. Computer Communications Vol 33 (8) pp. 1005-1012, May 2010.

[8] J. Szigeti et al. *Adaptive Multi-Layer Traffic Engineering with Shared Risk Group Protection*. In Proceedings of IEEE International Conference on Communications, ICC '08, pp. 5367-5371, 2008.

[9] J. Hawkinson and T. Bates. *Guidelines for creation, selection and registration of an Autonomous System, RFC 1930*. IETF standards Track, March 1996.

[10] D. Staessens et al. *A Quantitative Comparison of Some Resilience Mechanisms in a Multidomain IP-over-Optical Network Environment*. In Proceedings of IEEE International Conference on Communications, ICC '06, pp. 2512-2517, 2006.

[11] ITU-T Standardization Organization. *ITU-T Recommendation G.807/Y.1302, "Requirements for automatic switched transport networks (ASTN)*. July 2001.

[12] J. P. Vasseur et al. *Network recovery, protection and restoration of optical, SONET-SDH, IP and MPLS*. Elsevier, 2004.

[13] P. Demeester et al. *Resilience in multi-layer networks.* IEEE Communications Magazine, August 1998.

[14] D. Staessens et al. *Enabling High Availability over Multiple Optical Networks.* IEEE Communications Magazine, Vol 46 (6), pp. 120-129, June 2008.

[15] P. Cholda et al. *Reliability Assessment of Optical p-Cycles.* IEEE/ACM Transactions on Networking,Vol. 15, no. 6, pp. 1579-92, Dec. 2007.

[16] Akyamac et al. *Reliability in Single Domain vs. Multi Domain Optical Mesh Networks.* In Proc. NFOEC, Dallas TX, Sept. 2002.

[17] Lewis et al. *Introduction to Reliability Engineering.* John Wiley & Sons, 1987.

[18] J.W. Suurballe and R.E. Tarjan. *A quick method for finding shortest pairs of disjoint paths.* Networks, vol. 14, pp. 325336, 1984.

[19] IBM ILOG CPLEX Optimizer. *http://www-01.ibm.com/software/integration/optimization/cplex-optimization-studio/.* 2012.

[20] A. Farrel et al. *Inter-Domain MPLS and GMPLS Traffic Engineering Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions, RFC 5151.* IETF standards track, 2008.

[21] CY. Lee et al. *GMPLS Segment Recovery, RFC 4873.* IETF standards track, 2008.

[22] L. Berger et al. *Exclude Routes - Extension to Resource ReserVation Protocol-Traffic Engineering (RSVP-TE), RFC 4874.* IETF standards track, 2008.

[23] D. Awduche et al. *RSVP-TE: Extensions to RSVP for LSP Tunnels, RFC 3209.* IETF standards track, 2001.

[24] J.P. Lang et al. *RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery, RFC 4872.* IETF standards track, 2008.

5 Impact of transparency on the cost savings in future networks

5.1 Introduction

The technological advances detailed in Chapter 2 have opened up new perspectives in the design of cost-effective optical transport networks [1]. Introduction of transparency in the network allows for a reduction in optical-to-electrical-to-optical (OEO) regenerators (See Chapter 2). Furthermore, for survivable networks, capacity sharing also promises a reduction in network cost [2]. There has been a lot of research into the overall cost savings of transparent networks [3] [4] compared to traditional opaque networks. In this chapter, we perform an extensive study on the CAPEX reduction when introducing protection mechanisms into survivable networks with respect to transparency. We first investigate the influence of the node architectures discussed in Chapter 2 on the capability to perform restoration (Section 5.2). This study was performed on the GEANT2 topology (Figure 2.14) and on a number of generated topologies. The possible cost savings obtained through sharing backup resources in transparent and opaque networks are investigated in Section 5.3. We first perform dimensioning studies on the DT topology (Figure 2.12), and then extend the study with a randomized set of 1000 networks. From these studies we conclude that, while protection sharing significantly reduces the load in the network, transparent networks have less benefits in terms of node cost when the traffic is low (in terms of the number of wavelengths in the network). When we scale up the traffic demand, and ROADM node degrees increase, the

CapEx gain for transparent networks approaches the same levels (roughly 20 %) as for opaque networks. We also find that, for opaque networks, the CapEx gain through protection sharing is independent of the traffic scaling.

5.2 Impact of node directionality on restoration in transparent networks

In order to evaluate the impact of directionality on the capability of the optical layer to restore connections (without additional transponders), we simulated restoration for the pan-European network based on the Geant2 network topology, Figure 2.14, simulating behavior of the directional ROADM node architecture from Figure 2.8. The traffic matrix is routed over the shortest available physical path. We then simulate each link failure and check if the affected connections are restorable under the directionality constraints at the source and destination node and also at intermediate regeneration points. In the translucent case, we require that the regenerators used for the restored connection are the same as for the affected connection. Furthermore, connections can also be unrestorable due to the transparent length constraint. We assess the average fraction (between 0 and 1) of the affected traffic (meaning the lightpaths that are disrupted due to the considered failure) and of the total traffic that is unrestorable assuming different transparent lengths, ranging from regeneration at every node (opaque) to fully transparent. Note that for the opaque case, we simply use regeneration at every node, we do not consider a different node architecture.

Simulation results are summarized in Figure 5.1. The figure shows that in the opaque case all traffic is unrestorable in the optical layer after any failure. This was of course expected, since all paths are one hop, and all 1-hop and 2-hop paths are unrestorable. For a transparent length of 1000km, some paths were restorable under the directionality constraint, however, all of them exceeded the maximum transparent length (MTL). Increasing the transparent length will allow longer paths, both physically, but, more importantly, in hop count, which will give better restoration opportunities. For a 2500 km MTL (which is a realistic value for 10G long haul transmission), we notice that roughly 90 percent of affected traffic is still unrecoverable, of which only 6-7 percent is due to exceeding the transparent length on the restoration path. In the fully transparent case, the baseline performance is that almost 70 percent of the affected traffic after a link failure is unrecoverable in the optical layer.

Figure 5.2 shows the same results, but for the total traffic in the network. This graph takes into account the amount of traffic on each link, where the previous figure treats all links as equal. It shows that increasing the transmission distance increases the amount of traffic that can be restored significantly, where in the opaque

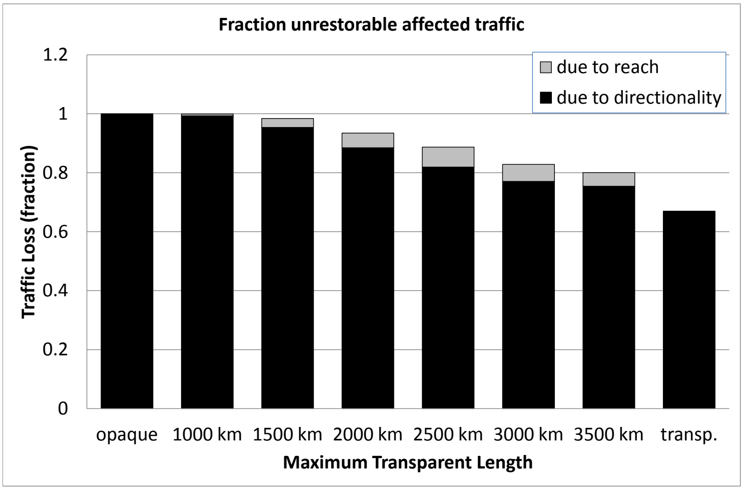


Figure 5.1: Fraction unrestorable affected traffic for GEANT2 topology.

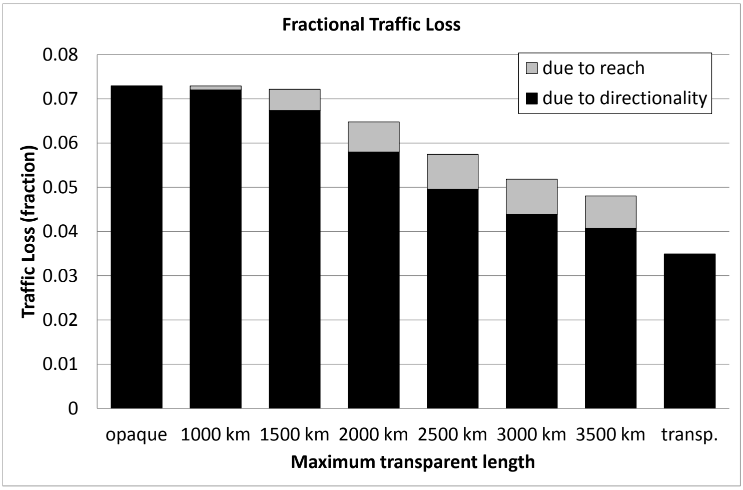


Figure 5.2: Fraction unrestorable traffic for GEANT2 topology.

case, on average 7.2% of the total traffic load is lost after a failure, compared to 3.4% in the transparent case. This is only partly due to the increased path length (less restoration paths are exceeding the transparent length) and mainly due to the removal of directional regeneration points.

There should be a direct relation between the amount of unrecoverable traffic and the average path length in the network. In order to evaluate this relation, we generated a large number of random (connected) network topologies with 10, 30 and 100 nodes and different average node degrees. These topologies range from trees to very dense mesh topologies. Note that for realistic telecommunication networks, typical node degrees are between 2 for ring networks and up to 3.5 for very dense mesh networks. This corresponds to average path lengths of 1.3 to 3.4 for 10-node networks, between 2.5 and 5 for 30-node networks and between 3.5 and 6 for 100-node networks. The traffic for these networks is assumed to be a full mesh of bidirectional lightpaths between all pairs of nodes.

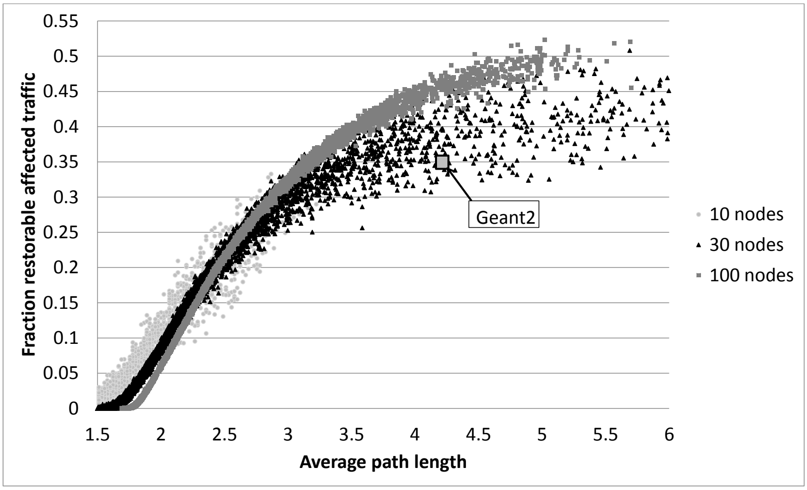


Figure 5.3: Fraction restorable traffic for random network topologies.

For each of these topologies, we calculated the average path length (hops) and the fraction of traffic that can be restored assuming full transparency. The results plotted in Figure 5.3 clearly show the relationship between the path length distribution in the network and the amount of restorable traffic. The longer the average path length in the network, the higher the fraction of traffic that is restorable. Note that the graph shows the fraction restorable traffic versus the affected traffic. Increasing the path length in the network will increase the fraction of the affected traffic that is restorable, but will also increase the total load on this link, meaning that the total number of failed connections will actually increase. Also, the larger the number of nodes in the network, the better the fraction of restorable traffic. This is even more valid if we take into consideration the average path lengths for realistic networks mentioned above. In these ranges, for 10 node networks, the fraction is roughly 10-25%, for 30-node networks 20-40% and for 100 node networks up to 50%. These results also coincide with our results for the Geant2 topology, which has roughly 35% restorability (34 nodes with an average path length of 4.12 hops).

5.3 Cost evaluation of backup resource sharing in future networks

The following sections present results for the cost estimation (Capital Expenditures or CapEx) for opaque and transparent networks, designed for a given traffic demand and using different protection schemes. Dimensioning the network requires us to calculate routes and assign the wavelengths to be used for each traffic demand, called routing and wavelength assignment (RWA). Optimized RWA for minimizing resource usage and blocking in wavelength-switched networks is NP-complete [5]. For the unprotected and dedicated protection solutions, we follow an R+WA scheme: we first calculate the route and then assign an available wavelength for that route using first fit wavelength assignment. The paths are calculated using Dijkstra's algorithm [6] in the unprotected case, the algorithm by Suurballe and Tarjan [7] is used for link-disjoint and node-disjoint cases. For mesh shared protection, optimized RWA becomes a very complex optimization problem which is well researched. ILP formulations [8] and approximation algorithms [9] have been proposed.

Because resource sharing optimization is a complex problem demanding considerable computation resources to find an optimum solution, we use a dimensioning for restoration as a compromise for an optimized mesh restoration scheme. The drawback of this approach is that it is not feasible to implement restoration on the transparent architecture due to directionality: in the architecture in Figure 2.8 we cannot reuse a transponder if its outgoing link fails, because it is tied to this one direction (See also the previous section). In order to really implement restoration, we would need additional transponders. This means we will have an underestimate of the transponder cost of restoration and 1:1 protection in the transparent solutions. In the approach we implemented, due to the possibility that the restoration path for a failed working path can use different outgoing links (as opposed to a single fixed one for 1:1 protection) for different failures along the path, we underestimate the transponder cost for restoration more than we underestimate the transponder cost for 1:1 protection. This means that, when comparing 1:1 protection to shared mesh protection / restoration in the transparent case, we have an overestimation of the CapEx benefits of protection sharing in the transparent case. Also note that the 1+1 protection scheme can be implemented on the ROADM architecture because all transponders are dedicated protected.

In summary, we consider the following protection schemes:

- *Unprotected.* All traffic is routed over the physical shortest paths, calculated using Dijkstra's algorithm, using 10G wavelengths.
- *Link / node restored.* This serves as a benchmark dimensioning for a shared mesh restoration scheme (1:1 shared protection). All traffic is routed over physical shortest paths. For each failure scenario (all possible single link failures for link restored and all possible link and node failures for node-restored) we calculate the required network resources required and determine the minimum capacity which is needed to cover all of the failure scenarios.
- *Link / node 1:1 dedicated protected.* All traffic is routed over physical shortest cycles, calculated using the Suurballe-Tarjan algorithm [7] for link-disjoint. For node-disjoint, we run the same algorithm on a modified directed graph where each node is split into two nodes, one containing the incoming edges, one containing the outgoing edges and a single directed edge is added between them from the node with the incoming edges to the node with the outgoing edges. The working path is the physically shorter half of the cycle.
- *Link / node 1+1 dedicated protected.* Uses the same paths as the 1:1 protected, only the traffic is duplicated and sent over both working and backup paths, meaning we also protect (and need to duplicate) the transponders.

As mentioned before, we use a two-step R+WA approach: we first determine the path using the algorithm details above, and then assign the appropriate wavelength(s) using first fit assignment. We evaluate the cost of these different protection schemes on the national backbone reference network from Chapter 2 (Figure 2.12). The most relevant characteristics are given in Table 2.2 and the traffic matrix is given in 2.3.

5.3.1 Link capacity usage

Figure 5.4 shows the total used link capacity in the network for the different recovery mechanisms. If wl_l^w is the number of working paths traversing link l and wl_l^b

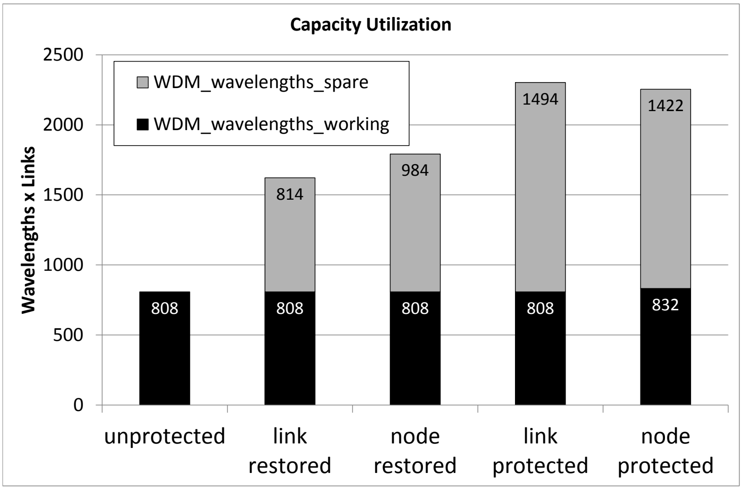


Figure 5.4: Link capacity utilization for the DTAG topology

is the number of backup paths traversing link l, the total wavelength consumption in (wavelengths x links, WL) for the network with m links is calculated as

$$\sum_{l=1}^{m} wl_l^w + wl_l^b \tag{5.1}$$

These values are valid for both the transparent and opaque architectures, as the routing schemes used for both architectures are the same. Note that for 1+1 protection the values are the same as for 1:1 protection and therefore not shown in the figure. Of course, in 1+1 protection the spare capacity is occupied; which is not the case for 1:1 protection, where the spare capacity can be used for low priority traffic.

We see that for the unprotected network we require 808 WL. If we want to be able to restore every lightpath in case of all possible link failures, we require an additional 814 WL (totaling 1622 WL) and for all possible link and node failures, 984 extra WL (totaling 1792 WL). Protection clearly requires a lot more resources. In the DT topology we used, the capacity for the working paths in the protection case is the same as the unprotected and restored cases, however, due to the absence of capacity sharing, the link protected-network requires 1494 WL extra and the node-protected network 1422 WL.

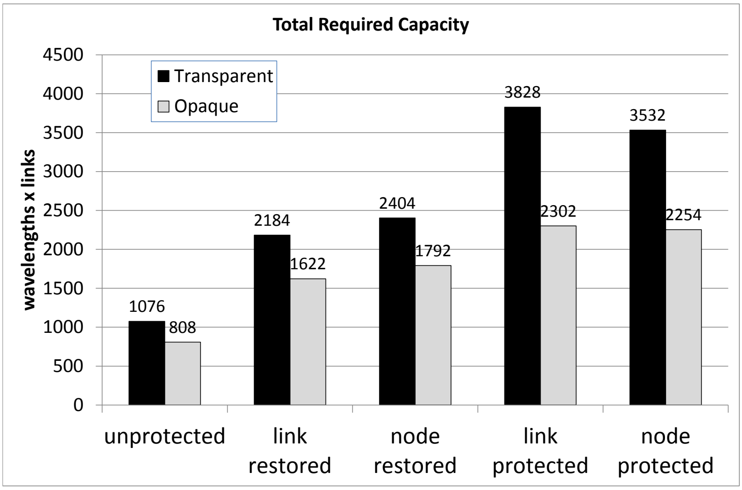


Figure 5.5: Link capacity requirement for the DTAG Topology

This figure does not tell the whole story. To accommodate all the active lightpaths, the transparent network needs more wavelengths due to the wavelength continuity constraint (see chapter 2). If the wavelength channels on a link are numbered starting from 1 in increasing order (for instance, according to the ITU DWDM 50Ghz frequency grid [10]), and the highest used wavelength channel on link l is wc_l, then the total required capacity is calculated according to

$$\sum_{l=1}^{m} wc_l \tag{5.2}$$

For the opaque solution this amounts to the sum of the working capacity and spare capacity from Figure 5.4, all wavelengths which are available are also used. We immediately see that, when compared to the opaque solution, the transparent solution requires a lot more resources. For the unprotected case, we require 1076 WL (or 33% more resources), meaning that the links are utilized only for 75% (the active lightpaths consume 808 WL) due to the wavelength continuity constraint. For the node restored case the situation is similar (35% extra resources or 74% link utilization). For the node protected case, the increase is much more prominent, requiring almost 66% extra resources due to the wavelength continuity constraint (60% link utilization). This is due to the longer paths required for the backup paths, since they have to be disjoint from the working paths. If we compare the so-

Equipment	Cost
WDM layer	
Transponder 10G grey	0.1
Transponder 10G 2000km	1.2
N degree ROADM ($N \leq 8$)	$N \times 9.2$
N degree ROADM ($9 \leq N \leq 19$)	$N \times 11.8$
OTN Layer	
Transceiver grey 10G	0.1
Transceiver 10G 2000km	1.1
Line card 10x10G	16
Basic node 8 slot	7
Basic node 16 slot	14.3
Basic node 32 slot	28.6
Basic node 64 slot	67
Basic node 128 slot	154

Table 5.1: Cost model

lutions for node protection to node restoration, we see a 21% (2404 WL vs. 3532 WL) decrease in link resource consumption for the opaque case and an even larger 32% reduction in WL consumption (2404 WL vs. 3532 WL) in the transparent case. We can attribute this to the wavelength continuity constraint in transparent networks.

There is one peculiarity to these results. The attentive reader will undoubtedly have noticed that, in contradiction to common sense, the link-protected solution consumes more resources than the node-protected solution. This is due to the four nodes in close vicinity of each other in the DT topology (the link length distribution is not smooth) and the fact that we use a physical length shortest cycle, which in some cases routes through these nodes. If we use hop count instead of physical length in the routing algorithm, this does not occur.

5.3.2 Node capital expenditures

Now that we have shown a significant reduction in wavelength consumption for restoration compared to protection for both the opaque and transparent network architectures, we turn our attention to the Capital Expenditures (CapEx) of the nodes. The considered transparent node is based on the well-known broadcast-and-select ROADM architecture (See Section 2.5.5.1. ROADMs which have degree $N \leq 8$ use 1x9 WSSs and ROADMs with degree $9 \leq N \leq 19$ use 1x20 WSSs. The opaque node is based on the OTN cross-connect from Chapter 2. Each fiber has 80 available wavelength channels.

The CapEx of the nodes is broken down in three main components:

- Tributaries. The transmission equipment (transponders or transceivers) towards the client host or network.
- Transmission. These are the source and destination transponders (transparent) / transceivers (opaque), any intermediate transponders/transceivers and the OTN linecards (opaque).
- Switching. These are the switching fabric and AWG/terminals in the ROADM or the backplane (basic node) in the OTN cross-connect.

The cost of the transmission links is not considered in this study, because they will be the same in both solutions. The used cost model is based on the models from [11] [12] [13] and is being updated in the STRONGEST [14] project. The cost values used in this text are given in Table 5.1.

Figures 5.6 and 5.7 show the CapEx results for the transparent and opaque solutions respectively. It is immediately clear that the restoration (i.e. the solution with backup capacity sharing) is cheaper than protection in both solutions. In the transparent network, the cost difference is in the switching. If the capacity increases (between different recovery methods), the capacity of some links may exceed the number of wavelengths (80), so some nodes need a parallel line system (an extra ROADM degree) in order to accomodate this increase in traffic. For the opaque solution, the main cost is in the transmission equipment because we need 2 transceivers in every intermediate node for each traversing connection. There is also an increase in the switching cost due to larger backplane (basic node) requirements for protection when compared to restoration. Also, in the opaque solution, we see that the cost of the tributaries is negligible compared to the overall node cost.

If we compare the gain by implementing protection sharing/restoration, we see that for the opaque solution, the CapEx reduction is roughly 36% (4059.3 Cost Units (CU) vs 5474.9 CU), while for the transparent solution it is roughly 24% (958 CU vs 1261.6 CU)[1]. This shows that a reduction in wavelength consumption is definitely not a direct indicator for a similar reduction in network cost. The

[1] This result contradicted our previous results from [15], where we found no such advantage for the traffic from Table 2.3 and only a little advantage if we doubled the amount of traffic in the traffic matrix. The reason for the discrepancy lies in the fact that we did not take into account the wavelength continuity constraint in our first work. As was shown in Figure 5.4 the continuity constraint leads to almost 33% increase in traffic for the restoration case and a 66% increase in traffic for the protection case, effectively increasing the reduction in node CapEx gained by resource sharing in transparent networks. While we expected a small increase from including the wavelength continuity constraint, we never expected such a significant one. This find lead us to perform more extensive research in order to find the relation between the traffic and the node cost for transparent and opaque networks.

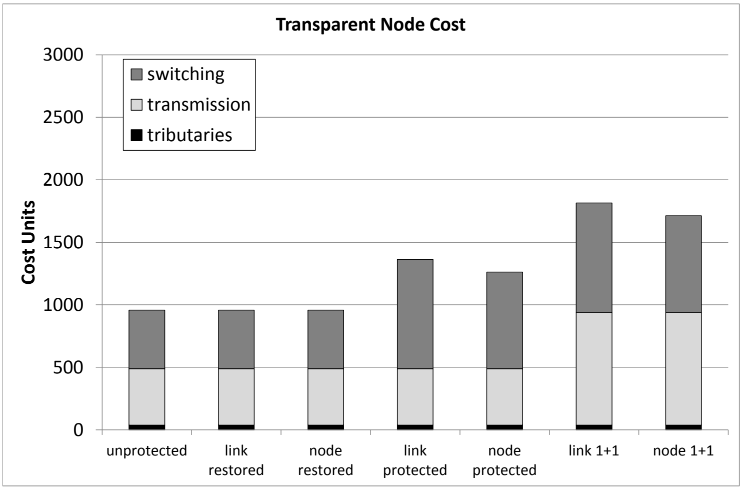


Figure 5.6: Transparent node cost

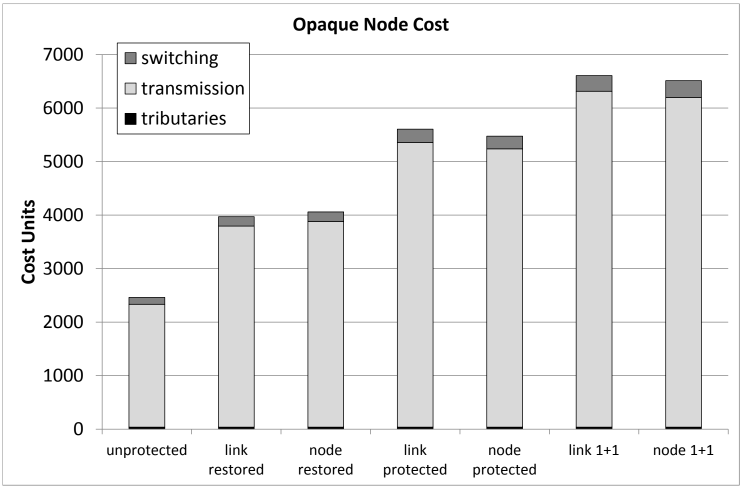


Figure 5.7: Opaque node cost

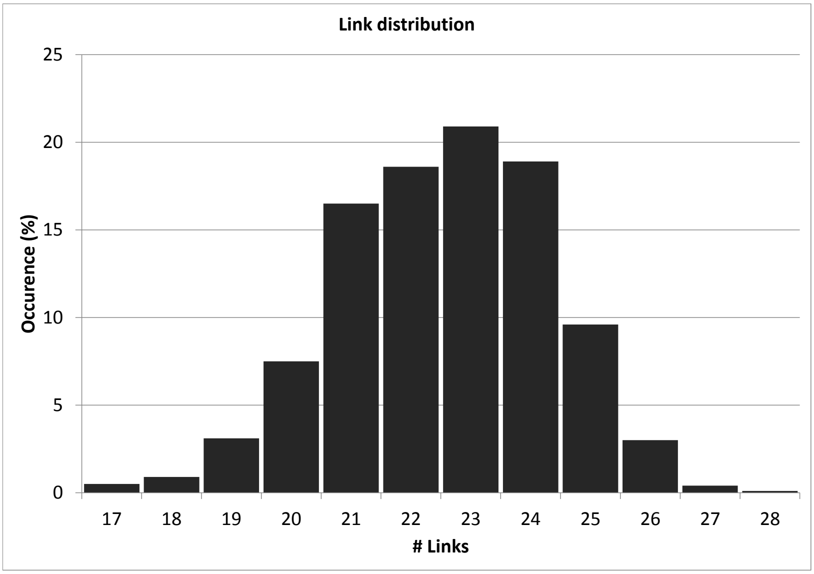


Figure 5.8: Link distribution for the generated topologies

CapEx reductions are far less outspoken than the wavelength consumption, moreover, where the wavelength consumption decrease was largest in the transparent network, the node cost decrease is larger in the opaque network. In the next section, we perform a thorough investigation how the node CapEx gain (through the introduction of resource sharing) scales with traffic demand and network meshedness in a randomized scenario.

5.4 Randomized study

In order to have a more meaningful analysis and evaluate the benefits of resource sharing more thoroughly, we extend our dimensioning study by using random generated 14-node networks as opposed to the single reference network from the previous section. We number the nodes 1-14 at random, and apply the traffic matrix from Table 2.3 to each of these networks to calculate and analyze the node CapEx.

We generated 2-node-connected planar graphs by randomly assigning 14 points to an 800km by 800km grid and computing the Gabriel graph [16] for these 14 points. We discarded all non 2-connected graphs until we had a population of 1000 random graphs. These graphs had a link distribution shown in Figure 5.8. It

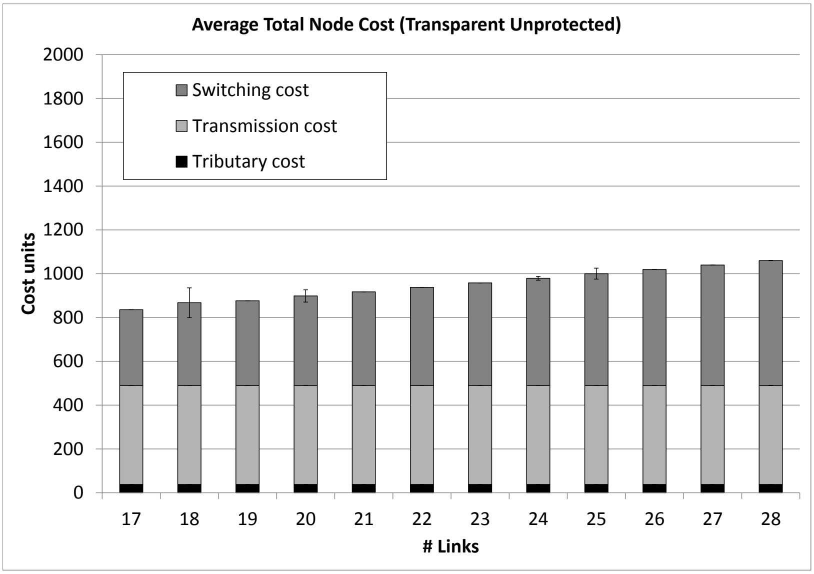


Figure 5.9: Average node cost, transparent unprotected

seems that the topology with 23 links (like the DTAG topology) is the most likely to occur.

5.4.1 Influence of the Topology

In this section we investigate the influence of the number of links and the topology on the node cost in both transparent and opaque networks. The results show that, for the traffic matrix from Table 2.3, for the transparent solution, the node CapEx increases for unprotected traffic, but is more or less stable if we apply resiliency. In the opaque solution, the node CapEx always goes down with increasing node degree.

Figure 5.9 shows the average node costs for the generated networks versus the number of links in the generated topologies for unprotected routing. 2σ confidence intervals are included (note that there is no variation for most of the unprotected networks). The cost for tributaries (451.2 CU) and the transponders (451.2) is the same for all solutions. Indeed, all networks (17-28 links) are transparent for all shortest paths. The cost of a transparent network goes up (from 836 CU to 1060 CU) with the number of available links. From Figure 5.9 we clearly see that this is due to an increase in switching cost (from 347 CU to 571 CU, a 63% increase), or more specifically, an increase in the degree of the ROADM node due to the

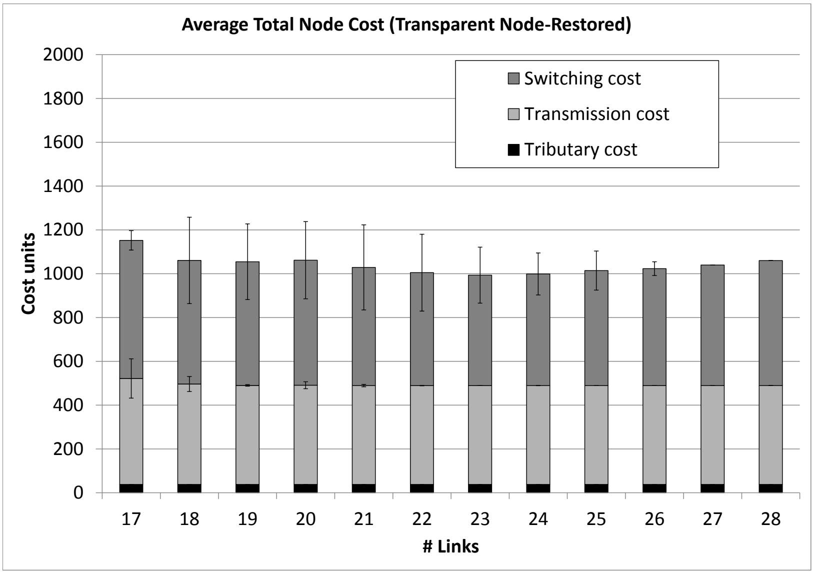


Figure 5.10: Average node cost, transparent node-Restored

increase in physical degree of the topology. For some networks (the 18,20,24 and 25) there is a slight variation in the cost of the switch due to some network topologies requiring parallel line systems. However, we can conclude that for this level of traffic, there is very little influence of the actual topology and only the number of links (and thus the average node degree) affects the node CapEx of the network.

Figures 5.10 and 5.11 show similar figures for node restoration and protection. We see that the monotonous increase of the cost vs. the number of links observed for the unprotected case is not present anymore and the node CapEx shows a more flat distribution with respect to the number of network links (the average node degree). In Figure 5.10, the transmission cost slightly decreases (484 CU to 451 CU) with an increase in links. This is somewhat expected, as the restoration path will be longer in sparse networks, requiring regeneration, which is implemented by terminating and continuing the traffic at an intermediate node, which means the need for additional transponders. The variation in the switching cost is also more present than in the unprotected case, meaning there is more dependence on the actual topology. We will investigate the effect of traffic increases in the next section.

Figure 5.11 shows the same behavior for the transmission cost. It goes slightly down with an increase in the number of links (from 493 CU to 451 CU). The transmission cost is slightly higher compared to the restored / unprotected case

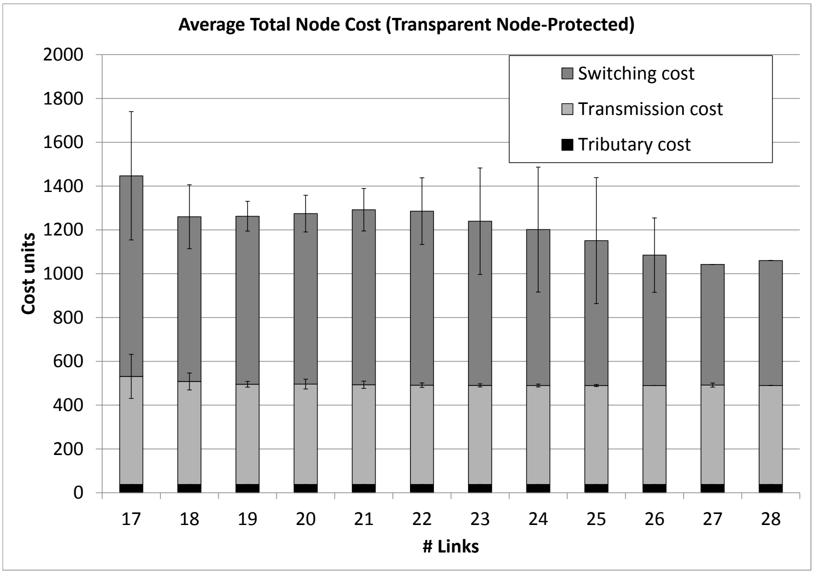


Figure 5.11: Average node cost, transparent node-protected

due to an additional increase in the length of the working/backup paths due to the use of a shortest cycle algorithm (vs. shortest path on the remaining topology after a failure in the restoration scenario).

If we look at the cost benefits of restoration vs protection (i.e. the difference between Figure 5.11 and Figure 5.10) we see that, for the 23 link network, in our generated topologies the gain is around 25% (993 CU vs 1240 CU). What is very peculiar is that the gain is higher for the medium meshed networks (21-24 link networks are all in the 20-25% range) than for the higher meshed networks (the gain for the 26 links network is already less than 10 %).

We now turn our attention to the opaque architecture. Figures 5.12 and 5.13 show the node costs for the node-restored and node-protected cases. We see that the cost of the network scales down with an increase in the number of links. This is because an increase in meshedness reduces the average hops on each path, which in turn reduces the number of O/E/O conversions and therefore the transmission cost. The cost reduction is almost 50%, with a node cost of 6666 CU for the 17 link network and a node cost of 3448 CU for the 28 link network. There is again little variation due to the actual topology as the 2σ confidence intervals are quite small, the larger values for 17, 18 and 27 (and infinite for 28) are due to the small data set for these networks.

When we compare the two solutions (restoration vs. protection), we again see

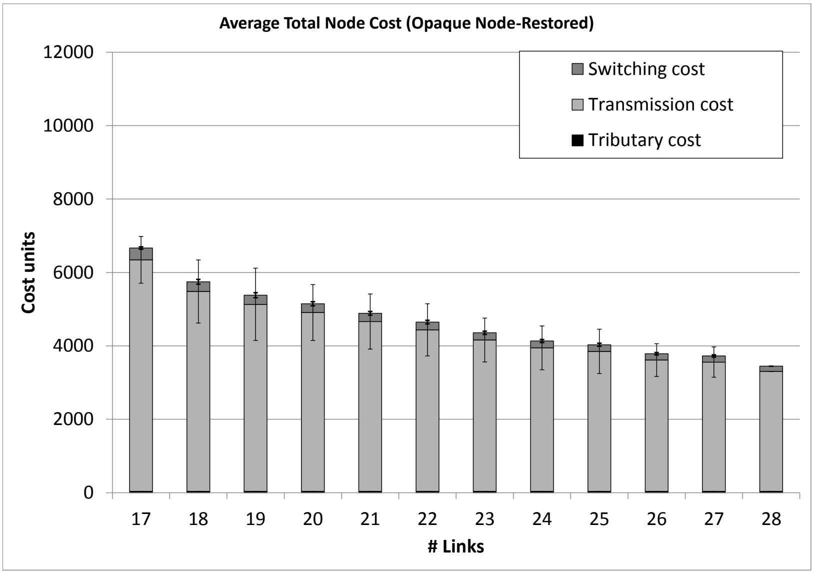


Figure 5.12: Average node cost, opaque node-restored

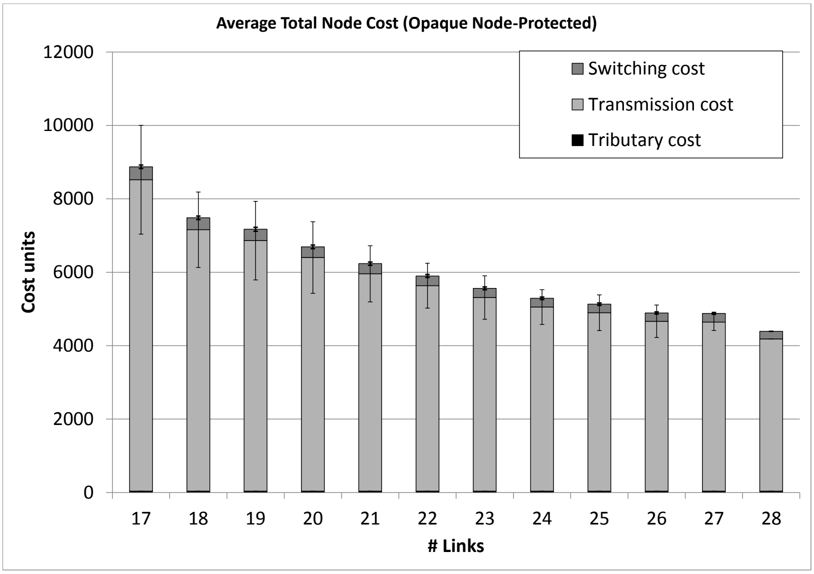


Figure 5.13: Average node cost, opaque node-protected

a significant node CapEx gain due to protection sharing which decreases slightly with the number of links in the network. The gain is 25% for a 19 link network (5380 CU vs 7171 CU) and 23 % for a 26 link network (3785 CU vs. 4890 CU).

5.4.2 Influence of Traffic Scaling

In order to evaluate the effects of the traffic load, we scaled the traffic from Table 2.3 from 50% to 500% in 50% increments. From a multiplier of 3-3.5x onwards, the ROADM degree of some node exceeds 19 (which is the limit set by the use of 1x20 WSSs), and the OTN backplane reaches its limits (128 slots) at 4.5-5x. We therefore limit our results to a traffic multiplier for 3x for the transparent case and 4.5x for the opaque case.

From Figure 5.14 it is clear that the increase in node cost with the number of links for transparent networks we noticed in the previous subsection is only valid for the low traffic cases where there is little increase in ROADM degrees through the necessity for additional parallel line systems. The slightly increasing slope for multiplier values 0.5x and 1x turns to a fairly constant line for a multiplier of 2x and becomes decreasing if we further increase the multiplier. The figure also shows that the dependency on the topology is independent of an increase in traffic. The 2σ confidence intervals on the cost become larger the further we scale the traffic, but the increase is linear, always around 14% of the traffic value. Note that we only show the confidence intervals for 0.5x, 1x, 2x, 3x, 4x and 5x in order to avoid cluttering the figure. The only exception to this 14% rule is the bottom line in Figure 5.14, where there is no variation. Due to the low amount of traffic, all traffic could always be routed transparently (except for the 17 and 18 node networks which show very small variation). Figure 5.15 shows the same data, but for the protection case.

We summarize the relative CapEx gain for the transparent networks in Fig. 5.16. We include the networks with 20-25 links and apply the traffic multiplier from 0.5x to 3x. We omit the other cases because of some ROADM degrees exceeding the limits set by the use of 1x20 WSSs as noticed before. What we learn from this figure is that the relative gain through protection sharing in transparent networks is very dependent on the traffic scaling. For low traffic there is almost no resource gain (less than 5% for the 23 node network). From the moment the traffic loads exceeds a certain threshold (here it's roughly at the 1x multiplier), the average ROADM degree in the network goes up and the relative decrease in traffic load needed to reduce the degree goes down. In turn, the probability of this happening goes up significantly. If we have a node with 3 neighbours in the physical topology, reducing it from a 14-degree to a 12-degree ROADM takes less of a rela-

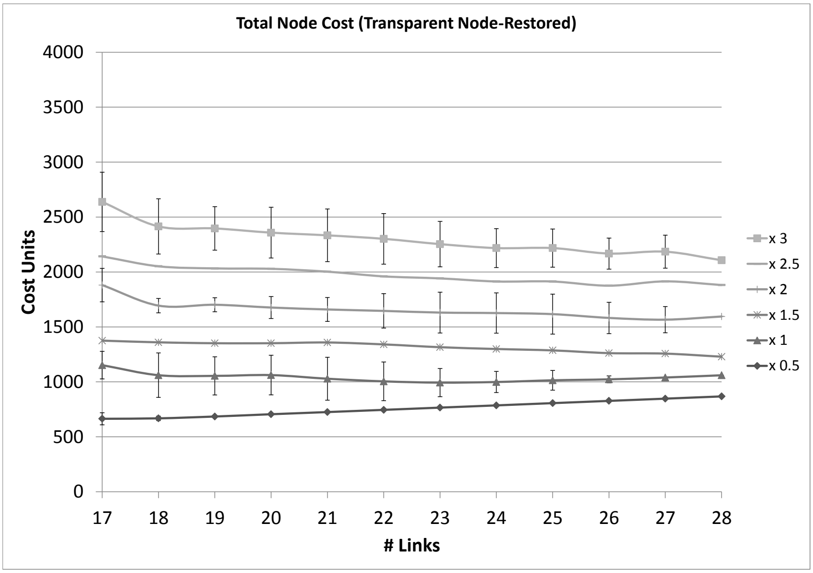


Figure 5.14: Total node cost for different traffic multipliers, transparent node-restored

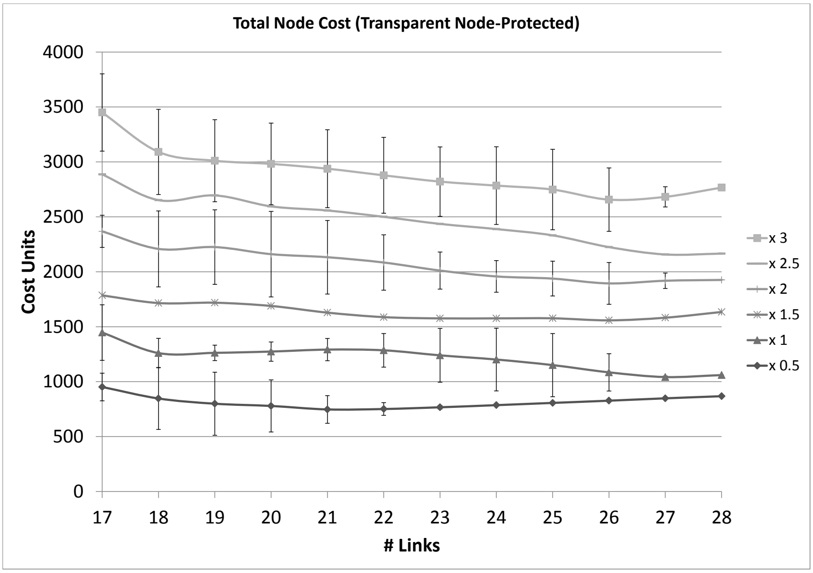


Figure 5.15: Total node cost for different traffic multipliers, transparent node-protected

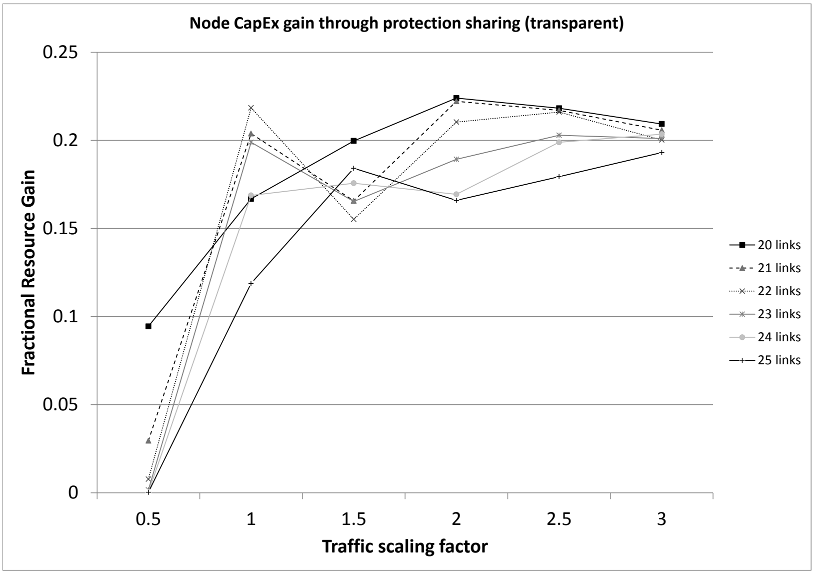


Figure 5.16: CapEx gain through resource sharing in transparent networks

tive traffic reduction than to reduce it from a 4-degree to a 3-degree ROADM. The wavelength continuity constraint is certainly an important contributor in speeding up this process. After this threshold is reached the CapEx gain of resource sharing is roughly 17-22%.

For opaque networks, the overall picture is quite different. As shown in Figs. 5.17 and 5.18, the total node cost always goes down with the number of links in the network, no matter the load. Also, it is independent of the load. For instance, in the node-restored case the relative gain from 20 links to 25 links is 21% for 0.5x traffic (2490 CU vs 3156 CU) and also 21% (12839 CU vs 16457 CU) for 4x traffic. It's also independent of the actual underlying topology, since the confidence intervals are very small (less than 3% overall). This decreasing trend has as a result that operators will be able to find an optimum between the additional link cost (for increasing the node degree) and the decreasing node cost (due to the shorter paths in the network). Remember that transparent networks with low traffic do not have this and have a decrease in node cost together with a decrease in link cost, always driving the optimum towards sparsely meshed networks. This may give transparent network operators additional incentives to prefer higher bandwidths per channel and more wavelengths per fiber instead of installing parallel line systems.

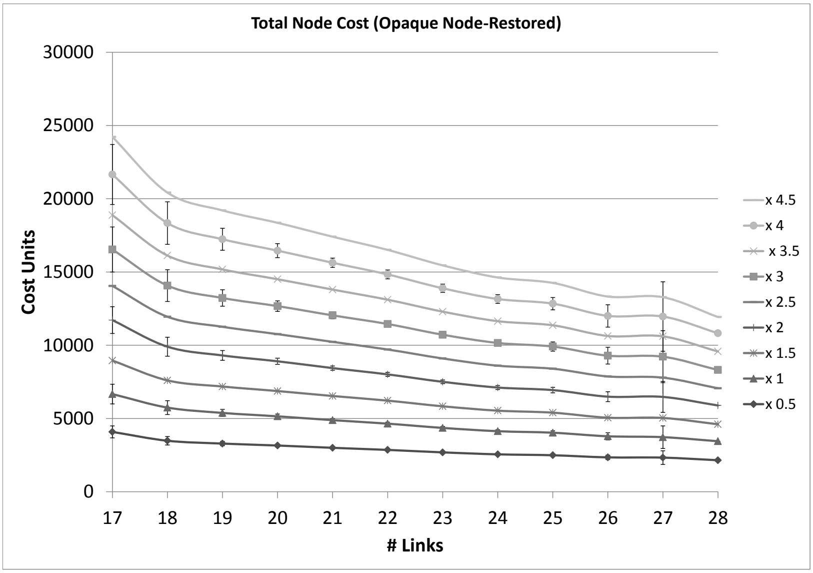


Figure 5.17: Total node cost for different traffic multipliers, opaque node-restored

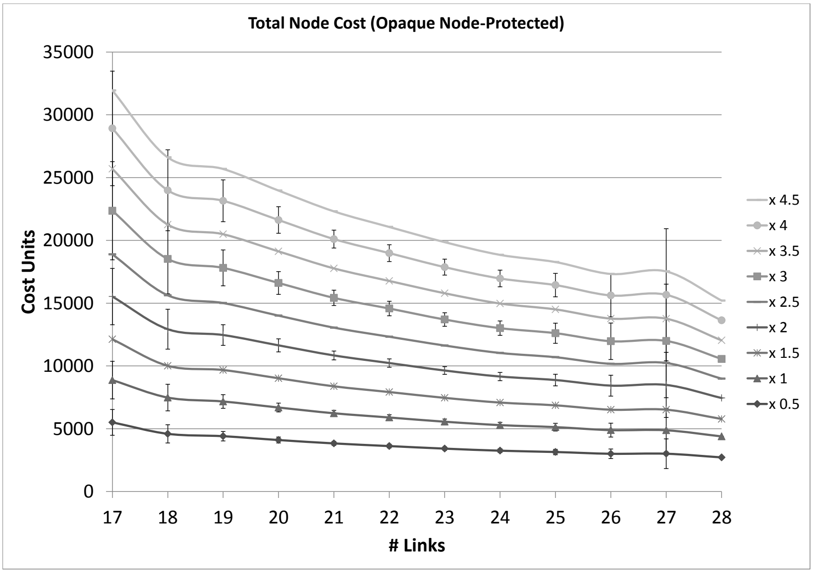


Figure 5.18: Total node cost for different traffic multipliers, opaque node-protected

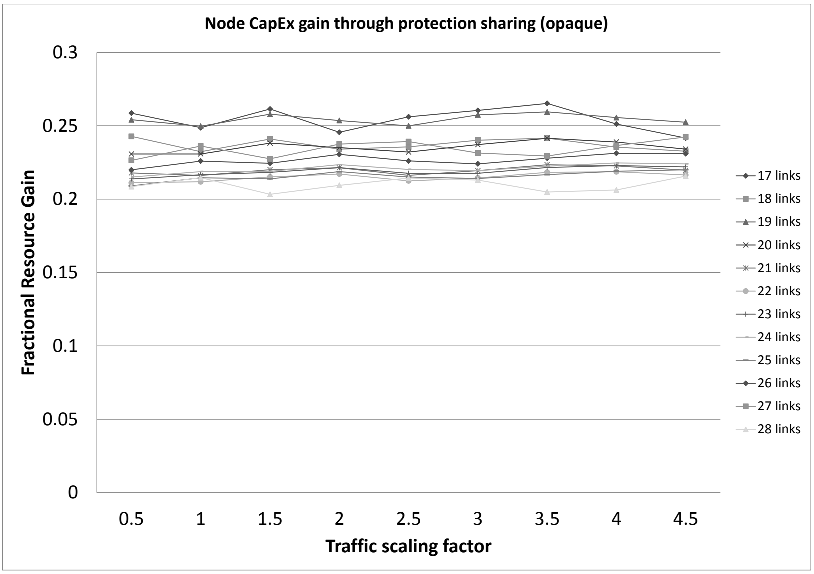


Figure 5.19: CapEx gain through Resource Sharing in opaque Networks

When comparing the overall cost reduction from Figure 5.16 and 5.19 we see that the node CapEx reduction for transparent networks in the high load case (17-22%) is definitely comparable to the CapEx reduction in the opaque architecture (21-25%). We find this quite surprising as the load reduction through resource sharing in transparent networks only affects the ROADM degree (See Figure 5.10) and therefore only the cost of the WSS drives this reduction. In opaque networks the cost reduction is driven through a reduction in the number of required transceivers, linecards and a reduction in the size of the switching fabric.

5.5 Conclusions and future work

In this chapter, we evaluated the impact of transparent node architectures on network survivability. First we investigated the impact of directionality on restoration capabilities in the network. Following a link failure in such networks, only a fraction of the affected traffic is restorable in the optical layer. The average fraction of affected traffic that is restorable is in direct relation to the average path length in the network and the number of nodes. Increasing the transparent length will increase the restoration opportunities by increasing the average path length and decrease the number of regenerators needed in the network. We also performed a thorough investigation into the possible CapEx saving through resource sharing in transparent and opaque transport networks. The number of links in the network has a small impact on this gain, with sparsely meshed networks having greater benefit than densely meshed networks. We find that the load has an important impact in transparent networks, where low load (i.e., few parallel line systems) means that the network does not benefit greatly from protection sharing. However, when the average required ROADM degree increases, the CapEx benefits approach the same levels as for traditional opaque networks. Opaque networks do not show a dependency on the load and always have a similar node CapEx gain from protection sharing. With the ongoing trend towards higher bit rates (400Gb/s and up), denser channel spacing and more efficient spectrum usage (Flexigrid), we think the balance for transparent networks will tip over towards the low load solution, meaning protection sharing may be less interesting to implement in such networks.

References

[1] Biswanath Ramamurthy et al. *Transparent vs. opaque vs. translucent wavelength-routed optical networks*. In Proc. Optical Fiber Communications, 1999.

[2] S. De Maesschalck et al. *Intelligent Optical Networking for Multilayer Survivability*. IEEE Communications Magazine, pp. 42-49, June 2002.

[3] A. Morea et al. *A critical analysis of the possible cost savings of translucent networks*. In Proceedings of DRCN '05, 2005.

[4] G. Li et al. *Resilience design in all-optical ultralong-haul networks*. J. Opt. Netw. 5, pp. 625-636, 2006.

[5] R. Ramaswami and K.N. Sivarajan. *Routing and Wavelength Assignment in All-Optical Networks*. IEEE/ACM Transactions on Networking Vol 3 (5), pp. 489-500, 1995.

[6] E.W. Dijkstra. *A Note on Two Problems in Connexion with Graphs*. Numerische Mathematik 1,pp. 269-271, 1959.

[7] J.W. Suurballe and R.E. Tarjan. *A quick method for finding shortest pairs of disjoint paths*. Networks, vol. 14, pp. 325336, 1984.

[8] P.-H. Ho et al. *Spare capacity allocation for WDM mesh networks with partial wavelength conversion capacity*. In High Performance Switching and Routing, 2003, HPSR. Workshop on, 2003.

[9] Y. Liu et al. *Approximating Optimal Spare Capacity Allocation by Successive Survivable Routing Yu*. IEEE/ACM Transactions on Networking, vol 13 no 1, February 2005.

[10] ITU-T Standardization Organization. *Spectral Grids for WDM, ITU-T recommendation G.694.1*. 2001.

[11] M. Gunkel et al. *A Cost Model for the WDM Layer*. In Proc. Photonics in Switching, 2006.

[12] R. Huelsermann et al. *Cost modeling and evaluation of capital expenditures in optical multilayer networks*. OSA Journal of Optical Networking, vol. 7, no. 9, 2008.

[13] M. De Groote et al. *Cost comparison of different Translucent Optical Network Architectures*. In Proc. Conference of Telecommunication, Media and Internet Techno-Economics, CTTE 2010, 2010.

[14] STRONGEST. *Scalable, Tunable and Resilient Optical Networks Guaranteeing Extremely-high Speed Transport.*

[15] D. Staessens et al. *Cost Efficiency of Protection in Future Transparent Networks.* In Proceedings ICTON '09, July 2009.

[16] K. R. Gabriel and R. R. Sokal. *A new statistical approach to geographic variation analysis.* Systematic Zoology (Society of Systematic Biologists) 18 (3) pp. 259270, 1969.

6
Conclusions

The past 2 decades the growth in Internet traffic is explosive, with a projected annual growth of 32% in the coming years. While Web browsing and Peer-to-Peer file sharing have been the predominant bandwidth consumers until just recently, online video (such as Youtube offering High Definition clips) is now the dominant driver for Internet growth, both in terms of average traffic and peak traffic. Even if consumption of video services is still different between average users and top users, which still use a lot of Peer-to-Peer solutions for instance, real-time online video services (i.e. streaming or progressive download of live events) are now mainstream. In 2010, only 3 percent of Internet traffic originated with non-PC devices, but by 2015 the non-PC share of Internet traffic will grow to 15 percent. Moreover, with the quick development of powerful mobile devices and the increased bandwith of 3G mobile access networks, users are also using more demanding applications on the go. These key factors will accelerate the development of mobile Internet usage in the coming years, with growth above 100% per year, with online video representing more than 60 to 70% of mobile Internet traffic. From the corporate and institutional point of view, almost all major strategic functions are dependent on communications between key offices, located in different countries and often even on different continents. Furthermore, there are recent advances in telemedicine and telesurgery which have huge bandwidth and reliability requirements. This fuels a quest for reliable interconnections over huge distances, often spanning multiple networks.

The popularity of the Internet Protocol and the huge bandwidth available in optical fiber are driving the network architecture towards multilayer networks, where an IP packet layer is supported by an optical transport network. These are the networks we focused on in this dissertation. Building on the available expertise in multilayer networks and reliability present at the IBCN research group, we answered the questions raised by the increased demands for bandwidth and reliability.

We investigated how we can establish highly reliable connections spanning multiple networks to enable demanding applications such as corporate VPN connections or High Definition Video-on-Demand being run over large distances. The solution presented not only targets high availability, but also tries to maximize resource efficiency. In order to do this, we extended the concepts from multilayer networks towards multilayer multidomain networks, and evaluated the schemes. Just as in the single domain case, the common pool solution, a highly efficient scheme for sharing protection resources in a network, was the most efficient for multidomain networks as well. We further refined the solution to a structure which has to be setup between 4 gateways in the intermediate domains. We proved that the solution can be calculated in any 2-connected network topology, which is quite important because this means that we can establish the connection per domain in a linear way without running into situations where a global solution exists but is not found because of a conflicting local solution in some domain (cfr. the trap topologies from Chapter 2). This greatly reduces the complexity of path establishment, removing the need for any backtracking mechanisms. The proof was constructive, which led immediately to a heuristic algorithm for calculating the common pool structure per domain. We also developed a mathematical optimum solution using Integer Linear Programming optimization techniques. Our results show that the heuristic already performs very well on average (well within 5% of the optimum), but the ILP can outperform the heuristic up to 30% in some particular cases, which shows there is definitely room for future improvement. We also designed a new routing object for RSVP-TE, the GSRO (Gateway Specification Routing Object) which allows signaling of the proposed solution in networks using a GMPLS control plane.

With the increased deployment of optical circuit switches based on ROADM designs, we also tried to answer how this architecture impacts protection and restoration in the network. First we have shown that directional ROADM designs severely limit the capability of the network to perform restoration. Due to the lack of flexibility, only a small amount of traffic can be restored in the optical layer. We also performed a thorough investigation into the possible CapEx savings through resource sharing in transparent and opaque transport networks. The number of links in the network have a small impact on this gain, with sparsely meshed networks having greater benefit than densely meshed networks. We found that the network load plays an important role in transparent networks, where low load (i.e.

few parallel line systems) means that the network does not benefit greatly from protection sharing. However, when the average required ROADM degree increases, the CapEx benefits approach the same levels as for traditional opaque networks. It will depend on the relative cost evolution of WSSs compared to transponders whether the cost benefits of protection sharing for transparent networks will increase or decrease. Opaque networks do not show a dependency on the load and always have a similar node CapEx gain from protection sharing. With the ongoing trend towards higher bitrates (400Gb/s and up), denser channel spacing and more efficient spectrum usage (Flexigrid), we think, at least for the near future, the balance for transparent networks will tip over towards the low load solution, meaning protection sharing may be less interesting to implement in the optical layer.

There are some recent advancements in the field of networking which are very interesting to consider. The work which is done in this thesis was based on fixed grid DWDM with homogenous channel rates (10 Gb/s). Recently a more flexible solution to improve spectrum usage was proposed, called Flexigrid. In a flexigrid network, channels can occupy spectral bands in multiples of 12.5 GHz (with 6.75 GHz proposed) which allows higher spectral efficiency and higher bitrates per channel (100+ Gb/s for long haul). Including flexigrid, especially in a mixed linerate (e.g. 10/40/100 Gb/s) scenario will further affect the cost-efficiency of protection sharing in the optical layer. Another area for further study is the use of Software Defined Networking, a recently defined paradigm which seperates control and forwarding functionality of a switch (which is one step further than seperating control plane and data plane in the network), facilitated by the OpenFlow protocol. The control functionality is centralized in an OpenFlow controller which configures the switch's forwarding hardware using the OpenFlow protocol. The solutions from Chapter 4 can be applied to OpenFlow networks, and instead of GMPLS and PCE, OpenFlow can be used to set up the connections.

A

Failure Localization in Transparent Optical Networks

D. Staessens, K. Manousakis, D. Colle, U. Mahlab, M. Pickavet, E. Varvarigos and P. Demeester

published in Proceedings of the Second International Workshop on Reliable Network Design and Modeling, October 19-20 2010, Moscow, Russian Federation

Abstract *In this paper, we address failure localization from both a practical and a theoretical perspective. After summarizing the state-of-the-art of failure localization algorithms and monitoring techniques, an overview of the most prevalent failures in optical core networks is presented. We review the role of the Optical Supervisory Channel and how it reports problems to the management plane. We analyze different equipment, investigating where most failures occur and how these failures can be monitored. We conclude that in-band OSNR monitoring is the most important monitoring technique for failure localization purposes. We give a general probabilistic model for failure localization and assess its limitations using the mutual information metric. We give a simple example for computing this mutual information and show that is it a valid metric for evaluation of the failure localization problem. For practical applications, with imperfect monitoring equipment and countless possible failures, the mutual information may be prohibitingly low. Initial analysis of the problem shows that we need intense and accurate monitor-*

ing in order to increase the mutual information for the problem and to be able to localize failures accurately.

A.1 Introduction

Modern telecommunications networks need to be able to detect and locate failures and degrations as fast and as accurately as possible, in order to restore lost traffic and repair the failure. While protection and restoration mechanisms can cope with traffic loss without exact knowledge of the failure type and location, most of the time spent reparing failures is spent in finding the precise cause.

Failures can be detected using various monitoring devices. These vary from simple photodetectors (detecting loss of light or attenuation), over OSNR (Optical Signal to Noise Ratio) monitors to Bit Error Rate (BER) monitoring, which is automatically performed at the termination point of each lightpath. More advanced monitoring techniques can specifically detect residual Chromatic Dispersion and other impairments. The failure localization problem is stated as, given a number of alarms in the network, where is the failures causing these alarms.

General probabilistic models for localizing network failures have been examined in [1], [2] and [3]. In [1] the network element failures are modeled in a dependency graph, where each node (element) has an a-priori probability to fail by itself (primary failure). When a node fails it will emit an alarm. A directed edge $e_i \rightarrow e_j$ indicated that element e_i depends on e_j and has a probability $P(e_i|e_j)$ to malfunction (and emit an alarm) due to the failure of e_j. The probabilities are assumed known and based on empirical and historical knowledge. [1] assumes that alarms only carry information about the emitting node, while [2] makes use of Alarm Reporting Functions in order to create classes of objects and [3] defines a hierarchical dependency graph consisting of services, protocols and functions and defining multiple failure modes per element. Both [2] and [3] transform their extended and hierarchical dependency graph into a simpler flat causality graph, mapping the extra information from the alarm messages into this graph. Note that we can consider the causality graph as a dependency graph. Also, each node has a single failure mode i.e. elements can only fail in one way (due to the primary fault) and emit only a specific alarm message (due to secondary malfunction). [1] examines the Maximum Mutual Dependency algorithm. The complexity is estimated to $O(N^3)$. [2] proposes an alarm domain extraction algorithm and [3] examines two algorithms, a combinatorial that uses a metric of goodness and an iterative heuristic (entitled Incremental Hypothesis Update) that uses a belief metric. The first one has $O(2^N)$ complexity which practically may be polynomial, while the complexity of the second one is $O(N^4)$.

A probabilistic approach is examined under a real environment [4]. They extract a hierarchical causality graph of tree topology and perform the reasoning by unfolding the hierarchy and just keeping the most probable problem.

In [5] authors propose an algorithm for locating multiple failures at the physical layer of a WDM network. Given the set of triggered alarms for each failure in the network, and a set of triggered alarms (may include false/missing alarms), find all possible failures which are capable of producing these alarms. The proposed algorithm does not rely on timestamps nor on failure probabilities as in [1].

Different techniques for distributed monitoring are described in literature. [6] showed the feasibility of a fault detection scheme for all-optical networks based on their decomposition into monitoring-cycles (m-cycles). In [7] authors formulate an m-cycle construction for fault detection as a cycle cover problem with certain constraints. They propose a heuristic spanning-tree based cycle construction algorithm that they apply to four typical networks. To detect and locate network faults, it is not necessary to put monitors on all links, lightpaths, or nodes. For example, some authors proposed a diagnosis method with sparse monitoring nodes (multiple monitors may be required) particularly for crosstalk attacks, which could be considered as special cases of network faults in a wide sense [8].

In [9] authors investigate the m-trail design problem. They conduct a bound analysis on the minimum length of alarm code required for unambiguous failure localization. Then, a novel algorithm based on random code assignment (RCA) and random code swapping (RCS) is developed for solving the m-trail design problem. The algorithm was verified by comparing with an Integer Linear Program (ILP), and the results demonstrated its superiority in minimizing the fault management cost and bandwidth consumption while achieving significant reduction in computation time.

Authors in [10] provide quantitative performance analysis for flat and hierarchically distributed monitoring and fault-localization in all-optical networks. They present an efficient heuristic and compare achievable improvements in monitor activation and fault-localization complexity for both schemes. A centralized, flat monitoring model consists of a central fault-manager which receives alarms from all monitors in the network and processes them simultaneously. Using such a model for monitoring large Transparent Optical Networks (TONs) can result in flooding the central manager with a large number of redundant alarms, delaying fault localization and service restoration.

In [11] authors propose the fault localization method using integrated network alarm correlation technique based on Consolidated Inventory Database (CID) which stores the network equipments details and the connection data among them. The proposed method collects the network alarms from various NMSes(Network Management System) which manages its own network domain. Authors claim that the analysis of alarm correlation based on the detailed end-to-end network view point is necessary to improve the effect of fault localization technique on complicated telecommunication networks. They propose fault localization method which covers complex networks, e.g., SDH networks, IP backbone networks, IP access domain networks, xDSL networks and etc.

This paper is further organized as follows. In Section A.2 we give a general overview of the network resources for failure management and an generic classication of failure types. In Section A.3 we summarize the most typical failures occuring in optical networks. In Section A.4 we give the probabilistic description of the failure localization problem. Section A.5 provides an example to evaluate the model and Section A.6 provides some directions for future work and concludes the paper.

A.2 Issues in network failure localization

Most networks use an Optical Supervisory Channel (OSC) for for remote node management, monitoring and control [12]. This OSC is typically a low bandwidth (STM-1) out-of-band (usually at 1510 nm), full duplex point-to-point communication and control channel. It is common practice to use the Digital Communication Channel (DCC) section of the STM-1 header or the General Communication Channel (GCC) of OTN for this purpose. In every managed node (e.g. amplifier, regenerator, cross connect) the channel is dropped, the relevant data is inspected, instructions are performed and possible replies are added. This reframing typically takes $100 - 200\mu s$.

There are many types of service disruptions in optical networks, which we can classify in two major types. On the one hand, we have *hard failures*, such as fiber cuts and failure of a network line card. Fiber cuts happen all too frequently, due to human error such as construction workers breaking a cable or due to natural causes, such as earthquakes. Line card failures can for instance happen due to short circuiting. These failures occur suddenly and have a severe impact on services, causing major loss of traffic. On the other hand, we have *soft failures* such as end-of-life of an amplifier. These are more subtle changes in performance, causing a wide spectrum of service degradations which are far more difficult to detect and locate.

We can differentiate between failures that are self-reported through the management systems, and those that are not. If some malfunctioning can be detected in a cost-efficient way, the equipment itself will implement a self-diagnostics subsystem and report these types of failures immediately. For other failures, such as noise increases, the detection requires OSNR monitoring, which is very expensive. These kind of degradations will usually not be self-reported.

Most hard failures (causing sudden loss of transmission) are self-reported, while only some soft failures are. Soft failures that are not self-reported may be very hard to detect and nearly impossible to accurately locate. We will now give an overview of the most prevalent network element failures and their consequences.

Equipment type	Failure mode	degradation	Monitor types	self-reported
VOA/DGE/DTE	unknown	wrong attenuation		yes
V-mux	total	attenuation max		yes
	total	attenuation fixed		yes
tunable filter+	drift of passband	noise due to XT (channel)	OSNR	no
tunable DCF	narrowing	distortion	(OSNR), BER	no
		wrong DCF length (ISI)	OSNR	no
switch, WSS	subsytem failure	noise due to XT	OSNR	no
	narrowing	FC (@ provisioning)		no
tx	end of life	drift	OSNR	no
	wrong power			yes
rx	complete	channel lost		yes
	electrical unit failure	noise		yes
fiber	bending	attenuation		no
	bad connector	attenuation/LOL		no
amplifier	low/high output			yes
	gain			yes
	gain tilt			no
	pump	noise (all channels)	OSNR / OSA	no

Table A.1: Failure modes in optical networks

A.3 Failures modes in optical networks

This section was compiled from IEC equipment specifications [13] and discussions with (sub)system design engineers.

A.3.0.1 Optical fibers and connectors

The most common failure in an optical network is a fiber cut. Fiber cuts are self-reported, because they generate loss of light, which is easily detected at neighboring managed sites and then reported to the management plane using the OSC. A lot more difficult to locate are fiber bending (macrobending) and lossy connectors due to dust or burning. Connector burning is commonly observed in high-power systems, for instance at a Raman pump laser, but could also occur due to amplifier transient effects. Usually transient effects are managed within the amplifiers, but after significant channel drop in a transparent network (for instance due to a fiber cut on a neighboring link) or malfunction of the transient management subsystem, it is possible for a transient to increase the power on a channel to disruptive levels.

Fiber bending and bad connectors cause loss over a wide spectrum, ranging from 0 to 20 dB. High loss will be self-reported like a fiber cut, but low loss due to minor bending or a little dust can be within design limits. This loss will lead to decreased OSNR. At an amplifier site, a lower power input signal is compensated by higher gain toward the output port, so that the net effect is a decreased OSNR of the output signal, which deteriorates with every subsequent amplification. The number of affected channels is dependent on the location of the bad connector or fiber bend. If the loss occurs before multiplexing, it will affect only a single channel. If it occurs after multiplexing, it will deteriorate all channels on the fiber. Depending on this location, localization techniques using out-of-band monitoring are not able to detect this failure.

A.3.0.2 Amplifiers

Amplifiers can cause different types of signal degradation. If an amplifier cannot reach its target output power due to malfunction of the gain control or power loss of the pump laser, this is usually detected by a photodiode and reported to the management system. Similarly, if the output power is too high this will be reported. However, variations of pump laser wavelength due to the aging or due to malfunctions of the temperature control system can increase optical noise. Most amplifier failures usually affect all channels, but if there is tilt in the amplifier gain, channels with higher amplification will show increased noise. This may make these failures difficult to detect using out-of-band monitoring techniques.

A.3.0.3 Variable optical attenuators

Another type of equipment that is widely used are Variable Optical Attenuators (VOAs). These are commonly used in arrays for Dynamic Gain Equalisation (DGE) in OXCs and multiplexers and tilt compensation in amplifiers. The effect of malfunction of these components is usually a change in power (when the VOA gets stuck in maximum gain or no gain) or a loss of control if the VOA gets stuck on its current gain level. The last failure of the VOA will not lead to an immediate signal degradation. All these failures are easily detected using photodiodes and therefore can be considered self-reported.

A.3.0.4 Tunable filters

The use of tunable filters in the network can also lead to OSNR degradations and increased BER. An application for filters is Dispersion Compensation in systems with multiple datarates, where higher datarates require more compensation. The filter will select the higher rate wavelength for transmission through additional DCF to compensate for residual chromatic dispersion. Narrowing of the passband can create signal distortion which will lead to BER increase, but may not be detected by OSNR measurement. Another problem is Filter Concatenation (FC), however this is only encountered at channel provisioning and is not a network failure in the strict sense. Drift of the filter passband may create noise due to crosstalk (XT), and if the DCF is of the wrong length, we may get Inter-Symbol Interference, which will again lead to increased OSNR.

A.3.0.5 Optical cross-connects

Optical Cross Connects exhibit similar problems as tunable filters since they use similar technology (e.g. MEMS). Attenuation problems, for instance due to misalignment of MEMS, are typically self-reported, but limited loss and XT leading to decreased OSNR are much more difficult to detect. Depending on the switch design, these failures affect single channels or all channels passing through it.

A.3.0.6 Transmitters and receivers

Most failures of transmitters and receivers are also easily detected. Wrong output power is self-reported since the transmitter usually uses a feedback loop to control its output power. If it cannot reach the correct output power, the unit sends an alarm through the OSC. A transmitter which reaches end-of-life and starts drifting will lead to misalignment with various filters, with distortion and possible OSNR decrease as a result. This failure is hard to locate. Receiver failures (destroyed receivers or electrical failures) are also self-reported.

A summary of these failure modes is given in Table A.1. From this section, we can conclude that the most important quantity to monitor in optical networks is noise, more specifically (in-band) OSNR. These monitors have a certain margin of error and are quite expensive. These factors make practical failure localization a difficult problem.

A.4 General problem statement

A.4.1 Definition

A network consists of a set of elements $E = \{e_1, \ldots, e_n\}$, which can fail with a certain probability $P_E(e_i) \in [0, 1]$. We define a network failure f_j as a set of element failures, so the set of network failures $F = \{f_1, \ldots, f_{2^n}\}$ is the power set of E. F includes the non-failure case. The probability of a network failure $P_F(f_i)$ can, in theory, be computed from the element failure probabilities and the dependency between these failures. Each network element failure can trigger alarms through different monitors. Call the set of alarms $A = \{a_1, \ldots, a_m\}$. An observation o_i is a set of alarms that are raised due to some network failure f_i with probability $P_{O|F}(o_j|f_i)$. The set of observations O is the power set of the set of alarms and has 2^m elements. The problem is to find the most likely network failure $f_x \in F$ which explains the observation $o_y \in O$, $f_x = \max_z \big(P_{O|F}(o_y|f_z)P_F(f_z)\big)$.

This general model describes the general problem of network failure localization. Every derived approach (i.e. a failure localization algorithm) will approximate the solution of this problem. The accuracy of the model will depend on the quality of the initial probabilities and the amount of information that is contained in the alarms. We will now assess the efficiency of the approach using the mutual information [14] metric. This metric gives a quantitative measure how sure we can be, given observation o_i, that network failure f_j is indeed the cause.

The above problem description is NP-complete and therefore computationally infeasible for large networks. In a network with n elements and m alarms, the number of probabilities is $2^n * 2^m = 2^{n+m}$.

A.4.2 Mutual information, self-information and entropy

Let $x_1, \ldots, x_k$ be the X sample space and $y_1, \ldots, y_l$ be the Y sample space in an XY joint ensemble. We want a quantitative measure of how much the occurence of y_j in the Y ensemble tells us about the occurence of the possibility x_i in the X ensemble. The occurence of $y = y_j$ changes the probability of $x = x_i$ from the *a priori* probability $P_X(x_i)$ to the *a posteriori* probability $P_{X|Y}(x_i|y_j)$. This measure is called the mutual information between y_j and x_i and is defined as

$$I_{X;Y}(x_i; y_j) = I_{Y;X}(y_j; x_i) = \log \frac{P_{X|Y}(x_i|y_j)}{P_X(x_i)} \tag{A.1}$$

The term *mutual* information comes from the symmetry of equation (A.1). The (weighted) average mutual information between X and Y is defined as:

$$I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{XY}(x_i, y_j) \log \frac{P_{X|Y}(x_i|y_j)}{P_X(x_i)} \tag{A.2}$$

where $P_{XY}(x_i, y_j) = P_X(x_i)P_{Y|X}(y_i|x_i) = P_Y(y_i)P_{X|Y}(x_i|y_i)$ is the probability of observing $X = x_i$ and $Y = y_i$ simultaneously. If an event x_i is fully specified by the occurence of y_j, i.e. $P_{X|Y}(x_i|y_j) = 1$ the mutual information between x_i and y_j becomes:

$$\begin{aligned} I_{X;Y}(x_i; y_j) &= \log \frac{P_{X|Y}(x_i|y_j)}{P_X(x_i)} \\ &= \log \frac{1}{P_X(x_i)} = I_X(x_i) \end{aligned} \tag{A.3}$$

and we call this the self-information of the event $x = x_i$. The entropy of an ensemble X is the (weighted) average self-information of the ensemble and is given by:

$$\begin{aligned} H_X(X) &= \sum_{i=1}^{n} P_X(x_i) \log \frac{1}{P_X(x_i)} \\ &= -\sum_{i=1}^{n} P_X(x_i) \log P_X(x_i) \end{aligned} \tag{A.4}$$

A.5 Efficiency of the probabilistic model

The efficiency of any failure localization in an optical network will strictly depend on the mutual information between monitors and failures. In the ideal case, self-reported failures have mutual information equal to the self-information, meaning that the probability of the reported failure, when we receive the alarm indicating this failure, is 100%. Of course, implementing perfect monitoring for every conceivable set of failures in the network is impossible.

From a theoretical viewpoint, all probabilities are considered as input for the model. Of course, from a practical perspective, this is where the real difficulties are encountered. The a priori failure probabilities for the equipment can be more or less estimated from experience [1], but the conditional probabilities for the alarms are far less straightforward to compute. Most models [1] [5] take these to be 1, i.e. if the equipment fails, the alarm(s) will be raised and vice versa. However, for real networks this is not the case, as we illustrated above.

Even small changes in these probabilities have a huge impact on the mutual information. This is intuitively understood by considering the following example. If you monitor some equipment with a failure probability of 10^{-4}, with 100% accuracy, when your alarm is raised you are 100% sure that this failure occured. However, if you monitor the same element with 99.99% accuracy, when you have an alarm, you have roughly 50% chance that the element failed and 50% chance it's a false alarm, since both events are equally likely. It are these a posteriori probabilities that are summed to compute the mutual information in Eq. (A.2).

A.5.1 Example

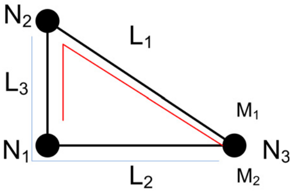


Figure A.1: Example for localization of link degrations in a triangle topology using 2 monitoring points.

In Figure A.1, we give a small example for a simple 3 node ring network with 2 monitors at the end of 2-hop paths. We make the following simplifying assumptions. First, the failures are statistically independent, second the monitors work perfectly. We only consider 3 possible failures associated with the three links. Call the three nodes N_1, N_2, N_3 with two monitors M_1 and M_2 located in node N_3. The links L_1, L_2, L_3 have length $5, 4, 3$ respectively and there are two lightpaths, one from N_1 to N_3 along $L_3 - L_1$, monitored by M_1 and from N_2 to N_3 along $L_3 - L_2$, monitored by M_2. Failure of L_1 triggers M_1, failure of L_2 triggers M_2 and failure of L_3 triggers both monitors. All multiple link failures will trigger M_1 and M_2, meaning we cannot fully distinguish between multiple link failures and the single failure L_3. We assume the probability of a failure per unit length to be 10^{-4}.

We can thus construct the following sets:

$$E = \{L_1, L_2, L_3\} \tag{A.5}$$

$$F = \{\emptyset, L_1, L_2, L_3, L_1L_2, L_1L_3, L_2L_3, L_1L_2L_3\} \tag{A.6}$$

$$A = \{M_1, M_2\} \tag{A.7}$$

$$O = \{\emptyset, M_1, M_2, M_1M_2\} \tag{A.8}$$

failure f	$P(f)$	$\log \frac{1}{P(f)}$	$P(f) \log \frac{1}{P(f)}$
$\emptyset$	0.9988005	0.0017316	0.001729
L_1	0.0004996	10.966794	0.005480
L_2	0.0003997	11.288867	0.004512
L_3	0.0002997	11.704049	0.003508
L_1L_2	$1.199\ 10^{-07}$	22.991183	$2.758\ 10^{-06}$
L_1L_3	$1.499\ 10^{-07}$	22.669111	$3.400\ 10^{-06}$
L_2L_3	$1.999\ 10^{-07}$	22.253930	$4.449\ 10^{-06}$
$L_1L_2L_3$	$6\ 10^{-11}$	33.956247	$2.037\ 10^{-09}$
	1		0.015239663

Table A.2: Entropy of the failures

The a priori probabilities are given in Table A.2, together with the quantities to compute the entropy. The table immediately confirms what was to be expected: the highest information content lies in the single failures and the absense of failures. In order to compute the mutual information, we need the a priori probability of a monitor triggering, i.e. the a priori probabilities of the observations. These are easily computed to be $P_O(\emptyset) = 0.99880047$, $P_O(M_1) = 0.00049965$, $P_O(M_2) = 0.00039968$ and $P_O(M_1M_2) = 0.0003002$.

Calculation of the average mutual information is given in Table A.3. The first column shows the mutual information between each failure and the observation. Since the conditional probability equals 1, this is equal to the self-information in the monitors (see Eq. (A.1). Note that the probabilities $P_{FO}(fo)$, needed for computation of the mutual information between the two sets, are completely dependent on the failures, so $P_{FO}(fo) = P_F(f)$ if we assume perfect monitoring.

We see, that in this simple example, the mutual information is lower than the entropy of the failures, meaning we cannot distinguish all failures. Should we

f	o	$I(f;o)$	$P_{FO}(fo).I(f;o)$
$\emptyset$	$\emptyset$	0.001731595	0.001729518
L_1	M_1	10.96679435	0.005479559
L_2	M_2	11.28886678	0.004511935
L_3	M_1M_2	11.70178869	0.003507378
L_1L_2	M_1M_2	11.70178869	$2.33966\ 10^{-06}$
L_1L_2	M_1M_2	11.70178869	$1.75457\ 10^{-06}$
L_2L_3	M_1M_2	11.70178869	$1.40351\ 10^{-06}$
$L_1L_2L_3$	M_1M_2	11.70178869	$7.02107\ 10^{-10}$
			0.015233882

Table A.3: Mutual information

have placed three monitors (with perfect accuracy), there would of course be no ambiguity. An algorithm which focusses on single failures would in this case perform almost as well as the complete solution.

In real networks we need to take caution with this example. First, failures occur with a large distribution of failure probabilities. Some dual failures may be more common than other single failures. Second, monitoring is not perfect, and we cannot choose to omit certain failure types. If we assume imperfect monitoring, say with inaccuray of 10^{-8}, we can compute the entropy for this scenario to be 0.015239663. For the mutual information, we find the value 0.015233699, or a ratio of 0.9996. This is an example where we can distinguish all single failures, and double failures are much less likely and contribute little to the total mutual information. An algorithm focussing on single failures has mutual information 0.015228207 (this is easily computed by omitting the contributions of multiple failures to the mutual information) or ratio 0.99925, and can be considered a good algrrithm.

If we take the same example (with inaccuracy 10^{-8}, but the second lightpath is from N_1 to N_3 along L_2, then M_2 triggers only when L_2 fails and failure of L_3 cannot be distinguished. In this case, we can compute the mutual information to be 0.014474206 or a ratio of 0.94977. This may seem like a good value, but we know that this example cannot locate all single failures, so this value is already an indication of inadequate monitoring.

In Figure A.2 we plot the mutual information versus the monitor accuracy. The monitor accuracy is shown as the logarithm of the accuracy (-4 meaning 0.9999 accuracy). This figure clearly shows a sudden drop in mutual information around 10^{-4}, exactly the range of the failure probabilities of the elements. This shows that mutual information is also a good indicator for the monitoring accuracy, and inversely shows that accurate monitoring is paramount in failure localization.

We find already in these simple examples that the mutual information is lower

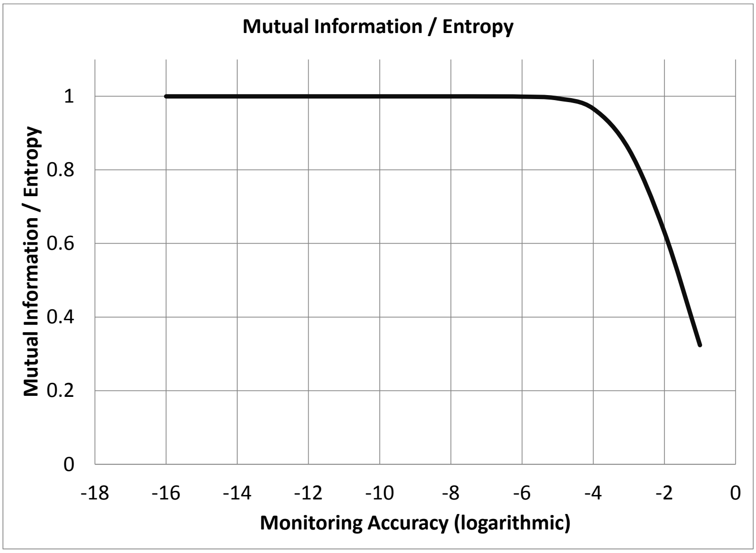


Figure A.2: Mutual information vs monitor accuracy

than the entropy of the failures, meaning we cannot distinguish all failures. Should we have placed three monitors (with perfect accuracy), there would of course be no ambiguity. In real networks failures occur with a large distribution of failure probabilities. Some dual failures may be more common than other single failures.

A.6 Conclusions and future work

We have summarized different possible failures in optical networks and how they can be monitored. From this summary, OSNR monitoring proves to be the most important form of monitoring to install in the network. We show that the mutual information between the monitors (i.e observations) and failures is a good metric for failure localization efficiency, both in the case of insufficient monitoring and inaccurate monitoring. In the ideal case, the mutual information between the monitors and the failures should equal the entropy of the failures.

In ongoing work, we will investigate the sensitivity of the mutual information to the number of monitors, their location and the accuracy of the monitoring. We will try to find exact boundaries for the mutual information where monitoring is accurate enough to locate all major failures. We will use this model for locating the optimum placement of monitors.

Acknowledgment

This work was supported by the European Commission within the BONE-project ('Building the Future Optical Network in Europe'), a Network of Excellence, and Strep DICONET ('Dynamic Impairment Constraint Networking for Transparent Mesh Optical Networks'), both funded through the 7th ICT-Framework Programme.

References

[1] I. Katzela and M. Schwartz. *Schemes for Fault Identification in Communication Networks.* IEEE/ACM Transactions on Networking, Vol. 3 (6), pp. 753-764, 1995.

[2] M. Choi J. Choi and S.H. Lee. *An Alarm Correlation and Fault Identification Scheme Based on OSI Managed Object Classes.* In Proc. IEEE Int'l Conference on Communications ICC '99, pp.1547-1551, 1999.

[3] M. Steinder and A.S. Sethi. *Non-Deterministic Diqgnosis of End-to-End Service Failures in a Multi-Layer Communication System.* In Proc. IEEE Int'l Conference on Computer Communications and Networks '01, pp. 374-379, 2001.

[4] D.L. Yang C.S. Chao and A.C. Liu. *An automated Fault Diagnosis System Using Hierarchical Reasoning and Alarm Correlation.* Journal of Network and Systems Management,Vol.9 (2), pp.183-202, 2001.

[5] C. Mas and P. Thiran. *An efficient algorithm for locating soft and hard failures in WDM networks.* IEEE Journal On Selected Areas In Communications, Vol. 18 (10), Oct. 2000.

[6] T. Wu and A.K. Somani. *Necessary and Sufficient Condition for k Crosstalk attacks localization in All-optical Networks.* In Proc. IEEE Globecom, pp. 2541-2546, 2003.

[7] C. Huang H. Zeng and A. Vukovic. *A novel fault detection and localization scheme for mesh all-optical networks based on monitoring-cycles.* Photonic Network Communication, Vol. 11, pp.277-286, 2006.

[8] T. Wu and A.K. Somani. *Attack monitoring and localization in all optical networks.* In Proc. of SPIE Opticommun 2002: Optical Networking and Communications, vol. 4874, pp. 235-248, 2002.

[9] J. Tapolcai et al. *On monitoring and failure localization in mesh all-optical networks*. In Proc. IEEE INFOCOM '09, 2009.

[10] S. Stanic and S. Subramaniam. *A Comparison of Flat and Hierarchical Fault-Localization in Transparent Optical Networks*. In Proc. Optical Fiber Communications, 2008.

[11] et al. J. Kim. *Fault Localization for Heterogeneous Networks Using Alarm Correlation on Consolidated Inventory Database*. In Springer LNCS , pp. 82-91, 2008.

[12] R. Ramaswami and K. N. Sivarajan. *Optical Networks, A Practical Perspective, 2nd ed.* Elsevier, 2001.

[13] The IEC. *IEC86C, Document 62343-6-6, DYNAMIC MODULES, Failure Mode Effect Analysis for Optical Units of Dynamic Modules*.

[14] R.G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

B

Benefits of Implementing a Dynamic Impairment Aware Optical Network: Results of EU project DICONET

M. Angelou, S. Azodolmolky, I. Tomkos, J. Perell, S. Spadaro, D. Careglio, K. Manousakis, P. Kokkinos, E. Varvarigos, D. Staessens, D. Colle, C. V. Saradhi, M. Gagnaire, Y. Ye

Published in IEEE Communications Magazine, Vol. 50 (8), August 2012, pp. 79-88

Abstract *Dynamic optical networking allows operators to effectively maximize the capacity of their physical infrastructure and cope with the rapid growth rates of the Internet traffic. In the framework of the European DICONET project we proposed and developed a comprehensive solution that utilizes the dynamicity as well as the valuable physical layer information of a reconfigurable WDM core network to provide a smooth transition from the quasi-static networking of today to an intelligent reconfigurable and physical impairment-aware architecture. In this work we discuss the benefits of implementing the DICONET solution and present some of the major achievements of the project that support both the planning and operation phase of a core optical network.*

B.1 Introduction

In old voice-centric telecom networks, planning and operation phases little resembled the equivalent functions of today's data-centric telecom networks. Nowadays the indisputable growth of data traffic with dynamic usage patterns generated by novel capacity-demanding applications drives the developments in communications worldwide. Core optical networks, occupying a fundamental piece in the Internet puzzle, evolved to networks that span over thousands of kilometres of fiber, carry high-capacity traffic and switch connections all-optically. The evolution trend followed a path towards higher spectral efficiency and lower total cost of ownership (TCO), facilitated first by the emergence of Wavelength Division Multiplexing (WDM) and Optical Add Drop Multiplexers (OADM) and later by the reconfigurable all-optical nodes. The evolution path moved from an optical network with optical-electronic-optical regeneration at every node (opaque) to a translucent network where the regeneration takes place only in a small number of sites or to a transparent network where the regenerators are totally omitted. This transition was driven and justified not only by the high-bandwidth provision but also by the minimized TCO stemming mostly from the elimination of the costly opto-electronic interfaces. In this context though, new and more complicated implications were introduced in order to commercially realize an optical network that is fully-dynamic, robust and cost-effective.

In old voice-centric telecom networks, planning and operation phases little resembled the equivalent functions of todays data-centric telecom networks. Nowadays, the indisputable growth of data traffic with dynamic usage patterns generated by novel capacity-demanding applications drives developments in communications worldwide. Core optical networks, occupying a fundamental piece of the Internet puzzle, evolved to networks that span over thousands of kilometers of fiber, carry high-capacity traffic, and switch connections all-optically. The evolution trend has followed a path toward higher spectral efficiency and lower total cost of ownership (TCO), facilitated first by the emergence of wavelengthdivision multiplexing (WDM) and optical adddrop multiplexers (OADMs) and later by reconfigurable all-optical nodes. The evolution path moved from an optical network with optical-electronic-optical regeneration at every node (opaque) to a translucent network where the regeneration takes place only in a small number of sites, or to a transparent network where the regenerators are totally omitted. This transition was driven and justified not only by high-bandwidth provision but also by the minimized TCO stemming mostly from elimination of the costly opto-electronic interfaces. In this context, though, new and more complicated implications were introduced in order to commercially realize an optical network that is fully-dynamic, robust and cost-effective. Optical transparency has a significant impact on network design and operation; the introduction of physical-layer considerations is mandatory in order to cope with the physical-layer effects that deteriorate the connections quality of transmission (QoT). These challenges can be overcome by introducing additional rules for WDM systems, performance monitoring, and control plane-driven reconfiguration capabilities.

The European research project Dynamic Impairment Constraint Networking for Transparent Mesh Optical Networks (DICONET) successfully addressed these challenges. The consortium of the DICONET project, which consisted of seven academic partners, four equipment manufacturers, and one telecom operator, envisioned a comprehensive multilevel solution based on novel cross-layer algorithms for core optical networks [1]. The key concept of the solution makes use of the accumulation of the physical-layer effects that degrade the quality of the optical signal. Linear and nonlinear impairments that either affect every WDM channel individually or cause interference to neighboring channels render the optical reach finite. Current long-haul WDM networks tackle the limitations induced by the optical medium with careful link design, dispersion management, and power budgets. Beyond these offline techniques, DICONET actually exploits the knowledge of the impact of the single-channel and multichannel effects to introduce an intelligent and high-performance network with impairment awareness in the planning and operational processes. In addition, DICONET applies to dynamic optical networks and utilizes the reconfiguration ability of the optical switching components (reconfigurable optical crossconnects, R-OXCs) to support highly dynamic traffic in a flexible and economic manner [2]. As opposed to the current networks that employ OXCs with fixed configuration, reconfigurable optical nodes offer a clear advantage to the operators as they do not have to overprovision their network with costly equipment meant to serve future traffic variations. Hence, the network can react on the fly to traffic changes or failures, without the need for on-site interventions. The goal of this article is to highlight the key features of the integrated DICONET solution via some of the achievements of the project. In the core of DICONET resides a set of crosslayer optimization algorithms designed to serve the network both during planning and operation. These algorithms are integrated in a common software platform, the DICONET Network Planning and Operation Tool (NPOT) [3], that considers the impact of physical-layer impairments (PLIs) on the decision making. The control plane effectively supports the DICONET networking solution through developed generalized multiprotocol label switching (GMPLS) protocol extensions, allowing the different entities to cooperate and run in an orchestrated manner. The project was completed with the implementation of the multilevel integrated solution in the DICONET testbed, practically realizing the vision for high end-to-end connectivity, dynamicity, and reliability.

The article is organized as follows. The following section gives an overview of the project by discussing the benefits of the DICONET solution. The remainder is dedicated to the main achievements that essentially enable the features of the solution, in turn including the NPOT, the developed cross-layer optimization modules, the control-plane related developments, the testbed implementation, and the achieved capital resource optimization. The article concludes with some remarks about future research challenges.

B.2 Benefits

The DICONET project proposes a comprehensive solution that tackles issues in the planning or offline phase as well as the operation or online phase of an optical core network. The planning phase includes processes that are directly linked to the network capital. Transponders, monitors, and regenerators are costly equipment, whose associated capital and operational cost justifies the need for resource optimization. In the DICONET approach this challenge is addressed by dedicated resource optimization modules that minimize and efficiently allocate the available resources (i.e., monitor and regenerator placement) offline: impairment-aware routing and wavelength assignment IA-RWA), exploiting the maximum transparent optical reach.

During operation, employing the dynamic and impairment-aware solution implies a situation where the network is fully aware of the physical status of its components and the QoT of the established connections. Optimum decision making is therefore achieved also during operation, utilizing intelligent online IA-RWA algorithms that serve the dynamic traffic. In addition, dynamicity and high-performance end-to-end connectivity strengthen the online operation as the light-path provisioning is achieved in low setup times. Apart from the fast connection establishment the DICONET solution is capable of rapidly localizing potential physical failures and restoring the affected traffic, rendering the network intelligent and robust. Besides, operators always seek for Quality of Service (QoS) as it is an important revenue-generating attribute.

As a whole, DICONET is designed to work independent of the scale of the network topology as its tools are applicable to both core networks where regeneration of the optical signal is not necessary (transparent) and networks of bigger scale, where some strategically selected regeneration sites are required (translucent). Furthermore recognizing the importance of the use of standardized protocols, Generalized Multi-Protocol Label Switching (GMPLS) is adopted in the DICONET approach to control the transport plane. Indeed the control plane entities that run the optical transport employ the full GMPLS protocol suite, yet properly enhanced to support the PLIs. In what follows the main building blocks that induce these benefits are presented.

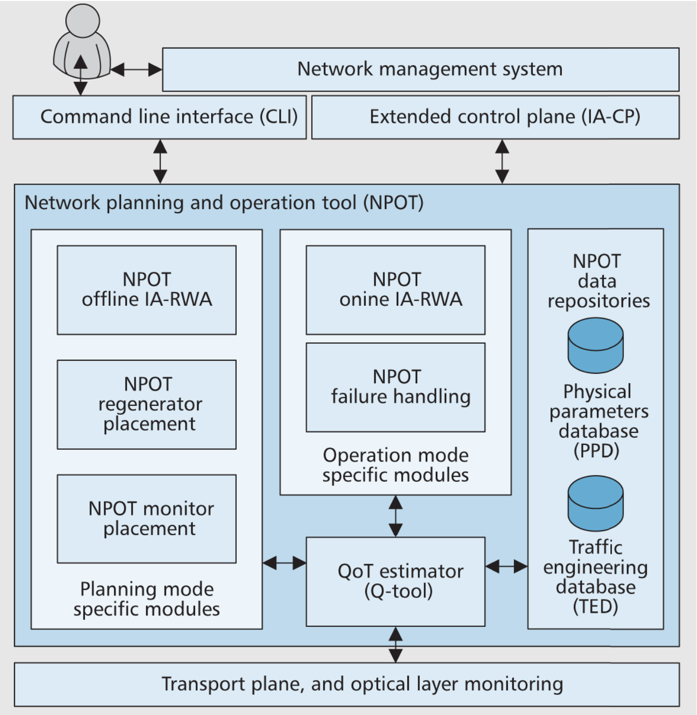

Network management system
Command line interface (CLI)
Extended control plane (IA-CP)
Network planning and operation tool (NPOT)
NPOT offline IA-RWA
NPOT regenerator placement
NPOT monitor placement
Planning mode specific modules
NPOT onine IA-RWA
NPOT failure handling
Operation mode specific modules
QoT estimator (Q-tool)
NPOT data repositories
Physical parameters database (PPD)
Traffic engineering database (TED)
Transport plane, and optical layer monitoring

Figure B.1: Anatomy of DICONET network planning and operation tool

B.3 Network planning and operation tool

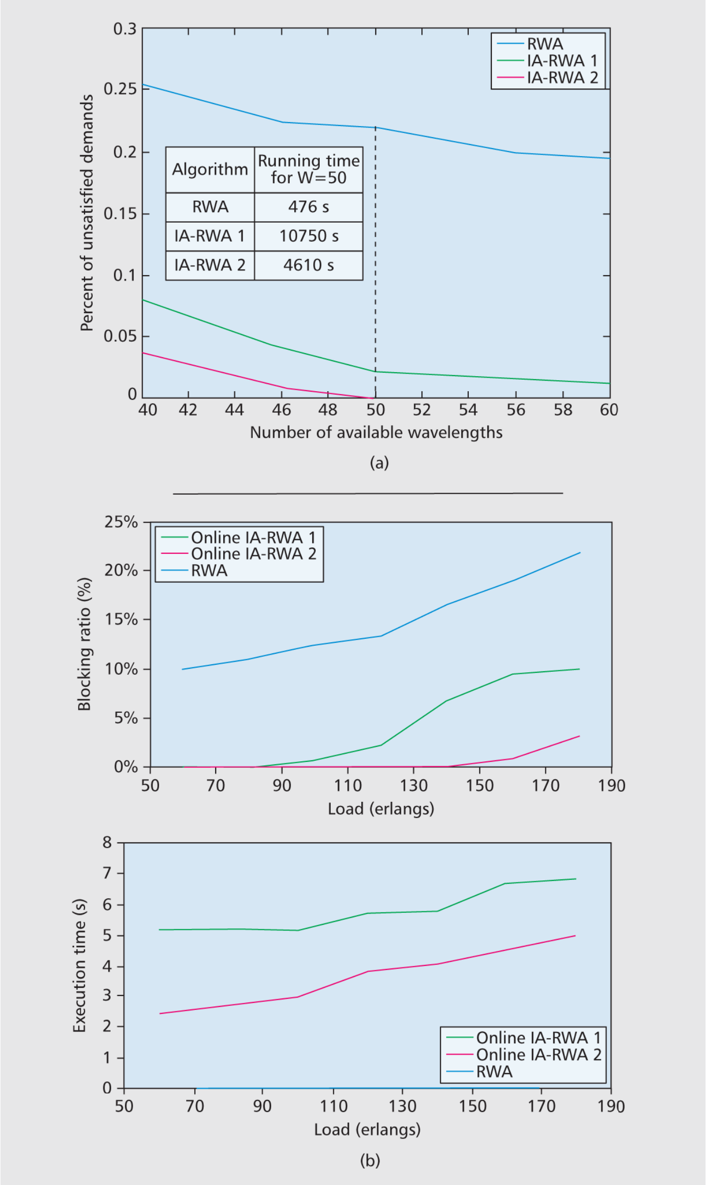
The key innovation of DICONET is the design and development of NPOT that integrates in a common platform cross-layer algorithms that make use of physical-layer assessments, serving the network during planning and operation citeSiamak2011. Following the development and testing of the various cross-layer techniques, the most suitable of each task was selected and all together were combined to act as the building blocks of NPOT. The most important of those are illustrated in the graphical representation in Fig. B.1. The planning mode of NPOT consists of the Optical Monitor Placement, the Regenerator Placement and the offline IA-RWA modules, supporting the network manager before the actual network operation. In the operation mode the tool includes the online IA-RWA and the Failure Localization modules. The global network information, including the physical layer, the topology and the traffic parameters, populate two data repositories that are kept as external databases (Physical Parameters Database (PPD) and Traffic Engineering Database (TED)). All the input data are introduced to the two databases in a simple XML format.

In the core of the NPOT is situated a QoT estimator. The various components of the tool consult the QoT estimator to make physical-layer aware decisions. The RWA process, whether online or offline, uses the QoT estimator either as a quality metric during the routing and wavelength assignment process or after the routing and wavelength assignment has taken place to evaluate and validate the computed solution. In turn the regenerator and monitor placement algorithms invoke the QoT estimator in order to find the optimum location for these components. The QoT estimator utilizes the updated information stored in the databases to estimate in a single figure-of-merit (i.e. Q-factor) the quality of the signal travelling on a lightpath. It is noteworthy that the modular design of NPOT allows any of its building blocks to be upgraded or replaced by other algorithms in a seamless way.

B.4 Cross-layer optimization modules

Network planning entails all the activities that are required to accommodate an initial traffic matrix with optimal resource allocation. The well-known problem of RWA constrained by the performance of the optical signal has received great attention from the research community [4]. IA-RWA refers to the process which given a set of demands, assigns a route and a wavelength to each of them always taking into account the physical impairments that degrade the QoT. During the planning phase, such an algorithm computes the routes and allocates the available resources in this case the optical channels- for a static traffic in order to find the optimal strategy to accommodate it. This offline operation takes place before a network actually starts to operate.

Throughout the project the consortium dedicated significant effort to develop and study offline IA-RWA algorithms for transparent and translucent networks [5], [6]. Considering PLIs in offline RWA has a certain particularity, as it involves the joint assignment of routes and channels to the connection requests, and interference among the selected lightpaths is inevitable once the solution has been found. Extensive comparative simulations for transparent topologies were performed using the various offline IA-RWA algorithms under realistic network and traffic parameters, exploiting a common QoT estimator tool developed within the project [3]. Our experiments showed the applicability of these algorithms to real-scale experiments, as they demonstrated good performance characteristics and implementation complexity, and relatively low execution times. Indicatively, we refer here to two of the developed impairment-aware algorithms and compare them with a standard k-shortest path RWA algorithm without physical-layer constraints in an effort to highlight the added value of impairment awareness. Figure B.2a includes the blocking ratio of the three offline RWA algorithms with respect to the number of available wavelengths. The two IARWA algorithms use linear programming techniques and account for the interference among lightpaths in their formulation. The first algorithm (IA-RWA 1) takes the physical layer indirectly into account by



Algorithm	Running time for W=50
RWA	476 s
IA-RWA 1	10750 s
IA-RWA 2	4610 s

(a)

(b)

Figure B.2: (a) Offline IA-RWA blocking ratio and running times as a function of the available wavelengths W; (b) online IA-RWA blocking ratio and average execution time per connection in seconds presented as a function of the network load, for a fixed number of wavelengths (W = 20).)

limiting the impairment-generating sources. The second algorithm (IA-RWA 2) uses noise-variance-related parameters to directly account for the most important physical impairments [7]. The experiment used as reference topology the generic national backbone network of Deutsche Telekom (DT) and assumed a realistic traffic matrix corresponding to the yearly traffic of 2009 (2.8 Tb/s). Each traffic demand is assigned a channel at 10 Gb/s. The simulations of each algorithm were executed on a PC with Intel Core2 Duo at 3.GHz and 4 Gbytes RAM. Evidently the plain RWA cannot compete with the IA-RWA methods and yields very high blocking for the entire range of available channels. In addition, offline RWA algorithms constrained by PLIs were also proposed for translucent networks, where regenerators are necessary in certain nodes to serve the traffic demands. Assuming a static traffic scenario with regenerators placed sparsely at certain a priori known locations, the solution consists of the routes and assigned wavelengths, and includes the decision of whether a connection will be served with or without the use of regenerators. In the former case also the sequence of regenerators is returned [6].

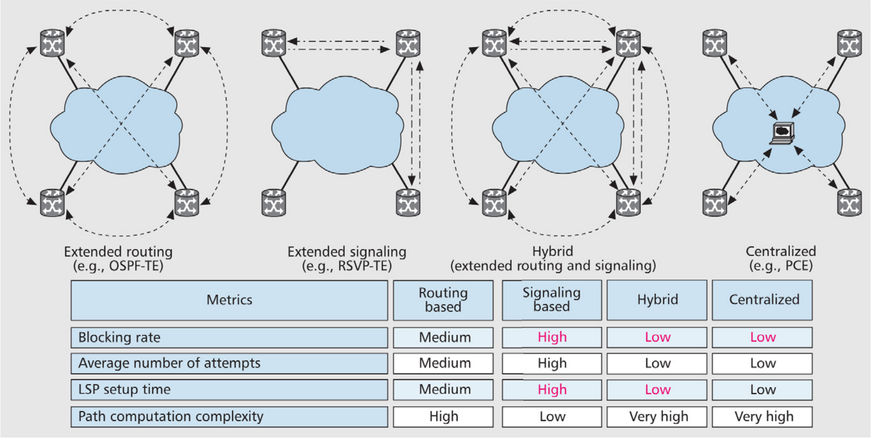
Before the RWA process, operators that employ the DICONET solution have additional cross-layer optimization tools at their disposal to support the planning phase. Regenerator and monitor placement refer to the modules developed to make optimized decisions on the number and location of the regenerating and monitoring equipment required in the network, by considering again the physical-layer performance [8]. This task has been specially focused on the regenerator placement techniques as those components imply significant capital and operational expenditures. Minimizing the particularly power-consuming opto-electronic interfaces of regenerators leads to the invaluable optimization of the total energy consumption of the network.

After the offline planning has been applied and the deployed network starts to operate, traffic demands may be requested or dropped in a dynamic fashion. Online IA-RWA algorithms specially designed for the operation phase process the new demands upon their arrival and one at a time, taking into consideration the current state of the network. Therefore, a new demand is served constrained by the traffic and physical layer characteristics present at the time of arrival. The objective here is to assign routes and wavelengths to these dynamic demands taking PLIs into account, so as to satisfy their QoT requirements without disrupting the QoT of the already established connections. The time needed for making a connection assignment decision should be short so that the connections establishment delay is also acceptably short. Similar to the offline case, we developed a number of online IA-RWA algorithms and performed simulation experiments to assess their performance under identical conditions and utilizing the same QoT estimator. Indicatively, Fig. B.2b illustrates the capabilities of two multicost algorithms against a simple shortest-path-based that does not consider the QoT. In multicost routing, a vector of cost parameters is assigned to each link, from which the cost vectors of the paths are calculated. The first algorithm (online IA-RWA 1) utilizes cost vectors consisting of impairmentgenerating source parameters, so as to be generic and applicable to different physical settings. These parameters are combined into a scalar cost that indirectly evaluates the quality of candidate lightpaths. The second algorithm (online IA-RWA 2) uses specific physical-layer models to define noise variance-related cost parameters, so as to directly calculate the Q-factor of candidate lightpaths [9]. The comparison scenario assumed the DT reference topology and a dynamic input traffic. Connection requests (each requiring bandwidth equal to 10 Gb/s) are generated according to a Poisson process with rate ($\lambda = 1$) (requests/time unit). The source and destination of a connection are uniformly chosen among the nodes of the network. The duration of a connection is given by an exponential ran-dom variable with average $1/\mu$ (time units). Thus, λ/μ gives the total network load in Erlangs. In each experiment 2000 connection requests are generated. Figure B.2b demonstrates two different performance metrics, blocking probability and execution time. Upon arrival of a traffic demand, fast response is essential together with accurate QoTaware routing decisions. Moreover, an effort was made to develop IA-RWA algorithms for translucent networks. We proposed algorithms that jointly address the route, lightpath, and regenerator selection problems, attempting to minimize the usage of the available regenerators [10].

Another important building block of the overall networking solution is responsible for monitoring the network for failures and locating the exact link that needs to be recovered. Upon a failure, following the fault localization process, the network utilizes its reaction mechanisms and restores all affected traffic. The online IA-RWA module takes over to compute new lightpaths for the connections that have been disrupted. The result is a robust and reliable core network with guaranteed QoS. In the framework of the project, localization techniques for failures that cause complete interruption of a connection (e.g., fiber cut) or merely QoT degradation were developed and studied [11]. These techniques are fed with monitoring data from supervising devices (e.g., bit error rate [BER], power, or optical signal-to-noise ratio [OSNR] monitors) spread throughout the network that feed the restoration mechanisms.

B.5 Control plane

Dynamic and impairment-aware networking relies heavily on a control plane enhanced with features that together with NPOT essentially enable the realization of this vision. Recently, the adoption of the GMPLS framework developed by the Internet Engineering Task Force (IETF) seems to prevail as the winning solution for the efficient control of an optical network. One of the main applications of GMPLS in the context of optical networks is the dynamic establishment and tear-



Metrics	Routing based	Signaling based	Hybrid	Centralized
Blocking rate	Medium	High	Low	Low
Average number of attempts	Medium	High	Low	Low
LSP setup time	Medium	High	Low	Low
Path computation complexity	High	Low	Very high	Very high

Figure B.3: Qualitative assessment of the various control plane options

down of lightpaths. DICONET utilizes the GMPLS protocol set but is not limited to its standard capabilities. One of the key tasks of the project concerned the extension of the GMPLS protocols to carry physical-layer information [3]. Whenever a change in the physical-layer status occurs, it needs to be communicated to all responsible entities that have to take actions. The availability of this up-to-date information is essential so as to accurately evaluate the effect of PLIs and decide on the feasibility of a lightpath in the optical domain.

Various control plane architectures were evaluated for implementation in the dynamic and impairment-aware network, including centralized and distributed solutions. The centralized approach implies a central point of control accessible by all network entities and aware of the complete network topology, resource availability, and physical-layer information. In the distributed case all network entities are involved in the control plane signaling and routing processes, but are deprived of global data knowledge. Three distributed architectures were considered: a signaling-based approach where the signaling component (i.e., Resource Reservation Protocol with Traffic Engineering extensions [RSVP-TE]) is extended to consider the PLI information, a routing-based approach where the routing component is extended (i.e., Open Shortest Path First with Traffic Engineering extensions [OSPF-TE]), and also a hybrid one that overcomes the limitations of the other two by extending both the routing and signaling protocols. The distributed approaches along with a centralized architecture that employs a path computation element-based (PCE-based) method underwent a qualitative comparison to explore the performance and applicability of the four options using performance and engineering metrics (Fig. B.3).

Two control plane schemes were eventually selected to implement and test for the purposes of the project. The two schemes differ with respect to the role of the NPOT in the overall multiplane architecture; one is hereafter referred to as the centralized (PCE-based) and the other as hybrid/distributed. In the former the NPOT is an engine common to all optical communication controllers (OCCs). The set of OCCs essentially realize the control plane, and each of them runs the full GMPLS protocol suite: RSVP-TE, OSPF-TE and Link Management Protocol (LMP). Apart from extending the OSPF-TE to disseminate the PLI information, the novelty of this approach lies in the PCE which in collaboration with the NPOT forms the so-called enhanced PCE (E-PCE) that deals with all the path computation and provision related actions. Path Computation Reply message of the standard Path Computation Element Protocol (PCEP) is extended and two novel messages, namely the Path Allocation Result and the Path Tear-down Result are defined to match the requirements of the PCE-based approach [12]. The standard RSVP-TE is deployed to establish, maintain, and tear down connections.

In the latter architecture all the network nodes run their own instance of the NPOT and extended versions of OSPF-TE and RSVP-TE. OSPF-TE is extended to disseminate wavelength availability information, while RSVP-TE carries the PLIs information for the QoT feasibility check. Due to the distributed nature of this implementation, upon receiving a new connection request from the network management system (NMS) the lightpath computation and the QoT estimation processes take place in the local NPOTs of the source and destination nodes. Prior to final integration and validation of the DICONET concept, emulation experiments were conducted to explore the capabilities of both selected control plane architectures under dynamic conditions and traffic load.

B.6 Implementation, testing and demonstration

Following the development of the different pieces, all were eventually integrated in a multiplane testbed spanning from the transport to the management plane. Both the centralized and distributed architectures were implemented in the 14-node experimental testbed bearing 1 or 14 NPOTs respectively. Each of the 14 OCCs consists of three different modules: the link resource manager (LRM), routing controller (RC), and connection controller (CC). Briefly, the LRM is a module responsible for the management of the resources available at the optical node through the Connection Controller Interface (CCI), while the RC and CC implement OSPF-TE and RSVP-TE, respectively. The transport plane of the DICONET testbed represents the same 14-node topology, and bears both emulated and physical optical nodes. Specifically, three 2-degree reconfigurable OADMs (ROADMs) based on wavelength selective switching (WSS) technology were used. Each ROADM has been equipped with one optical performance monitor able to perform both optical power and OSNR measurements. The DICONET testbed is also equipped with an

NMS which interfaces with the control and optical nodes and provides a graphical representation of the network that allows its manager to monitor the traffic or potential failures.

In [3] the testbed was used to experimentally test the performance of the two architectures in a scenario with lightpath requests arriving and departing dynamically. It was demonstrated that the distributed scheme yields lower setup times in highly dynamic traffic conditions (Fig. B.4), benefiting from the parallel lightpath establishments. The centralized scheme, on the other hand, experiences better blocking ratio justified by the sophisticated impairment-aware routing process it employs (Fig. B.5). The centralized nature of this architecture allows the routing engine to have a complete picture of the physical layer and traffic conditions, yet only one connection request may be served at a time, thus affecting the connection setup times.

In these initial NPOT implementations, a guard time was left between two consecutive route computations, leaving enough time for the GMPLS OSPF-TE protocol to disseminate the new PLI and wavelength availability information in order to update the PPD and TED databases. This forces the NPOT to remain idle some seconds between route computations (i.e., the OSPF-TE flooding time), which affects the overall lightpath setup times. In this article we implemented an alternative strategy to improve the lightpath setup time for the centralized scheme. Specifically, once a lightpath is selected, the involved wavelength is directly pre-reserved in order to avoid its usage in upcoming lightpath requests; this eliminates the waiting time for the OSPF-TE flooding. The state of the pre-reserved resources can be changed to “reserved” or “free” based on the information of the OSPF-TE updates. By applying this strategy, as shown in Fig. B.4, the lightpath setup time is significantly reduced, particularly for high offered load. In addition the centralized scheme with the pre-reservation strategy maintains the same low blocking ratio as depicted in Fig. B.5.

Reference	Description	Lightpath setup time
T. Tsuritani et al.	Centralized PCE based	Within 10s
F. Cugini et al.	Centralized PCE based	Within 10s
R. Martinez et al.	Distributed	8-9 s

Table B.1: Relevant works.

There have been various other research efforts that also employ an impairment-enabled control plane, focused on either centralized [13] [14] or distributed approaches [15]. Table B.1 reports the lightpath setup time reported in some of these works to highlight the potential of the DICONET architecture which not only utilizes algorithms that incur low blocking ratio but also manages to achieve low

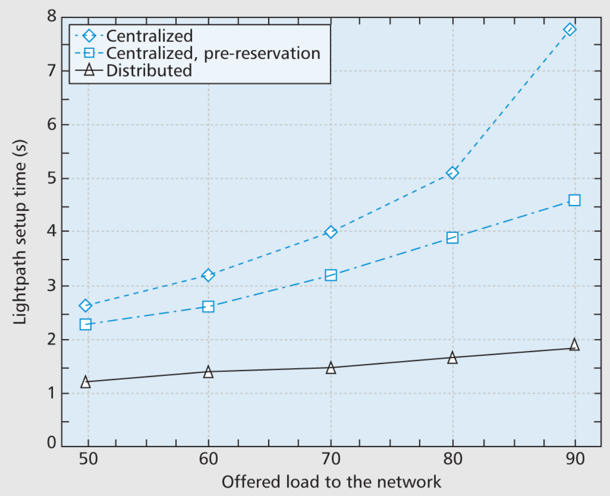


Figure B.4: Measured lightpath setup time of the two schemes for an operational scenario with the traffic load varying from 5090 Erlangs; the setup time of the centralized approach is improved using an alternative strategy (square marks, dashed line) where the wavelengths are pre-reserved to avoid waiting for the OSPF-TE flooding.

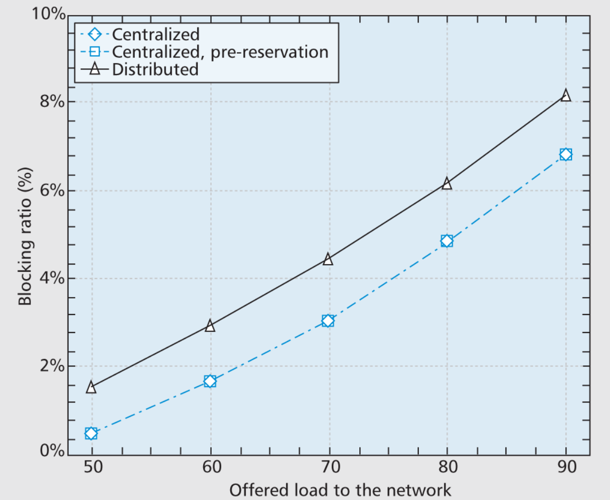


Figure B.5: Corresponding blocking ratio results of the two schemes.

setup times; lower than 5 s are achieved in the centralized case and lower than 2 s in the distributed case for all traffic scenarios.

In addition to the dynamicity offered by the integrated solution, DICONET enables continuous and robust operation by properly reacting to potential failures and restoring the affected traffic in a transparent fashion. In the effort to evaluate these capabilities we utilized the centralized configuration in a scenario where failures occur randomly throughout the network. The network was loaded with a set of predetermined active lightpaths, and then independent failures were caused by emulating link cuts in random locations. The results showed fast restoration times, despite the sequential processing of the disrupted lightpaths due to the centralized architecture. Indeed, it was shown in [16] that 72 percent of the lightpath restorations were performed in less than 5 s.

B.7 Capital resource optimization

In addition to having strong technical features, it is essential for any venture requiring investment to be coupled by a viable business case that highlights the advantages over other existing methods. Intelligent RWA, optimized component placement, failure localization and resilience, all integrated in a unified control plane, provide the network with an extended level of optimization. Indeed the networking solution that DICONET proposes is realized through a set of resource optimization algorithms that also introduce a cost benefit. The cost benefit that is gained with the DICONET solution was studied and quantified, focusing on the concepts of impairment awareness and reconfiguration ability.

The DICONET RWA algorithms were utilized to define the required network resources and compare an impairment-aware solution to an impairment-unaware solution in financial terms. Furthermore, in the analysis different commercially available reconfigurable node architectures were considered since R-OXCs account for a large fraction of the capital cost of the transport plane. Each type of R-OXC bears a different degree of flexibility in their add/drop capabilities (bearing or not the features directionless, colorless, contentionless) and therefore impact in a different way the network planning [17]. The capital and operational cost associated with each type is essentially determined by its physical implementation.

The analysis covered scenarios with transparent and translucent topologies [18], and utilized the offline IA-RWA and pure RWA algorithms developed in the project to estimate the cost of the required components (i.e., amplifiers, add/drop terminals, transponders, network interfaces, regenerators) of each planning solution and each node architecture. Figures B.6 and B.7 illustrate results indicative of this analysis. In particular, they correspond to the capital cost estimations of a transparent network for both cases impairment-aware (IA) and impairment unaware (IUA) with respect to the traffic load. Evidently, the IUA solution lacks the optimization capabilities of a QoTaware process and requires additional capital investment sooner than the IA solution as the traffic increases. For the same topology a comparison between a colorless (Fig. B.6) and a colored (Fig. B.7) node is presented. Colorless add/drop ports, unlike colored add/drop ports, do not

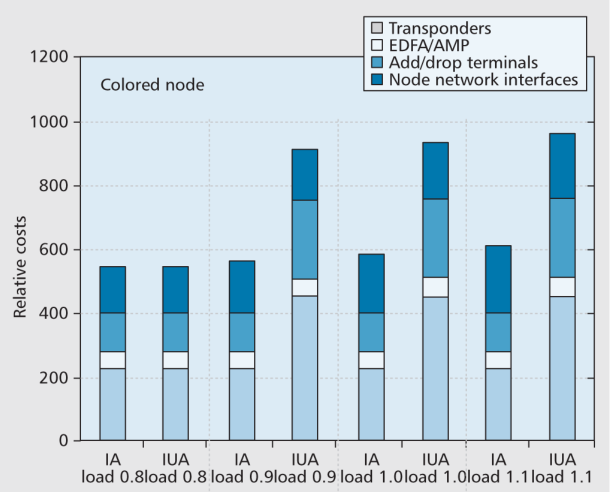


Figure B.6: CapEx of IA vs. IUA for two different types of node architectures, a colored and a colorless one.

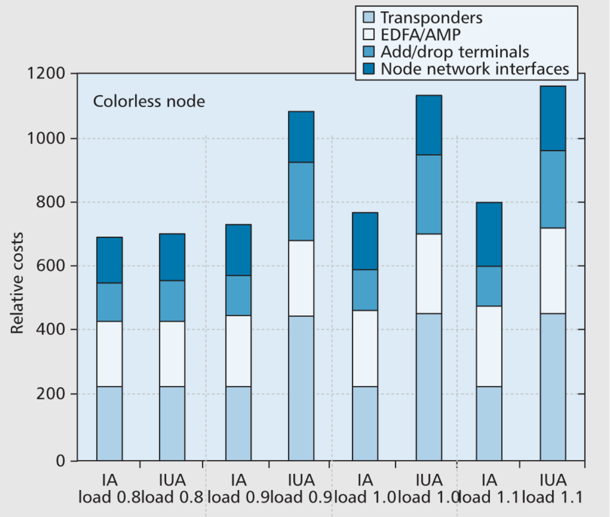


Figure B.7: CapEx of IA vs. IUA for two different types of node architectures, a colored and a colorless one.

have permanently assigned channels. This feature leads to higher capital expenditures for the case of the more flexible architecture (colorless) that are dominated by the number of add/drop terminals. Nevertheless, in the presence of dynamic traffic, deploying a flexible node facilitates online connection provisioning and minimizes manual interventions. Overall, the techno-economic analysis gave promising indications for the feasibility of the DICONET solution in commercial networks.

B.8 Conclusion

Extensive simulations, emulations, and experiments shaped DICONET to bear all the attributes the consortium sought after to fulfill the vision of dynamic and impairment-aware networking that maximizes the utilization of the existing WDM infrastructure. More than an architectural enhancement, DICONET empowers operators with useful tools applicable to both the planning and operational phases of a core optical network that is either transparent or translucent with features not limited to the control and management planes but that actually take advantage of the optical layer in an integrated cross-layer manner. Resource optimization, dynamicity, and resilience outline a network that offers not only quality of service but also cost effectiveness. DICONET effectively utilizes todays technologies to optimize the network and paves the way for a smooth migration to the next generation core. Driven by technological evolution, the future core network is going to enjoy an increased degree of dynamicity with higher bit rates, mixed transmission characteristics, and gridless network components. To also achieve resource optimization in the future flexible networks, issues such as the evaluation of signals with advanced transmission parameters and dynamic bandwidth allocation ought to be investigated.

References

[1] S. Azodolmolky et al. *A Dynamic Impairment-Aware Networking Solution for Transparent Mesh Optical Networks*. IEEE Communications Magazine, 2009.

[2] M. Ruffini et al. *Cost study of Dynamically Transparent Networks*. In proc. Optical Fiber Communications, paper OMG2, 2008.

[3] S. Azodolmolky et al. *Experimental Demonstration of an Impairment Aware Network Planning and Operation Tool for Transparent/Translucent Optical Networks*. IEEE/OSA Journal of Lightwave Technology, vol. 29, no. 4, pp. 439-448, Feb. 2011.

[4] S. Azodolmolky et al. *A Survey on Physical Layer Impairments Aware Routing and Wavelength Assignment Algorithms in Optical Networks*. Computer Networks Vol 53 (7) pp. 926-944, May 2009.

[5] S. Azodolmolky et al. *An offline impairment aware RWA algorithm with dedicated path protection consideration*. In proc. Optical Fiber Communications, paper OWI1, 2009.

[6] K. Manousakis et al. *Offline impairment-aware routing and wavelength assignment algorithms in translucent WDM optical networks.* Journal of Lightwave Technology, vol.27 (12), pp.1866-1877, June 2009.

[7] K. Christodopoulos et al. *Offline routing and wavelength assignment in transparent WDM networks.* IEEE/ACM Transactions on Networking, vol. 18, no. 5, pp. 1557-1570, Oct. 2010.

[8] M. Youssef et al. *Cross Optimization for RWA and Regenerator Placement in Translucent WDM Networks.* In Proc. IFIP ONDM, 2010.

[9] K. Christodopoulos et al. *Indirect and Direct Multicost Algorithms for Online Impairment-Aware RWA.* IEEE/ACM Transactions on Networking, vol. 19, no. 6, pp. 1759-1772, Dec. 2011.

[10] K. Manousakis et al. *Joint Online Routing, Wavelength Assignment and Regenerator Allocation in Translucent Optical Networks,.* Journal of Lightwave Technology, vol.28 (8), pp.1152-1163, Apr. 2010.

[11] E. Doumith et al. *Monitoring-Tree: An Innovative Technique for Failure Localization in WDM Translucent Networks.* In Proc. IEEE GLOBECOM '10, Dec. 2010.

[12] S. Spadaro et al. *Experimental Demonstration of an Enhanced Impairment-Aware Path Computation Element.* In proc. Optical Fiber Communications, paper OMW5, 2011.

[13] T. Tsuritani et al. *Optical Path Computation Element Interworking with Network Management System for Transparent Mesh Networks.* In proc. Optical Fiber Communications, paper NFW5, 2008.

[14] F. Cugini et al. *Implementing A Path Computation Element (PCE) to Encompass Physical Impairments in Transparent Networks.* In proc. Optical Fiber Communications, paper OWK6, 2007.

[15] R. Martinez et al. *Experimental GMPLS Routing for Dynamic Provisioning in Translucent Wavelength Switched Optical Networks.* In proc. Optical Fiber Communications, paper NTuB4, 2009.

[16] F. Agraz et al. *Experimental evaluation of path restoration for a centralised impairment-aware GMPLS-controlled all-optical network.* In ECOC '09. 35th European Conference on Optical Communication, 2009.

[17] K. Manousakis et al. *Performance Evaluation of Node Architectures with Color and Direction Constraints in WDM Networks,.* In Proc. IEEE GLOBECOM '10, 2010.

[18] M. De Groote et al. *Cost comparison of different Translucent Optical Network Architectures*. In Proc. Conference of Telecommunication, Media and Internet Techno-Economics, CTTE 2010, 2010.

References

[1] S. De Maesschalck et al. *Intelligent Optical Networking for Multilayer Survivability.* IEEE Communications Magazine, pp. 42-49, June 2002.

[2] International Telecommunications Union. *50 Years of Excellence.* http://www.itu.int/itudoc/gs/promo/tsb/88192.pdf, 2006.

[3] Wikipedia. *Wikipedia: Telegraphy.* http://en.wikipedia.org/wiki/Telegraphy, retrieved November 2010.

[4] J. Tyndall et al. *Notes of a course of nine lectures on light delivered at the Royal institution of Great Britain April 8-June 3, 1869.* http://www.archive.org/download/notesofcourseofn00tyndrich, 1870.

[5] A. Einstein. *Zur Quantentheorie der Strahlung (On quantum theory of radiation).* physikalische Zeitschrift, 1917.

[6] R. Gordon Gould. *The LASER, Light Amplification by Stimulated Emission of Radiation.* In The Ann Arbor Conference on Optical Pumping, the University of Michigan, 15 June through 18 June, 1959.

[7] E. Fred Schubert. *Light-Emitting Diodes, 2nd Ed.* 2006.

[8] ITU-T Standardization Organization. *ITU-T recommendation G.652/3/4/5.* 2009.

[9] Y. Ishii et al. *MEMS-based 143 wavelength-selective switch with flat passband.* In ECOC '09. 35th European Conference on Optical Communication, 2009.

[10] S. Gringeri et al. *Flexible Architecture for flexible transport nodes and networks.* IEEE Communications Magazine, 2010.

[11] R. D. Doverspike et al. *Future Transport Network Architectures.* IEEE Communications Magazine, Special Issue on Reliable Communication Networks, August 1999.

[12] A. Farrel and Y. Bryskin. *GMPLS: Architecture and Applications.* Elsevier, 2005.

[13] Ed. K. Kompella and Ed. Y. Rekhter. *Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS), RFC 4202*. IETF standards track, 2005.

[14] D. Katz et al. *Traffic Engineering (TE) Extensions to OSPF Version 2, RFC 3630*. IETF standards track, 2003.

[15] T. Li and H.Smit. *IS-IS Extensions for Traffic Engineering, RFC 3784*. IETF standards track, 2008.

[16] D. Awduche et al. *RSVP-TE: Extensions to RSVP for LSP Tunnels, RFC 3209*. IETF standards track, 2001.

[17] Ed. B. Jamoussi. *Constraint-Based LSP Setup using LDP, RFC 3212*. IETF standards track, 2002.

[18] Ed. J. lang. *Link Management Protocol (LMP), RFC 4204*. IETF standards track, 2005.

[19] A. Farrel et al. *A Path Computation Element (PCE)-Based Architecture, RFC 4655*. IETF standards track, 2006.

[20] J. P. Vasseur et al. *Network recovery, protection and restoration of optical, SONET-SDH, IP and MPLS*. Elsevier, 2004.

[21] E. Mannie and D. Papadimitriou. *Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS), RFC 4427*. IETF standards track, 2006.

[22] ITU-T Standardization Organization. *ITU-T Recommendation G.807/Y.1302, "Requirements for automatic switched transport networks (ASTN)*. July 2001.

[23] B. T. Doshi et al. *Optical network design and restoration. Bell Labs Tech. J. 4, pp. 58-84*. 1999.

[24] N. Wouters et al. *Survivability in a New Pan-European Carriers' Carrier Network Based on WDM and SDH Technology: Current Implementation and Future Requirements*. IEEE Communications Magazine, Aug 1999.

[25] C. Guillemot et al. *VTHD French NGI Initiative: IP and WDM Interworking with WDM Channel Protection*. In Proceedings IP-over-WDM conference, 2000.

[26] P. Demeester et al. *Resilience in multi-layer networks*. IEEE Communications Magazine, August 1998.

[27] F. Aslam et al. *Interdomain Path Computation: Challenges and Solutions for Label Switched Networks.* IEEE-Communications Magazine Volume 45 (10), pp. 94 - 101, Oct. 2007.

[28] Admela Jukan et al. *End-to-End Service Provisioning in Multi-granularity Multi-domain Optical Networks.* In Proceedings of ICC 2004, IEEE International Conference on Communication, June 2004.

[29] Srinivasan Seetharaman et al. *End-to-End Dedicated Protection in Multi-Segment Optical Networks.* Technical report, http://www.stanford.edu/s̃eethara/papers/e2eprot.pdf, 2003.

[30] Thomas Engel et al. *Increasing End-to-End Availability over Multiple Autonomous Systems.* In Proceedings of PDPTA'05, Las Vegas, USA, 2005.

[31] Thomas Schwabe et al. *Resilient Routing Using ECMP and MPLS.* In Proceedings of HPSR 2004, Phoenix, AZ, USA, 2004.

[32] D-L Truong et al. *Recent Progress in Dynamic Routing for Shared Protection in Multidomain Networks.* IEEE Communications Magazine, Vol 46 (6), pp. 112-119, June 2008.

[33] H. Drid et al. *A survey of survivability in multi-domain optical networks.* Computer Communications Vol 33 (8) pp. 1005-1012, May 2010.

[34] J. Szigeti et al. *Adaptive Multi-Layer Traffic Engineering with Shared Risk Group Protection.* In Proceedings of IEEE International Conference on Communications, ICC '08, pp. 5367-5371, 2008.

[35] J. Hawkinson and T. Bates. *Guidelines for creation, selection and registration of an Autonomous System, RFC 1930.* IETF standards Track, March 1996.

[36] D. Staessens et al. *A Quantitative Comparison of Some Resilience Mechanisms in a Multidomain IP-over-Optical Network Environment.* In Proceedings of IEEE International Conference on Communications, ICC '06, pp. 2512-2517, 2006.

[37] D. Staessens et al. *Enabling High Availability over Multiple Optical Networks.* IEEE Communications Magazine, Vol 46 (6), pp. 120-129, June 2008.

[38] P. Cholda et al. *Reliability Assessment of Optical p-Cycles.* IEEE/ACM Transactions on Networking,Vol. 15, no. 6, pp. 1579-92, Dec. 2007.

[39] Akyamac et al. *Reliability in Single Domain vs. Multi Domain Optical Mesh Networks.* In Proc. NFOEC, Dallas TX, Sept. 2002.

[40] Lewis et al. *Introduction to Reliability Engineering*. John Wiley & Sons, 1987.

[41] J.W. Suurballe and R.E. Tarjan. *A quick method for finding shortest pairs of disjoint paths*. Networks, vol. 14, pp. 325336, 1984.

[42] IBM ILOG CPLEX Optimizer. *http://www-01.ibm.com/software/integration/optimization/cplex-optimization-studio/*. 2012.

[43] A. Farrel et al. *Inter-Domain MPLS and GMPLS Traffic Engineering Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions, RFC 5151*. IETF standards track, 2008.

[44] CY. Lee et al. *GMPLS Segment Recovery, RFC 4873*. IETF standards track, 2008.

[45] L. Berger et al. *Exclude Routes - Extension to Resource ReserVation Protocol-Traffic Engineering (RSVP-TE), RFC 4874*. IETF standards track, 2008.

[46] J.P. Lang et al. *RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery, RFC 4872*. IETF standards track, 2008.

[47] Biswanath Ramamurthy et al. *Transparent vs. opaque vs. translucent wavelength-routed optical networks*. In Proc. Optical Fiber Communications, 1999.

[48] A. Morea et al. *A critical analysis of the possible cost savings of translucent networks*. In Proceedings of DRCN '05, 2005.

[49] G. Li et al. *Resilience design in all-optical ultralong-haul networks*. J. Opt. Netw. 5, pp. 625-636, 2006.

[50] R. Ramaswami and K.N. Sivarajan. *Routing and Wavelength Assignment in All-Optical Networks*. IEEE/ACM Transactions on Networking Vol 3 (5), pp. 489-500, 1995.

[51] E.W. Dijkstra. *A Note on Two Problems in Connexion with Graphs*. Numerische Mathematik 1,pp. 269-271, 1959.

[52] P.-H. Ho et al. *Spare capacity allocation for WDM mesh networks with partial wavelength conversion capacity*. In High Performance Switching and Routing, 2003, HPSR. Workshop on, 2003.

[53] Y. Liu et al. *Approximating Optimal Spare Capacity Allocation by Successive Survivable Routing Yu.* IEEE/ACM Transactions on Networking, vol 13 no 1, February 2005.

[54] ITU-T Standardization Organization. *Spectral Grids for WDM, ITU-T recommendation G.694.1.* 2001.

[55] M. Gunkel et al. *A Cost Model for the WDM Layer.* In Proc. Photonics in Switching, 2006.

[56] R. Huelsermann et al. *Cost modeling and evaluation of capital expenditures in optical multilayer networks.* OSA Journal of Optical Networking, vol. 7, no. 9, 2008.

[57] M. De Groote et al. *Cost comparison of different Translucent Optical Network Architectures.* In Proc. Conference of Telecommunication, Media and Internet Techno-Economics, CTTE 2010, 2010.

[58] STRONGEST. *Scalable, Tunable and Resilient Optical Networks Guaranteeing Extremely-high Speed Transport.*

[59] D. Staessens et al. *Cost Efficiency of Protection in Future Transparent Networks.* In Proceedings ICTON '09, July 2009.

[60] K. R. Gabriel and R. R. Sokal. *A new statistical approach to geographic variation analysis.* Systematic Zoology (Society of Systematic Biologists) 18 (3) pp. 259270, 1969.

[61] I. Katzela and M. Schwartz. *Schemes for Fault Identification in Communication Networks.* IEEE/ACM Transactions on Networking, Vol. 3 (6), pp. 753-764, 1995.

[62] M. Choi J. Choi and S.H. Lee. *An Alarm Correlation and Fault Identification Scheme Based on OSI Managed Object Classes.* In Proc. IEEE Int'l Conference on Communications ICC '99, pp.1547-1551, 1999.

[63] M. Steinder and A.S. Sethi. *Non-Deterministic Diqgnosis of End-to-End Service Failures in a Multi-Layer Communication System.* In Proc. IEEE Int'l Conference on Computer Communications and Networks '01, pp. 374-379, 2001.

[64] D.L. Yang C.S. Chao and A.C. Liu. *An automated Fault Diagnosis System Using Hierarchical Reasoning and Alarm Correlation.* Journal of Network and Systems Management,Vol.9 (2), pp.183-202, 2001.

[65] C. Mas and P. Thiran. *An efficient algorithm for locating soft and hard failures in WDM networks.* IEEE Journal On Selected Areas In Communications, Vol. 18 (10), Oct. 2000.

[66] T. Wu and A.K. Somani. *Necessary and Sufficient Condition for k Crosstalk attacks localization in All-optical Networks.* In Proc. IEEE Globecom, pp. 2541-2546, 2003.

[67] C. Huang H. Zeng and A. Vukovic. *A novel fault detection and localization scheme for mesh all-optical networks based on monitoring-cycles.* Photonic Network Communication, Vol. 11, pp.277-286, 2006.

[68] T. Wu and A.K. Somani. *Attack monitoring and localization in all optical networks.* In Proc. of SPIE Opticommun 2002: Optical Networking and Communications, vol. 4874, pp. 235-248, 2002.

[69] J. Tapolcai et al. *On monitoring and failure localization in mesh all-optical networks.* In Proc. IEEE INFOCOM '09, 2009.

[70] S. Stanic and S. Subramaniam. *A Comparison of Flat and Hierarchical Fault-Localization in Transparent Optical Networks.* In Proc. Optical Fiber Communications, 2008.

[71] et al. J. Kim. *Fault Localization for Heterogeneous Networks Using Alarm Correlation on Consolidated Inventory Database.* In Springer LNCS , pp. 82-91, 2008.

[72] R. Ramaswami and K. N. Sivarajan. *Optical Networks, A Practical Perspective, 2nd ed.* Elsevier, 2001.

[73] The IEC. *IEC86C, Document 62343-6-6, DYNAMIC MODULES, Failure Mode Effect Analysis for Optical Units of Dynamic Modules.*

[74] R.G. Gallager. *Information Theory and Reliable Communication.* John Wiley & Sons, 1968.

[75] S. Azodolmolky et al. *A Dynamic Impairment-Aware Networking Solution for Transparent Mesh Optical Networks.* IEEE Communications Magazine, 2009.

[76] M. Ruffini et al. *Cost study of Dynamically Transparent Networks.* In proc. Optical Fiber Communications, paper OMG2, 2008.

[77] S. Azodolmolky et al. *Experimental Demonstration of an Impairment Aware Network Planning and Operation Tool for Transparent/Translucent Optical Networks.* IEEE/OSA Journal of Lightwave Technology, vol. 29, no. 4, pp. 439-448, Feb. 2011.

[78] S. Azodolmolky et al. *A Survey on Physical Layer Impairments Aware Routing and Wavelength Assignment Algorithms in Optical Networks.* Computer Networks Vol 53 (7) pp. 926-944, May 2009.

[79] S. Azodolmolky et al. *An offline impairment aware RWA algorithm with dedicated path protection consideration.* In proc. Optical Fiber Communications, paper OWI1, 2009.

[80] K. Manousakis et al. *Offline impairment-aware routing and wavelength assignment algorithms in translucent WDM optical networks.* Journal of Lightwave Technology, vol.27 (12), pp.1866-1877, June 2009.

[81] K. Christodopoulos et al. *Offline routing and wavelength assignment in transparent WDM networks.* IEEE/ACM Transactions on Networking, vol. 18, no. 5, pp. 1557-1570, Oct. 2010.

[82] M. Youssef et al. *Cross Optimization for RWA and Regenerator Placement in Translucent WDM Networks.* In Proc. IFIP ONDM, 2010.

[83] K. Christodopoulos et al. *Indirect and Direct Multicost Algorithms for Online Impairment-Aware RWA.* IEEE/ACM Transactions on Networking, vol. 19, no. 6, pp. 1759-1772, Dec. 2011.

[84] K. Manousakis et al. *Joint Online Routing, Wavelength Assignment and Regenerator Allocation in Translucent Optical Networks,.* Journal of Lightwave Technology, vol.28 (8), pp.1152-1163, Apr. 2010.

[85] E. Doumith et al. *Monitoring-Tree: An Innovative Technique for Failure Localization in WDM Translucent Networks.* In Proc. IEEE GLOBECOM '10, Dec. 2010.

[86] S. Spadaro et al. *Experimental Demonstration of an Enhanced Impairment-Aware Path Computation Element.* In proc. Optical Fiber Communications, paper OMW5, 2011.

[87] T. Tsuritani et al. *Optical Path Computation Element Interworking with Network Management System for Transparent Mesh Networks.* In proc. Optical Fiber Communications, paper NFW5, 2008.

[88] F. Cugini et al. *Implementing A Path Computation Element (PCE) to Encompass Physical Impairments in Transparent Networks.* In proc. Optical Fiber Communications, paper OWK6, 2007.

[89] R. Martinez et al. *Experimental GMPLS Routing for Dynamic Provisioning in Translucent Wavelength Switched Optical Networks.* In proc. Optical Fiber Communications, paper NTuB4, 2009.

[90] F. Agraz et al. *Experimental evaluation of path restoration for a centralised impairment-aware GMPLS-controlled all-optical network.* In ECOC '09. 35th European Conference on Optical Communication, 2009.

[91] K. Manousakis et al. *Performance Evaluation of Node Architectures with Color and Direction Constraints in WDM Networks,.* In Proc. IEEE GLOBECOM '10, 2010.