UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Ph.D. Thesis

End-to-End Quality of Service Provisioning in Multilayer and Multidomain Environments

Author:
Víctor López Álvarez
Telecommunication Engineer

Supervisors:
Prof. Javier Aracil Rico
Dr. José Alberto Hernández Gutiérrez

Madrid, 2009

DOCTORAL THESIS:	End-to-End Quality of Service Provisioning in Multilayer and Multidomain Environments
AUTHOR:	Víctor López Álvarez
SUPERVISORS:	Prof. Javier Aracil Rico Dr. José Alberto Hernández Gutiérrez

The committee for the defense of this doctoral thesis is composed by:

PRESIDENT:	Prof. Maurice Gagnaire
MEMBERS:	Dr. Franco Callegati
	Dr. Mikel Izal Azcárate
	Prof. Josep Solé Pareta
SECRETARY:	Dr. Jorge López de Vergara

In memory of my cousin Miguel Ángel Gutiérrez Álvarez

Summary

This Ph.D. thesis deals with the integration of Traffic Engineering (TE) mechanisms in multi-layer and multi-domain Quality of Service (QoS) aware backbone networks. The advent of reconfigurable optical equipment has given raise to a ultrahigh bandwidth dynamic transport layer that can be controlled by the Generalized Multiprotocol Label Switching (GMPLS) protocol. Internet backbone networks are migrating to an IP over WDM network where the QoS provisioning is not yet defined. It is necessary to offer the desired QoS to each service without incurring in resource overprovisioning.

Multi-layer capable routers are able to route the incoming traffic using the IP layer, but also they can ask for extra resources to the optical layer in order to establish end-to-end connections. The delay of end-to-end connections is negligible compared with the delay suffered at the IP layer. On the other hand, the IP layer is more efficient, thanks to the statistical multiplexing, thus reducing the required equipment. This thesis proposes three solutions to deal with this Multi-layer Traffic Engineering (MTE) problem.

The reader may, firstly, find a discussion about the feasibility of the integration of MTE mechanisms in real backbone networks. The main strengths and drawbacks of such integration are introduced. Some case studies about the integration of MTE algorithms in backbone networks are studied to show its feasibility. However, “Threshold-based” algorithms show that there are two opposite objectives: performance optimization and resource utilization.

The second MTE solution is based on the Bayesian decision theory. This solution uses a novel methodology to deal with the multi-layer decision problem. The incoming traffic has several QoS requirements which invalidate the service offer if not fulfilled. On the other hand, the IP layer equipment is already deployed and at the operator is willing to use. Consequently, the incoming traffic is routed using the IP layer until the QoS requirements are degraded. At this point, the Bayesian decisor asks for extra optical resources in order to satisfy the incoming traffic demands. This methodology is validated from a single-node scenario to multi-domain networks throughout this thesis. To provide a realistic network scenario, the state of the art of traffic characterization is also reviewed.

Finally, the third solution proposed is the extension of Flow-Aware Networking (FAN) to IP over WDM networks. FAN allows QoS provisioning at IP layer, but it is not multilayer aware. The Multi-layer FAN (MFAN) architecture defines a node structure to decide when the optical resources should be used and which flows are more suitable to be transmitted through the optical domain. This MFAN solution does not require a complex monitoring infrastructure, since it uses some of the modules of standard FAN.

Resumen

Esta tesis doctoral está centrada en la integración de mecanismos de Ingeniería de Tráfico en redes multicapa y multidominio para la provisión de calidad de servicio. Con la llegada de equipamiento óptico reconfigurable y la definición de GMPLS es posible utilizar el gran ancho de banda de la capa óptica de manera dinámica. Además las redes troncales están migrando a una arquitectura de IP sobre WDM donde la provisión de calidad de servicio aún no está definida. Es necesario ofrecer la calidad de servicio deseada sin recurrir al sobredimensionamiento.

Los routers multicapa permiten enviar el tráfico entrante usando la capa IP, pero también pueden solicitar recursos extra a la capa óptica para establecer conexiones extremo a extremo. El retardo de estas conexiones extremo a extremo es despreciable en comparación con el retardo sufrido en la capa IP. La capa IP es más eficiente gracias a la multiplexación estadística, reduciendo el equipamiento necesario para la provisión de servicios. Esta tesis propone tres soluciones para resolver este problema de Ingeniería de Tráfico multicapa.

El lector puede encontrar primero una discusión sobre la viabilidad de integrar estos mecanismos de Ingeniería de Tráfico multicapa en redes troncales reales. Las ventajas e inconvenientes de esta integración se muestran en esa discusión. Se han realizado algunos estudios sobre la integración de estos mecanismos en redes troncales para mostrar su viabilidad. Sin embargo, los algoritmos basados en umbral muestran que hay dos objetivos contrapuestos: el rendimiento y la utilización de los recursos.

El segundo método de Ingeniería de Tráfico multicapa está basada en la teoría de decisión bayesiana. Esta solución utiliza una metodología novedosa para tratar el problema de decisión multicapa. El tráfico entrante tiene requisitos de calidad de servicio que si no son cumplidos hacen que el servicio no tenga ningún valor. El equipamiento IP está desplegado actualmente en la red, por lo que el operador quiere utilizarlo. Por lo tanto, el tráfico entrante es enrutado utilizando los recursos de la capa IP hasta que los requisitos de calidad de servicio son degradados. En este momento, el decisor bayesiano solicita recursos extra a la capa óptica para dar servicio a las nuevas demandas. Esta metodología es validada desde escenarios con un único nodo hasta escenarios con múltiples dominios en esta tesis. Para ofrecer un escenario de red real, el estado del arte sobre la caracterización de tráfico ha sido revisado.

Finalmente, la tercera solución propuesta es una extensión de Flow-Aware Networking (FAN) para redes IP sobre WDM. FAN ofrece calidad de servicio a nivel IP, pero no está orientado a redes multicapa. La arquitectura Multi-layer FAN (MFAN) propuesta define la estructura de nodo para decidir cuándo se deben usar los recursos ópticos y cuáles deben ser los flujos enviados a la capa óptica. La solución de MFAN no requiere una infraestructura de monitorización compleja, sino que utiliza los módulos definidos por FAN.

Acknowledgments

Firstly, I would like to thank my family for their support, for their understanding and for putting up with me, every time, every moment in this hard, tedious and non-stop work. Everyday, when I arrive home, I find love, talk and calm, all that I need. Especially, I would like to remind my cousin Miguel Ángel who taught me to have a smile for everybody and to “live deep and suck out all the marrow of life”[1].

Also, I would like to thank all my friends for understanding my situation, for their help, advises and funny times that help me disconnect during this complicated period.

When I was finishing my M.Sc. degree in Telecommunications Engineering, I was looking for an interesting research topic. I knew I wanted to do research, but I could not find any opportunity, so I decided to go to Finland as an Erasmus student. With my Erasmus granted confirmed, I submitted my CV to Telefónica I+D and Jesús Felipe Lobo Poyo thought that I could join them, so I decided to stay in Spain. Thank you Jesús for that opportunity which changed my life. I also would like to thank my colleagues from Telefónica: Juan Fernández Palacios, Óscar González de Dios, Javier Jiménez, Carlos García and Raúl Duque, who introduced me to the research community and helped me during this Ph.D., because we still collaborate. Let me also thank the members of the Space Research Group (SRG) at Universidad de Alcalá (UAH), with whom I worked during my M.Sc. thesis and we have collaborated during my Ph.D. period.

Very importantly, I would like to thank my supervisors Jose Alberto Hernández and Javier Aracil for their close collaboration, their advises and for helping me to grow as a researcher and as a person. Thank you Javier for giving me the opportunity to become a researcher and a teacher. In the same way, let me thank my colleagues from the Networking Research Group: José Luis García, Felipe Mata, Jorge López de Vergara, Sergio López, Antonio Martínez, Walter Fuertes, Alfredo Salvador, Luis de Pedro, Javier Ramos, Pedro Santiago and, the last to join us, but not the least, Bas Huiszoon. Thank you for all this

[1] From the film Dead Poets Society

time at the laboratory and for making the lab a nice place to work. This also definitely includes my other colleagues from the laboratory Iván González, Gustavo Sutter, Elías Todorovich, Juan González and Eduardo Boemo. All my work would not have been possible without the support of the Universidad Autónoma de Madrid and the Departamento de Informática.

A great experience during this Ph.D. period has been my internship at "Télécom Paris - École Nationale Supérieure des Télécommunications" in Paris. Thank you Prof. Gagnaire for hosting me in your group. In Paris, I found a very good co-worker and person: César Cárdenas. Thank you César for our long talks about so many research topics, your illusion for research and for encouraging me to work harder. I would like to thank my friend and colleague Óscar Salazar for the good time in Paris.

Finally, Marta, this thesis is partially done by you. Your company, your support and our talks encourage me to become a better worker and person. You make me stronger. Thank you for being my partner.

Contents

Summary **i**

Resumen **ii**

Acknowledgments **v**

Acronyms **xix**

1 Introduction **1**

1.1 Motivation 1

1.2 Objectives 2

1.3 Thesis structure 2

2 Feasibility of the Multi-layer integration in network operators **5**

2.1 Evolution of IP and transport networks 6

2.2 Multi-layer networking in the Next Generation Transport Networks 8

2.2.1 Control Plane Interconnection 10

2.2.2 Data Plane Integration 16

2.3 Multi-layer Traffic Engineering 17

2.4 Proposed Multi-layer Traffic Engineering Solutions 20

2.5 Case studies 22

2.5.1 Definition of basic MTE algorithms 22

2.5.2 Assumptions 23

2.5.3 Performance of the proposed algorithms 24

2.5.4 Impact on IP Equipment 28

2.6 Multi-layer Traffic Engineering Conclusions 30

3 Bayesian decisor of a multi-layer capable router 31
3.1 Motivation and model assumptions 31
3.1.1 On the use of Utility functions 34
3.1.2 Traffic model 35
3.2 Analysis 39
3.2.1 Problem statement 39
3.2.2 The utility function $U(x)$ 41
3.2.3 The utilization cost of electronic and optical switching 43
3.3 Experiments and results 44
3.3.1 Scenario definition 44
3.3.2 Basic decisor behaviour 44
3.3.3 Study of threshold $T_{\max}$ 45
3.3.4 Analysis with different R_{cost} values 46
3.3.5 Influence of the Hurst parameter H 47
3.3.6 Impact of the mean and variance of the LSPs 50
3.4 Dynamic behavior of the risk process 52
3.4.1 Analysis 52
3.4.2 Numerical example 54
3.5 Conclusions 56

4 End to end service provisioning in Multi-layer networks 57
4.1 Problem statement: Multi-hop scenario 57
4.1.1 Risk function definition 58
4.1.2 Cost of using resources 59
4.1.3 Utility functions definition 62
4.1.4 Illustrative example 65
4.2 Numerical results and discussion 65
4.2.1 Basic behavior of the algorithm 66
4.2.2 Decisor dynamics experiment 67
4.2.3 Influence of the utilization cost (R_{cost}) 68
4.2.4 Study of delay QoS threshold ($T_{\max}$) 70
4.2.5 Hurst parameter 71
4.2.6 On the influence of the path length in the decision 72
4.3 Conclusions 73

5 Evaluation of the Bayesian decisor in real networks 75
5.1 Risk definition in the full network topology . . . 76
5.1.1 Topology Definition . . . 76
5.1.2 Redefining the Risk function . . . 77
5.2 Problem solution . . . 79
5.2.1 Optimization Problem . . . 79
5.2.2 Heuristic Algorithm . . . 79
5.3 Performance evaluation in a backbone topology . . . 81
5.3.1 Scenario definition . . . 81
5.3.2 Validation of the algorithms . . . 82
5.3.3 Performance with the resource utilization cost variation (R_{cost}) . . . 83
5.3.4 Performance with the QoS parameters variation ($T_{\max}$) . . . 88
5.4 Detailed explanation of the Bayesian decisor behavior . . . 91
5.4.1 Utilization cost parameter variation (R_{cost}) . . . 92
5.4.2 QoS parameter variation ($T_{\max}$) . . . 94
5.5 Conclusions . . . 96

6 Analysis of the Bayesian decisor in a multi-domain scenario 99
6.1 Control plane information exchanged in multi-domain networks . . . 100
6.1.1 Routing in multi-domain networks . . . 101
6.1.2 Signaling in multi-domain networks . . . 102
6.1.3 Advances on multi-domain/multi-layer architectures . . . 103
6.2 Algorithms to characterize end-to-end delay . . . 105
6.2.1 Scenario definition . . . 107
6.2.2 Mean delay estimation . . . 109
6.2.3 Percentile estimation . . . 111
6.2.4 Fourier transform based algorithm . . . 113
6.2.5 Weibull mixture algorithm . . . 116
6.2.6 Performance comparison of the algorithms . . . 117
6.3 Risk function definition for multi-domain networks . . . 118
6.3.1 Problem statement . . . 118
6.3.2 Numerical results . . . 121
6.4 Conclusions . . . 124

7 A Multi-layer solution for Internet Service Providers 127

7.1 Motivation . . . 128

7.2 Quality of Service at the IP layer . . . 130

7.2.1 Integrated Services . . . 131

7.2.2 Differentiated Services . . . 132

7.3 Review of Flow Aware Networking . . . 135

7.4 Extension of FAN to an IP over WDM network architecture . . . 138

7.5 Experiments and results . . . 141

7.5.1 Simulation scenario description . . . 141

7.5.2 Admission control in the optical queue . . . 143

7.5.3 Implicit classification of FAN . . . 143

7.5.4 Flow routing policies over the optical queue . . . 144

7.6 Conclusions and future work . . . 150

8 Conclusions 153

8.1 Main contributions . . . 153

8.2 Assessment of the objectives . . . 154

8.3 Future work . . . 156

Conclusiones 157

References 163

List of Publications 175

List of Figures

2.1 Backbone Technologies Evolution 6

2.2 Next Generation Transport Network network reference 8

2.3 ASON architecture 9

2.4 Control architecture with an integrated model 12

2.5 Control architecture with separated control planes 13

2.6 Comparison between the Peer and the Segmented GMPLS models as a function of the control plane area of influence 15

2.7 Transponder configuration 16

2.8 Multi-layer Traffic Engineering Problems 18

2.9 Cut-through and hop-by-hop traffic 18

2.10 Traffic offloading over the optical transport network 19

2.11 Blocks Definition Offloading Problem 22

2.12 Backbone network topologies 24

2.13 Load of the NOBEL reference network with ECMP routing algorithm . . . 25

2.14 Load of the NOBEL reference network with ECMP routing algorithm and MTE algorithm (Longest) 26

2.15 Number of Overloaded links ($BW > C_{min}$) in each topology 27

2.16 Number of created By-passes 28

2.17 Percentage of off-loaded traffic over the total traffic 29

2.18 Number of IP cards required for the IP over WDM solution and MTE algorithms 29

3.1 Multi-layer capable router scenario 32

3.2 Decision make in a multi-layer capable router 33

3.3 Utility functions 35

3.4 Fluid queue model 37

3.5 Variation of the bitrate (b/s, averaged over intervals of 0.5 s) in the Bellcore trace 38
3.6 Bellcore trace FBM model validation 39
3.7 Multi-layer decision problem 40
3.8 Risk function example with $N = 60$, U_{mean} and $R_{\text{cost}} = 2$ 45
3.9 Bayes risk curves and optimal decisions (minimum values of risk) for several $T_{\max}$ values assuming hard real-time utility (left) and elastic utility (right) functions. Dashed line = Utility. 46
3.10 Variation of relative cost (R_{cost}) for U_{mean} 47
3.11 Variation of relative cost (R_{cost}) for U_{exp} 48
3.12 Variation of relative cost (R_{cost}) for U_{step} 48
3.13 Hurst parameter variation 49
3.14 Hurst parameter variation (risk level curves) 50
3.15 Mean LSPs variation 51
3.16 Variance LSPs variation 51
3.17 Risk versus optimal number of electronic LSPs for a total number of LSPs in the range $30 - 60$ 54
3.18 Evolution of $V_j(n)$ from different states 55

4.1 Multi-hop scenario under study 58
4.2 R_{cost} designed rule example 61
4.3 Risk Computation Example 65
4.4 Level curves examples without cross-traffic (left) and with cross-traffic (right) 66
4.5 LSPs sent through the electronic and optical layers at nodes 1 and 2 when the load in the first node increases 68
4.6 LSPs sent through the electronic and optical layer when the load in the second node increases 69
4.7 Variation of the $T_{\max}$ parameter (U_{exp}) 70
4.8 Variation of the $T_{\max}$ parameter (U_{step}) 71
4.9 Variation of the decision when the path length changes 72
4.10 LSPs sent in each hop when the load in the first node increases (left) electronic layer (e_i) (right) optical layer (o_i) 73

5.1 Blocks Definition Bayesian Decisor Problem 76
5.2 Five node network 77

5.3 Path definition example 78

5.4 15-node NSFNET topology 82

5.5 Decision when the traffic is increasing from node 1 to 5 (Umean) 83

5.6 Paths used when the traffic load increases 84

5.7 Number of paths used in each domain and their mean occupation when R_{cost} varies (U_{mean}) 85

5.8 Number of paths used in each domain and their mean occupation when R_{cost} varies (U_{step}) 86

5.9 Number of paths used in each domain and their mean occupation when R_{cost} varies (U_{exp}) 87

5.10 Number of paths used in each domain and their mean occupation when $T_{\max}$ varies (U_{step}) 89

5.11 Number of paths used in each domain and their mean occupation when $T_{\max}$ varies (U_{exp}) 90

5.12 Eight Network Topology 91

5.13 Decisor elections when R_{cost} parameter varies (U_{mean}) 92

5.14 Paths time-line when R_{cost} parameter varies 93

5.15 Decisor elections when R_{cost} parameter varies (U_{step}) 94

5.16 Decisor elections when $T_{\max}$ parameter varies (U_{step}) 95

5.17 Paths time-line when $T_{\max}$ parameter varies 96

5.18 Decisor elections when $T_{\max}$ parameter varies (U_{exp}) 97

6.1 Multi-domain Scenario: Autonomous Systems description 100

6.2 IP address space in multi-domain networks 101

6.3 End-to-end signaling in multi-domain networks 102

6.4 Path establishment in a PCE based architecture 103

6.5 Measurement scenarios 106

6.6 Network scenario - NSFNet T1 Backbone 107

6.7 A path with an end-to-end flow and cross traffic 107

6.8 Histograms of the CAIDA trace interarrival times and packet lengths. . . . 108

6.9 Structure of the simulated nodes 108

6.10 Relative Error of the mean estimation 110

6.11 LOT algorithm steps to compute coefficients (by courtesy of [1]) 114

6.12 Histogram and pdf estimated by Fourier method 115

6.13 FFT coefficients required to pass the chi-square test (significance level 1%) 115

6.14 Histogram and pdf estimated by the mixtures of Weibull distributions method 117
6.15 Monitoring information in a single-domain 120
6.16 View of a node in the AS 3 of the overall network 120
6.17 Multi-domain reference scenario 122
6.18 Number of LSPs routed in the electronic and optical layer when the first node increases its traffic 123
6.19 Number of LSPs routed in the electronic and optical layer when the second node increases its traffic 125

7.1 End-to-end transparent lightpaths and hop-by-hop opaque lightpaths. . . . 129
7.2 Implementation Reference Model for Routers 133
7.3 DiffServ components 134
7.4 Ingress and Egress nodes (DiffServ) 134
7.5 FAN Admission Control Flow Diagrams. 136
7.6 Admission Control regions. 136
7.7 Priority Fair Queueing architecture. 137
7.8 Multi-layer Flow-Aware Networking (MFAN) node architecture. 139
7.9 MFAN Admission Control Flow Diagram. 140
7.10 Optimal decision of the flows. 141
7.11 MFAN Simulation Scenario. 142
7.12 Blocks Definition MFAN Problem 143
7.13 Average goodput and delay in the optical queue with and without admission control. 144
7.14 Fair Rate and Priority Load evolution in a TCP-dominated scenario (Newest-flow policy and $OQ_{th} = 80\%$). 145
7.15 Fair Rate and Priority Load evolution in a UDP-dominated scenario (Newest-flow policy and $OQ_{th} = 80\%$). 145
7.16 Total number of UDP and TCP flows switched in the optical layer (over 100 seconds). TCP-dominated scenario. 146
7.17 Mean delay of the UDP packets in the optical queue (Confidence intervals=95%). TCP-dominated scenario. 147
7.18 TCP flows goodput in the optical queue (Confidence intervals=95%). TCP-dominated scenario. 148
7.19 Total number of UDP and TCP flows switched in the optical layer (over 100 seconds). UDP-dominated scenario. 148

7.20 Mean delay of the UDP packets in the optical queue (Confidence intervals=95%). UDP-dominated scenario. 149

7.21 TCP flows goodput in the optical queue (Confidence intervals=95%). UDP-dominated scenario. 150

7.22 Policies performance evaluation when the percentage of TCP varies. ($OQ_{th} = 70\%$) 151

List of Tables

2.1 Comparison between the integrated and separated control planes 15

3.1 Estimated Parameters Bellcore trace 38

4.1 Optimal decisions with the variation of the R_{cost} parameter 69
4.2 Optimal decisions with the Hurst variation 71

5.1 Path length with the variation of the R_{cost} parameter 88
5.2 Path length with the variation of the T_{max} parameter 91

6.1 Mean delay estimation 110
6.2 M/D/1 ($\rho = 10\%$) Percentile estimation 112
6.3 G/G/1 ($\rho = 50\%$) Percentile estimation 112
6.4 Comparison of each algorithm's size, complexity and accuracy 119

Acronyms

AFL Active Flow List.

AMP Active Measurement Point.

AS Autonomous System.

ATM Asynchronous Transfer Mode.

BER Bit Error Rate.

CoS Class of Service.

CR-LDP Constraint-based Label Distribution Protocol.

CSR Concatenated Shortest path Routing.

DFT Discrete Fourier Transform.

Diffserv Differentiated Services.

DS Differentiated Service.

ECMP Equal Cost Multi-Path.

FBM Fractional Brownian Motion.

FFT Fast Fourier Transform.

FR Fair Rate.

FR Frame Relay.

GMPLS Generalised Multi-Protocol Label Switching.

GPS Global Positioning System.

IETF Internet Engineering Task Force.

Intserv Integrated Services.

IP Internet Protocol.

IPDVD IP Delay Variation Distribution.

IPPM IP Performance Metrics.

IPTD IP Packet Transfer Delay.

IPTV Internet Protocol Television.

IS-IS Intermediate System to Intermediate System.

LMP Link Management Protocol.

LSP Label Switched Path.

MP Measurement Point.

MTE Multi-layer Traffic Engineering.

NIC Network Interface Cards.

NNI Network-Network Interface.

NTP Network Time Protocol.

OBS Optical Burst Switching.

OCC Optical Connection Controllers.

OCS Optical Circuit Switching.

OPEX Operational Expenditure.

OSPF-TE Open Shortest Path First-Traffic Engineering.

OWD One-Way Delay.

OWDD One-Way Delay Distribution.

P2P Peer-to-Peer.

PCC Path Computation Client.

PCE Path Computation Element.

PCECP PCE Communication Protocol.

PDRR Priority Deficit Round Robin.

PFL Protected Flow List.

PFQ Priority Fair Queue.

PHB Per-Hop Behaviour.

PIFO Push In First Out.

PL Priority Load.

PMP Passive Measurement Point.

PSTN Public Switch Telephony Network.

PTP Precision Time Protocol.

QoS Quality of Service.

RIF Reference Implementation Framework.

ROADM Reconfigurable Optical Add Drop Multiplexers.

ROI Return of Investment.

RSVP Resource Reservation Protocol.

RSVP-TE Resource Reservation Protocol-Traffic Engineering.

S-GMPLS Segmented GMPLS.

SFQ Start Fair Queue.

SLA Service Level Agreements.

SLS Service Level Specification.

SP Shortest Path.

STE Single-layer TE.

TDM Time Division Multiplexing.

VoIP Voice over IP.

WDM Wavelength Division Multiplexing.

WSS Wavelength Selective Switches.

Chapter 1

Introduction

This chapter provides an overview of this Ph.D. thesis, its objectives and its structure.

1.1 Motivation

Current IP backbone networks are migrating to an IP over WDM model, where the service provisioning not only depends on the independent behavior of each layer, but also on their influence. This situation has boosted the research on the integration of both layers, enhancing the information exchange to provide the required service to current applications.

The research community has done a great effort to define a common control plane for the IP and optical layers. Such common signaling protocols allow the dynamic reconfiguration of the optical layer, when the traffic conditions vary or when there is a failure in the network. Such automatic light-path provisioning ease the integration of Traffic Engineering (TE) algorithms in multi-layer and multi-domain networks.

Quality of Service (QoS) provisioning is mandatory when disparate applications, such as Grid, Peer-to-Peer (P2P), real-time or Voice over IP (VoIP), share the same network infrastructure. Keeping delay at low levels in such applications is mandatory, but the impact of such delay-sensitive applications in optical network is still an open research topic.

These situations motivate the research on multi-layer and multi-domain networks in order to provide end-to-end Quality of Service (QoS) to end users.

1.2 Objectives

The aim of this thesis is to identify and provide mechanisms to offer Quality of Service in multi-layer and multi-domain backbone networks. The following specific objectives are defined:

1. **Characterization of the main traffic sources in current IP over WDM networks.** The traffic characterization is a required research field, which helps us to find a model to simulate the traffic behavior and to propose mechanisms to offer Quality of Service.

2. **Definition of techniques to improve the multi-layer integration and offer Quality of Service.** Since the IP and the optical layers will coexist in the future, the conflicts between both layers must be solved in order to provide Quality of Service to the end users.

3. **Identify the monitoring parameters and techniques to operate in multi-domain networks providing Quality of Service**. The aim of this objective is to extend the proposed mechanisms to multi-domain scenarios.

4. **Integration of the proposed solutions with the current protocols and architectures.** This objective tries to increase the impact of the contributions of this thesis. The mechanisms proposed in this work may be integrated with the current standardized protocols.

The reader may notice that these specific objectives are treated throughout this Ph.D. thesis and their evaluation is reported at the end of this document.

1.3 Thesis structure

This document describes three solutions to provide Quality of Service (QoS) in multi-layer networks. Chapter 2 demonstrates the feasibility of Multi-layer Traffic Engineering (MTE) strategies in real network operators. An overview of the main strengths and drawbacks of the integration of IP and WDM are discussed along this chapter. Moreover, a case study is carried out proposing a Multi-layer Traffic Engineering (MTE) solution, which avoids congestion at the IP layer. Let us denote this solution by “Threshold-based” algorithm.

This chapter evaluates the economic impact of such a “Threshold-based” algorithm, showing that the optical and electronic resources can be traded-off in other to find an algorithm that not only provides QoS, but also it does not extensively request optical connections.

Such conclusion motivates the main MTE solution of this thesis: the “Bayesian decisor”. The customers request for a determined service, which is usually provided using the IP layer. The IP layer utilization lets the network operator supply multiple user’s demands thanks to the statistical multiplexing. However, the IP layer adds some delay to the packets, thus jeopardizing the proper service delivery. On the other hand, optical resources offer an optimal service in terms of delay, but the chaotic establishment of light-paths yields to their sub-optimal utilization. Chapter 3 defines the problem of transmitting the incoming Label Switched Paths (LSP) into the IP or optical domain in terms of this Bayesian decision theory for a single-node. Chapter 4 extends this model to end-to-end paths, redefining some terms of the Bayesian decisor for such a multi-hop scenario. Chapter 5 further evaluates the behavior of the algorithm in real backbone topologies.

In all cases the Bayesian decisor is evaluated in single-domain environments. Chapter 6 analyzes current protocols to operate in multi-domain scenarios and validates some end-to-end delay monitoring algorithms for intra-domain scenarios. Once the operation protocols and the monitoring are shown feasible, chapter 6 extends the Bayesian decisor to such a multi-domain scenario.

Chapter 7 defines the third MTE proposal of this thesis “Multi-layer Flow-Aware Networking”. The Bayesian decisor solution operates with LSPs, which are requested to the operator and they are allocated using the optimal resources. Flow-Aware Networking (FAN) is a QoS architecture, which allows the implicit traffic classification in elastic and streaming flows. Such implicit classification allows the Internet Service Provider (ISP) to detect the type of service without any packet marking. Chapter 7 extends Flow-Aware Networking (FAN) to operate in multi-layer networks. Thanks to this integration, it is possible to ask for extra optical resources when the IP layer can not deal with the incoming traffic demands.

Chapter 8 concludes this thesis with the evaluation of the objectives proposed in this chapter and the proposal of future research lines.

Chapter 2

Feasibility of the Multi-layer integration in network operators

In this chapter, the dissertation is introduced and motivated not only from an academic point of view, but also as a real problem for the network operators. Traditionally, the network operators have traditionally divided its backbone network in two, the IP network and the transport network. Network planning and engineering tasks are performed independently in both domains. This chapter aims to describe the motivation behind Multi-layer Traffic Engineering (MTE), demonstrate its feasibility and quantify its advantages in terms of cost effectiveness. For such purpose, this chapter compares the soltuions to implement the information exchange between the IP and the WDM layer and it justifies, by means of a case study, how the addition of MTE algorithms has significant benefits for global networking.

The remaining of the chapter is organized as follows: section 2.1 presents the evolution of the backbone networks, how and why they are migrating its legacy multi-protocol networks to IP over Wavelength Division Multiplexing (WDM) networks. Section 2.2 defines the architecture of Next Generation Transport Networks and it considers several solutions to control and manage such multi-layer IP over WDM architecture. Section 2.3 defines the Multi-layer Traffic Engineering problem, whereafter section 2.4 gives a brief overview of the proposed Multi-layer Traffic Engineering solutions by the scientific community. A case study is discussed in section 2.5 showing the trade-off problem of resources utilization and network performance. Finally, section 2.6 summarizes the main ideas of the MTE problems.

2.1 Evolution of IP and transport networks

Synchronous Digital Hierarchy (SDH) is the standard used for fiber-optic transmission systems in current backbone networks. SDH is an electronic circuit switched technology that uses Time Division Multiplexing (TDM). The main advantages of SDH with respect to previous technologies are: (1) interfaces with optical fiber; (2) transmission rates up to 40 Gbps; (3) operational support; (4) fast restoration (50 ms); and (5) grooming of multiple technologies [2]. Originally, the network operators used Frame Relay (FR) and the Asynchronous Transfer Mode (ATM) technologies for offering services to companies, but these two are being replaced by IP at the time of writing. The Synchronous Optical Network (SONET)/SDH technology has its origins in the Public Switch Telephony Network (PSTN), where reliability was a crucial issue. However, SONET/SDH adds some problems in data transmission: Shortest-Past algorithm are not always used; circuit reservation is static; and protection mechanisms force the operator to reserve fiber for protection. Figure 2.1 depicts the legacy multi-protocol stack for backbone networks and evolutionary solutions of such stack. This multi-protocol stack allows the division of functionalities: IP layer acts as the universal communication protocol, which permits the interconnection of any kind of equipment worldwide. ATM protocol is used as an access technology to aggregate end-user traffic coming from DSL connections. Finally, SDH performs the rest of important tasks such as signal monitoring, provisioning, grooming and restoration.

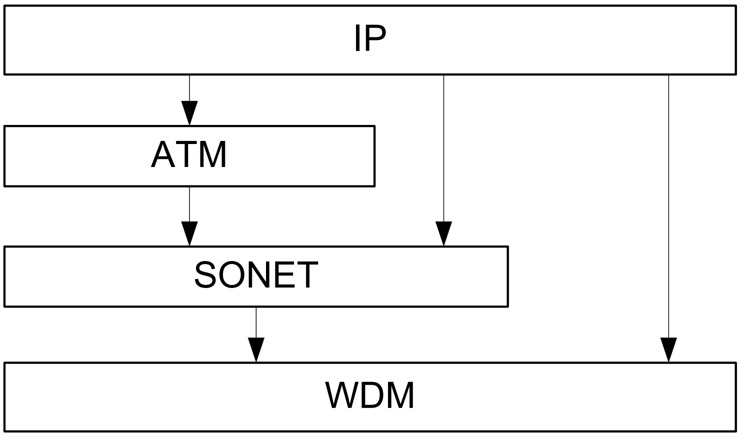


Figure 2.1: Backbone Technologies Evolution

The main drawback of this multi-protocol stack is the functional overlap between layers. When a layer detects a failure, all layers activate their restoration mechanisms leading to incoherent state in some cases. The addition of each layer header, not only includes non-negligible delay but also degrades the network utilization. Moreover, the cost of maintaining four technologies is huge, since every layer requires the specific maintenance of each its equipments. Finally, this architecture presents scalability problems, mainly due to the increase of cut-through traffic, which needs to be processed at the IP layer.

Wavelength Division Multiplexing (WDM) appears as a promising approach to use the vast amount of fiber bandwidth, since it multiplies the fiber transmission capacity. According to [3]: "A single strand of fiber offers a total bandwidth of 25 000 GHz". Such amount of bandwidth is incredibly higher than the potential bandwidth of any other transmission media. WDM transmission have been studied since 1977 [4]. Latest optical transmission testbeds show that it is feasible to reach 25.5 Tbps of bandwidth capacity [5]. First optical equipments were static and they required manual intervention to change its configuration. SDH equipment lets the operator to establish or release connection between its nodes dynamically. Consequently, a great research effort has been done in optical networking to make them reconfigurable by the companies and the scientific community. This effort has led to the creation of Reconfigurable Optical Add Drop Multiplexers (ROADM), which not only are capable of extracting or inserting lambdas but also changing the extracted or inserted lambda. For further reading about this topics, the authors in [6] describe the main optical switching technologies and [3] explains the principal optical network architectures.

Once the optical equipment is able to dynamically change its configuration, protocols are needed to automatically perform this task. This automatic provision of light-paths is possible thanks to Generalised Multi-Protocol Label Switching (GMPLS) [7] and the Automatic Switched Optical Network (ASON) [8]. Section 2.2 introduces some of these GMPLS and ASON concepts.

At the beginning of this section, five advantages of SDH technologies were outlined. However, with the evolution in optical equipment, functionalities (1) "interfaces with optical fiber" and (2) "transmission of 40 Gbps" are fulfilled just within the optical equipment. Moreover, ASON includes in its framework the management plane (Section 2.2) , whose functionalities overlap the (3) "operational support" provided by SDH. Although the protection of IP traffic is not as crucial as that of the telephone lines, Fast Reroute [9] mechanisms allow IP traffic (4) "fast restoration (times 50 ms)". Finally, (5) "grooming functionality" is directly done by the IP layer, merging the flows belonging to different sources and/or destinations. Moreover, IP over WDM is a cost-effective solution compared with SDH technologies [10]. According to these ideas, the transmission of IP traffic reliably over a fiber optical network, also called "IP over WDM network", has a great potential to become the architecture of the future backbone networks.

2.2 Multi-layer networking in the Next Generation Transport Networks

Next Generation Transport Networks (NGTN) comprise a control, a management and a data plane [11]. The data plane is used for the transmission of information packets, while the management plane deals with global operations, including accounting, security evaluation, monitoring reports, etc. The control plane is in charge of the decentralized management issues such as the exchange of routing information, monitoring of link state and the set up and tear down of connections. Additionally, such control plane manages the Service Level Agreements (SLA) and monitors the QoS offered to connections. Figure 2.2 shows the network reference that summarizes the main functionalities of each plane.

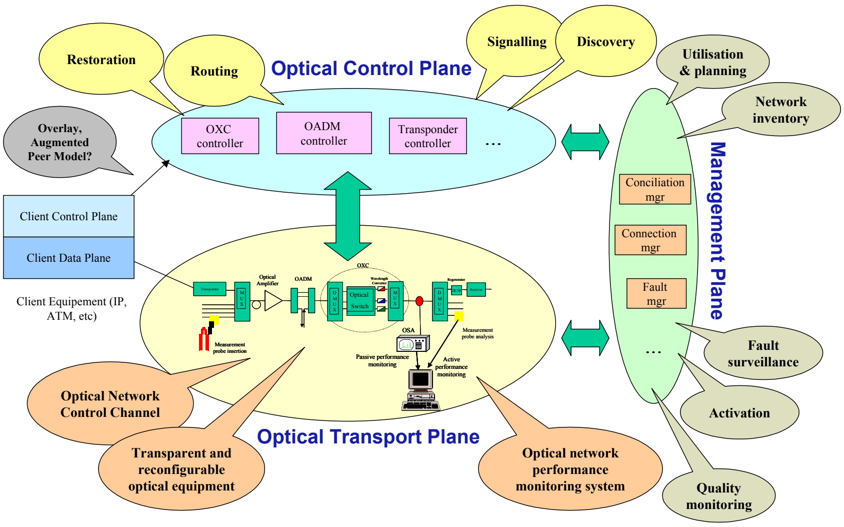


Figure 2.2: Next Generation Transport Network network reference

The International Telecommunication Union (ITU) has defined the Automatic Switched Optical Network (ASON) [8] a standard for the management of NGTN. Such standard is based on the definition of an architecture and its requirements to deal with optical networks. Figure 2.3 shows the main aspects of the ASON architecture. ASON does not define how the function should be implemented, but just the interface between the equipments. From the ASON's point of view, there are four kinds of entities: optical nodes, Optical Connection Controllers (OCC), Element or Network Management (EM/NM) and client equipment. The equipments in the data plane are the optical nodes, which are control by OCC entities which are located in the control plane. The EM/NM entities lie in the management plane. The client equipment is connected to the optical network via the control and the data planes.

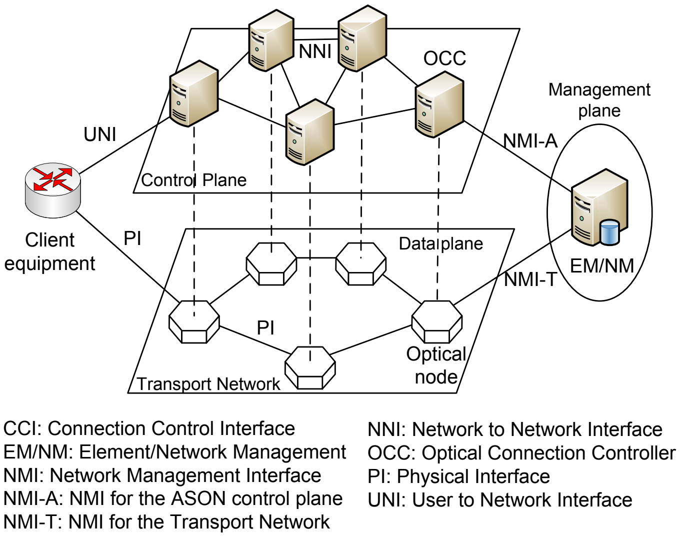

NNI
OCC
Management plane
UNI
NMI-A
Control Plane
Client equipment
PI
Data plane
EM/NM
NMI-T
PI
Optical node
Transport Network
CCI: Connection Control Interface
EM/NM: Element/Network Management
NMI: Network Management Interface
NMI-A: NMI for the ASON control plane
NMI-T: NMI for the Transport Network
NNI: Network to Network Interface
OCC: Optical Connection Controller
PI: Physical Interface
UNI: User to Network Interface

Figure 2.3: ASON architecture

Once the entities are defined, the interfaces between them are also defined. Optical nodes are connected through Physical Interfaces (PI). OCCs are capable of managing the optical node through the Connection Control Interface (CCI). If OCCs want to exchange information among them, they have to use the Network-Network Interface (NNI). Note that this interface exchanges only control plane information. The interconnection between the control and the management plane entities is done through the Network Management Interface for the ASON Control Plane (NMI-A), while the connection with the data plane uses a Network Management Interface for the Transport Network (NMI-T). Finally, the User Network Interface (UNI) is the interface between the client equipment and the optical network.

Figure 2.2 shows the connections between the IP and optical layers. They are both connected through the control plane to share control information and, obviously, through the data plane to transmit information. A transponder connects the IP and the optical layer in the data plane, where the information of the IP traffic in the electronic domain is converted to an optical signal [3]. There is a proposal [2] to fully integrate the optical and the IP equipment on the same box. However, there are advantages and drawbacks of such a full integration of the optical and electronic equipment, which are discussed throughout the next section.

2.2.1 Control Plane Interconnection

In the previous section ASON has introduced. Some of the interfaces that ASON defines are related with the control plane. However, while the ITU defines ASON, the Internet Engineering Task Force (IETF) has defined the Generalised Multi-Protocol Label Switching (GMPLS) [7]. This parallel standardization process is called “The battle of the optical control plane” [12]. The authors in [12] clearly explain the work carried out in the standardization process, a comparison between the proposals and its feasible implementation in real networks. Next, we briefly outline the general ideas of GMPLS. The authors in [13] provide more information about recent optical networking standardization activities for optical networks, including MPLS.

GMPLS extends the ideas of Multi-Protocol Label Switching (MPLS) [14] to the optical domain, which allows the utilization of existing protocols for IP layers in the optical domain. GMPLS requires an extension to the protocols, which leads to new protocol standards. GMPLS is not a protocol but a framework that includes the following protocols:

- **Two signaling protocols**, which allow the reservation of resources for service provisioning: Resource Reservation Protocol-Traffic Engineering (RSVP-TE) [15] and Constraint-based Label Distribution Protocol (CR-LDP) [16].
- **Two internal routing protocols**: Open Shortest Path First-Traffic Engineering (OSPF-TE) [17] and Intermediate System to Intermediate System (IS-IS) [18].
- **A discovering and monitoring protocol**: Link Management Protocol (LMP) [19].

The GMPLS proposal is carried out by IETF, looking for “de facto” standards which fit within the required functionalities of optical networks. In light of this, the last step is to provide inter-layer flexibility by adding a control layer between the IP router and the optical switch, leading to multi-layer capable routers [20], where, on demand, traffic can be sent either over the IP or the optical layer (through an already existing light-path or a newly created one). Note that this approach is evolutionary since the IP routers are already deployed in the network, and new demands are absorbed by the optical layer. This is a key issue for current network operators which have already deployed IP equipment and they strive for Return of Investment (ROI).

It is worth noticing that once this new light-path is established, a new connection is discovered at the IP layer, updating the information of its own routing protocols. On the other hand, optical and IP equipment could share routing information so that they do not have to run the same protocol twice. This issue is related with the integration (or not) of the control plane and it is treated in the next pages.

Integrated Control Plane

Full integration of the control plane is the only solution to optimize global resources usage and to reduce network management complexity, since this model takes into account the information available at all layers. The definition of an integrated control plane is done by the IETF in [21], where the peer model from the GMPLS architecture is proposed. In the peer model each node has a unique vision of the whole multi-layer network. There is only one control plane for all equipments in the domain, so the information is shared among all them (see Figure 2.4). The optical and IP/MPLS equipment are interconnected at the control plane level by means of signaling and routing adjacencies, existing only one integrated instance of routing and signaling protocols. The implementation of the control plane using this model offers two options:

- Both layers are integrated in a single node, with a control module running a single instance of the common control plane. In this case, the existence of a network with dual-layer integrated technologies is supposed, and generally from a single vendor.

- There are different network elements with different switching capabilities and different control modules, but participating in the same control plane instance, and therefore with peering relationships to each other. In this case, the support for different vendors at each layer would depend on the full support of the standards by each of them.

A unified control plane has also several drawbacks, the main one is that its computation requirements increase. Since a common control plane has to control more elements than a single-layer control plane, this implies bigger databases and a less scalable solution in general. Second, route calculation algorithms and traffic engineering mechanisms are significantly more complex, because they have to take into account not only the information at the IP layer but also the optical layer. This way, complexity of integrated routing calculation increases quadratically with the network size, which can cause a slower route calculation as the network grows.

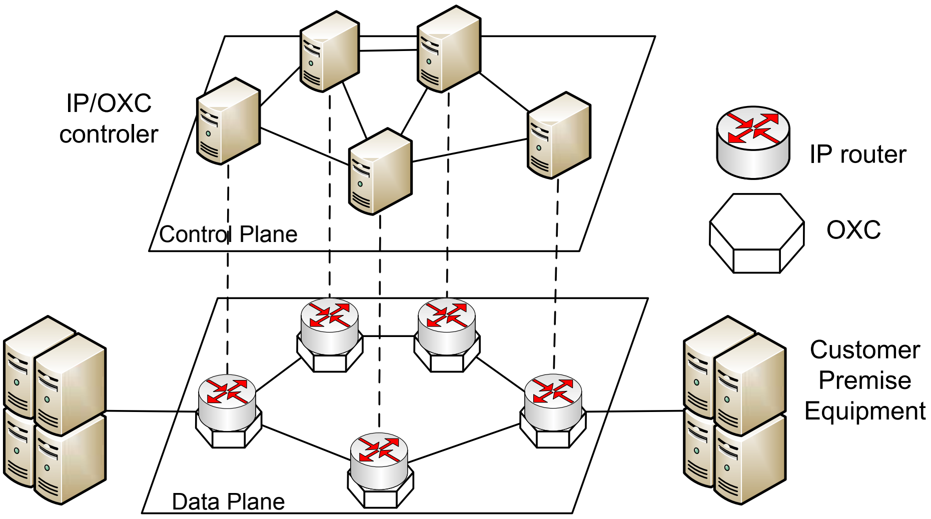


Figure 2.4: Control architecture with an integrated model

If the routing computation time is in the order of seconds, a normal operation scenario can tolerate such delays for connection establishment. However, this time is crucial in a scenario with failures. A single failure can trigger restoration mechanisms of a large number of connections, which may complicate the on-the-fly restoration in the integrated control plane model. This problem could be alleviated with pre-calculated backup paths, but the complexity of the system is increased along with the cost of the equipment if we are looking for an efficient and scalable solution. This issue makes become necessary to evaluate the maintenance cost of a complex control plane against two simplified control planes.

To sum up, the peer model achieves a more efficient multi-layer operation and a better resource usage with less operational effort. The coordinated action among the different layers is guaranteed, especially in the case of failure. In short, the integrated control plane is more powerful although it requires a very careful design and operation to avoid higher computation times in its routing and recovery algorithms.

Separated Control Planes

The main drawback of an integrated control plane is the increment of the computation time. Consequently, the main advantage of two separated control planes is that a single layer control plane is lightweight and thus simpler than an integrated control plane. Moreover, maintaining separated control instances allows a faster control plane with better performance in all its activities. The implementation of a simpler algorithm with less nodes and restrictions allows a more scalable solution. However, since this solution does not use global network information path computation may lead to sub-optimal routing from a multi-layer traffic engineering point of view.

The photonic network would implement its own control plane, including traffic engineering and distributed restoration. This model fits with the proposal of the IETF's definition of an "Overlay Model" [21], where there are two separated control instances one for each layer. (see Figure 2.5). Both domains have different switching capabilities and they can implement different protocols and algorithms. The interconnection between the control planes is done via the UNI interface [22].

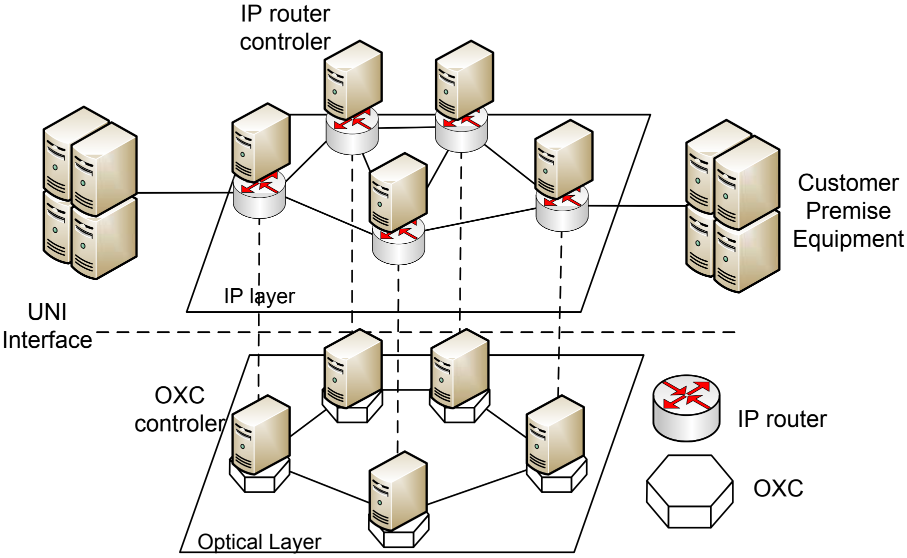


Figure 2.5: Control architecture with separated control planes

In light of this an external manager is required to further carry out multi-layer traffic engineering. Such external manager is in charge of collecting the information from both layers, evaluating it and signaling both layers with the most appropriate configuration. This interaction can be seen as the management of the optical resources to provide a virtual topology to the upper layer. This centralized solution reduces the instability problem, but it does not scale neither it is robust.

The main drawback of the multi-protocol stack (Section 2.1) is that some functionalities overlap among the layers. The same situation appears here when both layers do not share state information. Global resource optimization is complicated, since the network operation and management can be performed by different business or organizational units. When coordination is not properly done, instabilities of the system may appear. This problem is enhanced when there is a failure recovery process. IP and optical layer can trigger their own mechanisms to restore connections carrying out contradictory actions or reducing the efficient use of the resources. When the overlay routing process is done the information is separated, so a greater blocking probability is achieved. Resource optimization occurs just at the client layer, which adapts to the virtual configuration provided by the server layer. On the other hand the server layer has to adapt to the client's requirements. If the client layer knows that the server layer can change its configuration, it may ask for changes on-demand trying to adapt the configuration of both domains. This interaction benefits from new traffic situations due to network planning, congestion, failures, etc.

To put this all in a nutshell, the two main alternatives to implement this coordination are:

- **Joint management system:** both layers remains separate, but there is a joint entity in the management plane, which is in charge of implementing multi-layer traffic engineering algorithms. If both layers do not share the management plane, an external multi-layer traffic engineering agent, which is capable of triggering and indirectly coordinating both control planes, should be placed in the network.

- **Virtual topology on demand:** The network elements from different layers are co-located and they interface through a UNI interface with a "weak" interaction. The multi-layer traffic engineering capabilities are limited, but on-demand requests allow the client layer to change the virtual topology.

Intermediate Solutions: Augmented or Segmented Control Planes

The augmented model is a combination of the peer and overlay models [21]. Some of the edge equipment are peers in the same domain, while other nodes just use the UNI interface to interact with the service layer. The amount of information is shared following an agreement between both layers. This model has appeared as a compromise solution between the first two models.

There is a fourth option proposed by Cisco [23] which is known as Segmented GMPLS (S-GMPLS). The integrated model does not separate the administrative domains of the IP and the optical layer. This is a problem in the deployment of multi-vendor solutions, since this requires not only the interoperability of the functionalities, but also many other aspects. Moreover, the peer model assumes that the transport nodes implement the whole

GMPLS protocol stack, which would not be mandatory if network operator uses a centralized routing algorithm.

To cope with these limitations of the integrated model, Cisco proposes an integration of control based on a stepped migration. The goal is to get some benefits without having to implement the whole GMPLS model. Figure 2.6 compares the peer model and the S-GMPLS, showing the reach of the optical domain control in each case. In the S-GMPLS model, only border routers receive information from optical devices and other routers. The rest of the optical devices or IP routers only have information of the topology. Cisco's proposal is similar to the augmented model solution.

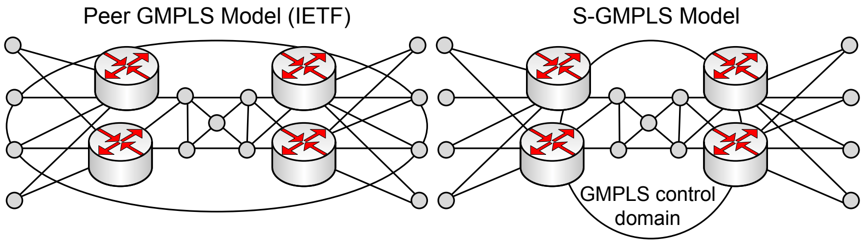


Figure 2.6: Comparison between the Peer and the Segmented GMPLS models as a function of the control plane area of influence

Conclusions concerning the control plane

Once the main approaches for the control plane has been introduced, it is easy to see that the decision of choosing one control plane depends on the network operator infrastructure and goals. Table 2.1 summarizes the main advantages and drawbacks of both the integrated and separated control planes.

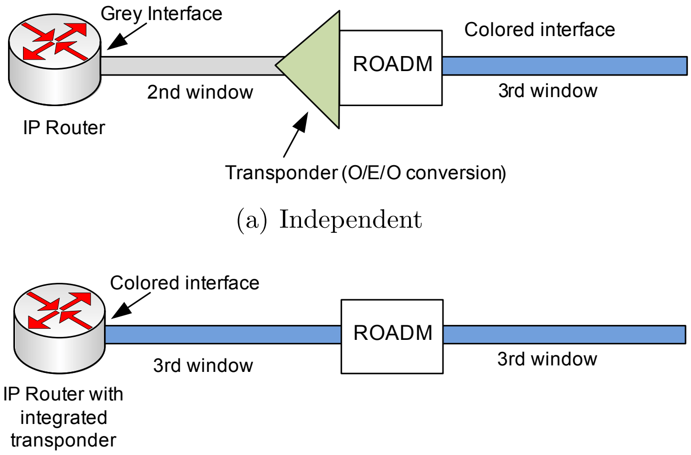
Control plane	**Integrated**	**Separated**
Multi-layer traffic engineering	Complete	Limited
Complexity	High	Low
Performance	Relatively slow	Fast
Stability	Average	High
Architecture	Distributed	Either
Restoration	Integrated	Possible instability
Online restoration	Slow	Fast

Table 2.1: Comparison between the integrated and separated control planes

2.2.2 Data Plane Integration

Multi-layer networking implies not only the integration of the both layers' control planes, but also the cooperation of the equipment in the data plane. In this light, traditional IP vendors, such as Cisco [2] or Juniper, defend the integration of DWDM transponders in IP routers. This equipment unification reduces the amount of components in the network and the optical transponders cost.

Usually the interface between any kind of equipment and DWDM equipment is done using a short range laser operating on the second window. This is known also as "grey" light. This kind of interfaces are incorporated to the SDH or Ethernet interfaces that an IP router has. The DWDM equipment is in charge of taking such second window signal, converting it to the electronic domains and transmitting it over the third window. This output light is also known as "colored". This colored light is multiplexed over a fiber link. Figure 2.7(a) depicts this configuration of independent transponders. This connection model uses three interfaces with their cost. Figure 2.7(b) shows the interconnection model if the long distance wavelength is done in the router. The O/E/O conversion is avoided and the number of interfaces is reduced to one, and, consequently the equipment cost is reduced.



(a) Independent

(b) Integrated

Figure 2.7: Transponder configuration

The main problem of such an approach is that it may imply interoperability problems with different IP or optical equipment. There are many vendors which use their proprietary solution for the physical layer, enhancing the properties of the currently standardized protocols. Therefore, when standard DWDM interfaces are used with such proprietary solutions some interoperability problems may appear. To conclude, the integration of transponders in the IP routers can help to improve multi-layer networking, but may also lead to serious interoperability issues.

2.3 Multi-layer Traffic Engineering

"Internet traffic engineering is defined as that aspect of Internet network engineering dealing with the issue of performance evaluation and performance optimization of operational IP networks" [24]. According to this definition, traffic engineering are techniques that help the network operator to obtain a better network utilization with subsequent performance improvement. Usually the target of traffic engineering is to avoid congestion by controlling how the traffic is routed. Figure 2.8 illustrates a schema with the main blocks in a Multi-layer Traffic Engineering (MTE) problem. These include five blocks:

- **Traffic demands:** These are the information that customers need the network transport. There are many features that determine a given traffic demand. A fixed matrix with end-to-end peak rates can be the input to the problem, but more complicated processes may be defined, for instance, with random traffic changes.

- **Network Equipment:** This is all kind of physical devices used in the process of transporting the information, including routers, ROADMs and the cables or fibers that connect them and their software. Operators invest on infrastructure and MTE tools help to this network capacity task, but once the equipment is deployed on the network, MTE algorithms must deal with the resources available only.

- **Objective:** This is a crucial aspect of every engineering problem. It is very important to properly define what aspects the MTE problem must either to solve or improve. Congestion is usually the target of many TE problems but, in multi-layer networks, it is mandatory to define the objective in both layers.

- **MTE module:** this module is a software that is able to get the operator requirements, traffic information and network status, and provide an optimum or feasible solution based on this information.

- **Network configuration:** this is the solution to the previous problem and it is signaled to the network equipment, so it can apply the proper changes.

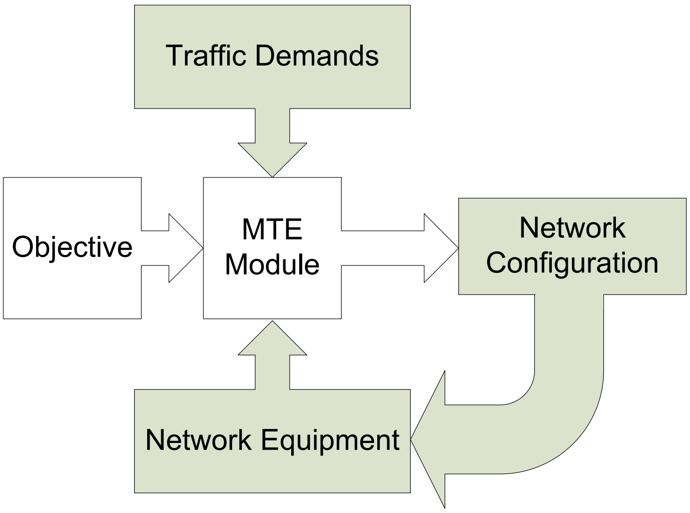


Figure 2.8: Multi-layer Traffic Engineering Problems

The previous definition is also valid for Single-layer TE (STE). Multi-layer TE adds more flexibility to the network, at the expense of increasing the problem's complexity. Multi-layer capable nodes can establish end-to-end connections (light-paths) between any two nodes in the network on-demand. Thus, traffic can be groomed and sent over a direct light-path, either by-passing the intermediate nodes or following the hop-by-hop connections. Such traffic offloading is depicted in Figure 2.9.

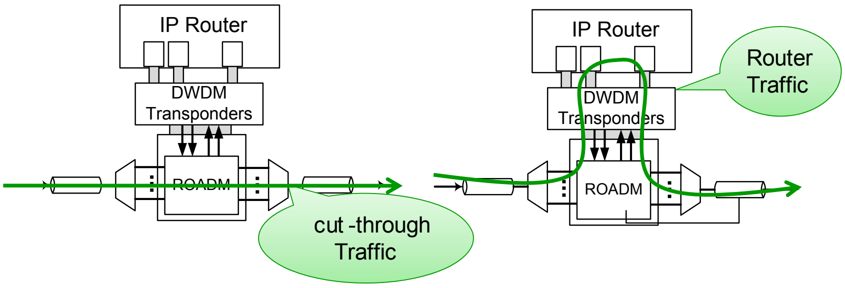


Figure 2.9: Cut-through and hop-by-hop traffic

An important remark concerning traffic offloading in the optical domain is that it requires a reconfigurable optical technology. In the past, the first optical networks were static, such that network managers had to go to the location where the optical equipment was connected and then change the connection. This solution is not suitable for multi-layer networking. At present optical devices can remotely and automatically change its configuration and a control plane is defined, thus the technology has became mature enough to establish by-pass connections that do traffic offloading at the optical layer, as Figure 2.10

illustrates.

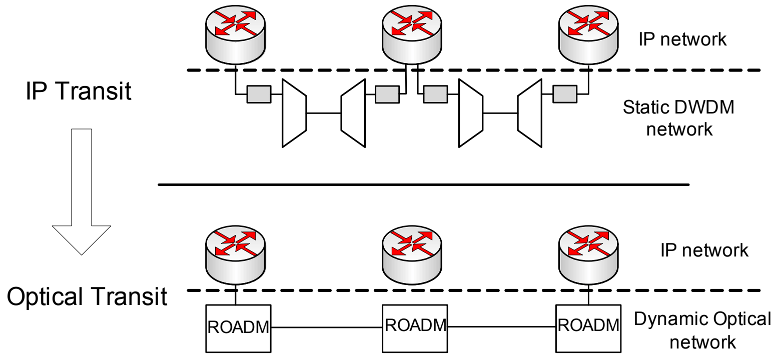


Figure 2.10: Traffic offloading over the optical transport network

While in pure IP networks, all TE mechanisms are exclusively done at the IP layer, in IP over WDM architectures these can be carried out either at the IP/MPLS or optical layer.

According to the previous definition of the MTE problem, there are many types of MTE problems. Every block has its own restrictions or features that makes a given MTE problem different. A taxonomy of Traffic Engineering Systems is done in [24], let us summarize the most important ideas:

- **Time-dependent vs State-dependent vs Event-dependent.** Time-dependent algorithms use historic information of the system in order to apply their policies. State-dependent based algorithms use the state of the network to make the decisions. Usually, this state information can not be obtained with the historical data. Event-dependent algorithms use learning models to find an optimal solution.

- **Offline vs Online.** Offline algorithms are used when there are no real-time requirements, while online algorithms are triggered periodically or when an anomaly occurs.

- **Centralized vs Distributed.** Centralized solutions have all the information in a unique entity, which usually makes it easy to find the optimum solution. However, these do not scale and lack robustness.

- **Local vs Global Information.** A TE is local when it does not require information from the whole network but a portion. Global algorithms needs information from the network, which can be a problem because of the delay in monitoring and measurement systems.

- **Descriptive vs Prescriptive.** Descriptive solutions just evaluate the possible options to carry out, while prescriptive algorithms recommend an option among them.
- **Closed Loop vs Open Loop.** Closed Loop algorithms are those which not only make a decision but also use feed-back information to improve such a decision. Open loop algorithms do not use this feed-back information.
- **Tactical vs Strategic.** Tactical algorithms try to solve a certain problem, while Strategic algorithms make a systematic evaluation of the system to keep it in a proper state in the medium or long term.

Whenever there is a possibility of improving the network performance in a multi-layer network, there is a MTE problem. To solve such a problem, the MTE algorithm requires information concerning traffic demands and available network equipment and resources. Once the problem is clear the MTE module must be designed paying attention to the alternatives shown in the previous classification. This will provide a new network configuration if there is a feasible solution or a recommendation to upgrade the infrastructure and satisfy the objectives.

2.4 Proposed Multi-layer Traffic Engineering Solutions

Multi-layer Traffic Engineering is a topic where the scientific community has paid a lot of attention. A usual term for MTE problems is “grooming”. Grooming is “essentially a network design problem of resource allocation” [25], where there are traffic demands to be allocated in a network topology and the grooming process determines, based on its objectives, the network configuration and traffic allocation. Let us remark that this schema is similar to the one previously presented in Figure 2.8. The first research works done in the area are focused on grooming SONET/SDH connections on ring topologies [26, 27, 28]. Then, this problem was studied in WDM mesh networks with SONET/SDH circuits [29, 30]. This SONET/SDH networks are bidirectional so their demands are symmetric. Moreover, the granularity of such problems is fixed to the standard SDH/SONET circuits capacity. These works are usually focused on defining an optimization problem where an objective function, such as a cost function, is either minimized or maximized. For further reading, reference [25] is a excellent book which covers the problem of traffic grooming.

When the client layer of the optical domain is not a SDH network, but an IP/MPLS network, the traffic demands model (Figure 2.8) changes and so does the problem itself.

For instance, the traffic does not have to be bidirectional but unidirectional. An example of such traffic profile is web-browsing, where a user just sends a request and a whole web-page is downloaded with all its content (images, videos, etc.). This asymmetric nature of the IP traffic yields to dynamic grooming problems [31, 20, 32, 33]. An example of how a model can help in the problem solution is shown in [31, 34]. The authors in [31] assume that the traffic demands are scheduled, so the demand arrival time and the duration for such demands are known in advance. This scheduled traffic model is deterministic because all the demands are known in advance, but, on the other hand, is dynamic because it changes with the evolution of the traffic load in the network over time. In the short term (seconds), the authors in [34] assumed that there is a set of scheduled lightpaths and also random demands. These random demands are scheduled after the fixed demands are allocated.

The authors in [20] apply the idea of by-passing the IP traffic through the optical network based on a monitoring system, which detects which links are exceeding a given minimum capacity threshold (C_{min}). When these links are detected, they are by-passed. This work also deploys a three-node test-bed to illustrate the feasibility of such ideas from the technical point of view. Authors in [35, 36] show the feasibility of such type of MTE algorithms not only with the path establishment but also the reversion to the IP layer when the end-to-end connection is not yet required. However, these mechanisms do not look for a global optimization objective, but just avoid the congestion at the IP layer. The authors in [32] use a cost function that penalizes the low loaded IP links with a medium cost and high loaded links (more than 60% of the link capacity) with a high cost. The links at medium load levels (from 20% to 60%) get a lower cost. Using this function, they look for the global optimal configuration that mimizes the cost, yielding to the optimal set of connections between the IP and the optical layer.

From an architectural point of view, the by-pass decisor can be centralized as proposed in [20, 37] or distributed as pointed out in [38, 39]. A third option to carry out such MTE mechanisms is not to include the MTE entity inside the optical network, but in the edges nodes. The authors in [40] differentiate between edge and core routers, and the information exchange is just done among the edge routers which are, consequently, in charge of making such MTE decisions. This approach uses the optical layer as a connectivity service layer. As previously discussed in Section 2.2.1, the best solution depends on many factors, but the feasibility of the control plane interaction with MTE algorithms is possible.

2.5 Case studies

2.5.1 Definition of basic MTE algorithms

Let us consider a network topology $T = [V; L]$ consisting of a vertex set V and an arc set L (see [41]). It is assumed that each vertex represents a multi-layer capable router. Each arc $(x, y) \in L$ has associated a non-negative real number $l(x, y)$, which refers to its load. Let us define D as the demand matrix, where each value associated to the position (i, j) denotes the traffic demand from node i to node j, $d(i, j)$ $i, j \in V$. The objective of this MTE problem is to by-pass all connections in the optical domain which exceed a given capacity threshold (C_{min}). This threshold can be chosen based on QoS parameters or economic cost, such that if exceeded this means that the cost of the optical offloading is lower than the IP transmission. Figure 2.11 summarizes the ideas of this MTE problem. The output of this problem are the remaining IP load and the by-passes matrix Bp, where $bp(i, j)$ is the amount of traffic offloaded in the optical domain from node i to node j.

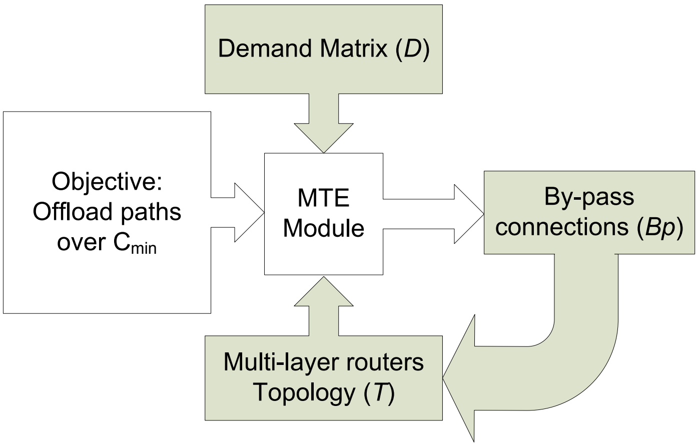

Demand Matrix (*D*)
Objective: Offload paths over C_min
MTE Module
By-pass connections (*Bp*)
Multi-layer routers Topology (*T*)

Figure 2.11: Blocks Definition Offloading Problem

This MTE algorithm first maps the IP traffic over the logical topology provided by the optical layer using a given routing algorithm. Once the load in each link is known (L matrix), the algorithm runs through the links of each node, finding n consecutive links where $l(i, j) > C_{min}$. Let us call this set of consecutive links as candidate paths set CP for optical by-passing. Note that the shared bandwidth among the links $l_1, \ldots, l_n$ is $BCP = min(l_1 - C_{min}, \ldots, l_n - C_{min})$. Once the CPs are found, the algorithm has to decide the order to extract them. We propose two cases: "Single Threshold Largest By-Pass" and "Single Threshold Longest By-Pass".

The “Single Threshold Largest By-Pass” algorithm sorts the candidates path by its bandwidth (BCP) and, if there are paths with the same bandwidth, by length. On the other hand, “Single Threshold Longest By-Pass” algorithm firstly orders them by length and, secondly, by bandwidth. Once the list is sorted, path candidates are extracted.

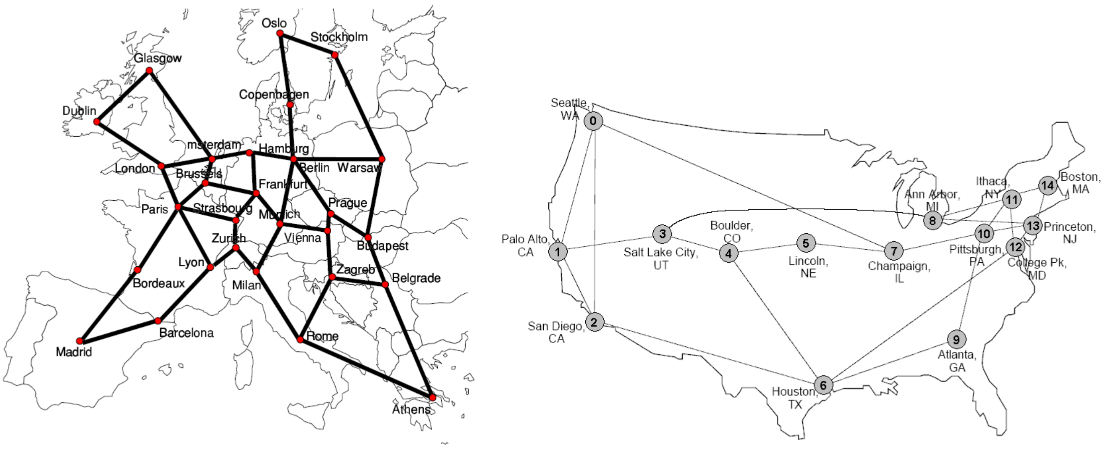
Algorithm 1 Single Threshold Longest By-Pass Algorithm

```
$L \leftarrow Map\_Demand\_WDM\_Layer(D, T, Routing)$
**for all** $V$ *as* $v$ **do**
    $(CP_v, BCP_v) \leftarrow Cand\_Paths(v, C_{min})$
    $CP \leftarrow CP \cup CP_v$
    $BCP \leftarrow BCP \cup BCP_v$
**end for**
**for** $i = maxlength(CP)$ *to* 2 **do**
    **for all** $CP(length == i)$ *as cp* **do**
        **if** $Bcp > C_{min}$ **then**
            $L \leftarrow update\_load(cp)$
            $(CP, BCP) \leftarrow update\_cand\_list(L)$
        **end if**
    **end for**
**end for**
```

This case study compares these MTE algorithms with the IP over WDM solution. When the IP over WDM technology is used, no information exchange is done between the layers, so the IP layer just chooses a routing algorithm and sends its traffic based on this protocol. The other extreme solution is to create a full mesh topology between all nodes in the network. This means that for each end-to-end demand, a lightpath is established between the source and the destination. Consequently, a network with N nodes would require $N(N-1)/2$ ligthpaths, which is 378 ligthpaths for NOBEL’s network 2.12(a) and 105 ligthpaths for the NSFNET 2.12(b). Obviously, this solution is feasible only when the end-to-end network load is too high, but it is out of the scope of this study.

2.5.2 Assumptions

This case study is done using two well-known backbone network topologies: the NOBEL reference network (Figure 2.12(a)) [42] and the NSFNET network (Figure 2.12(b)). The NOBEL reference network has 28 nodes, which are connected by 84 links. This yields an average nodal degree of 3 and the average end-to-end distance in number of hops of 3.4311. On the other hand, the NSFNET network has 15 nodes, 46 links, its average nodal degree is 3.0667 and the average end-to-end distance in number of hops is 2.0889.



(a) NOBEL reference network

(b) NSFNET network

Figure 2.12: Backbone network topologies

In order to study network performance, a random traffic matrix is computed with a uniform distribution $[0, 1]$. Once the matrix distribution is computed, its values are scaled such that the amount of traffic between every end-to-end nodes is multiplied by an increasing factor in order to simulate different traffic loads.

2.5.3 Performance of the proposed algorithms

The objective of this algorithm is to detect those links that exceed a bandwidth threshold (C_{min}) in order to by-pass them using the optical layer. Figure 2.5.3 shows the amount of traffic in each link in a pure IP over WDM architecture. The ECMP routing algorithm is used to send the traffic over the IP layer. Essentially, Equal Cost Multi-Path (ECMP) is the same algorithm as Shortest Path (SP), but instead of providing a single shortest path, ECMP provides a set of paths with minimum distance, thus enabling load balancing at the IP level. Once the traffic is allocated over the IP over WDM topology, one of the previously defined algorithms on section 2.5.1 is applied over such overloaded topology. For instance, the link from 8 to 9 is overloaded in the IP over WDM allocation (Figure 2.5.3), while the MTE solution (Figure 2.5.3) alleviates such congestion situation. It is important to remark that not all IP links are off-loaded through the optical layer. The reason is that if the amount of traffic between two sibling nodes is higher than C_{min} no by-pass can be created. At least there must be two consecutive overloaded links to by-pass any IP router.

Moreover, these two consecutive links must have shared routes in order to define a by-pass traversing both links with a shared amount of traffic higher than C_{min}.

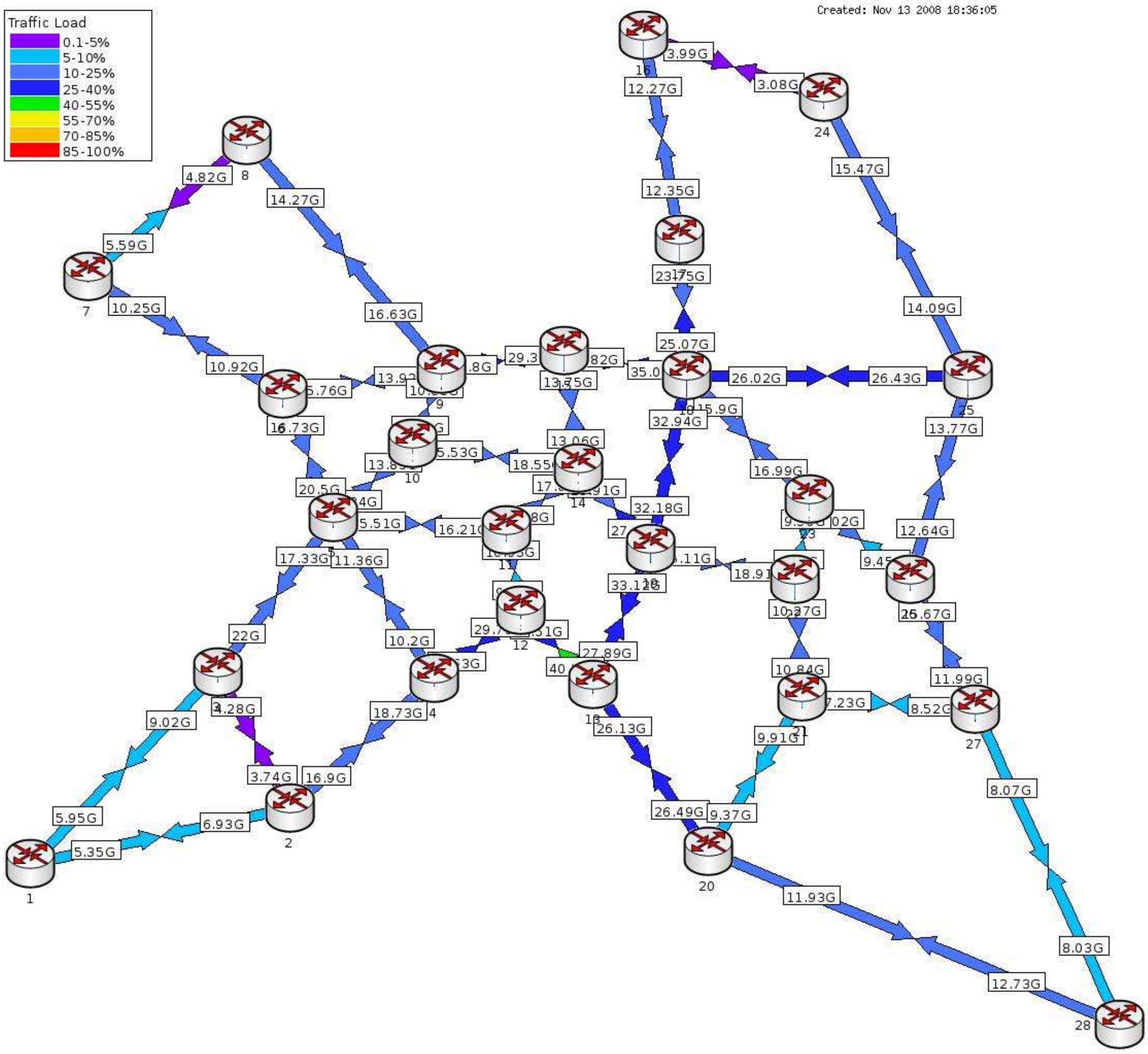


Figure 2.13: Load of the NOBEL reference network with ECMP routing algorithm

Figure 2.15(a) and 2.15(b) illustrates the amount of overloaded links in the IP over WDM architecture and when the MTE algorithms are applied. Let us remark, that not only SP routing, but also ECMP is depicted. The X axis shows an increasing traffic rate until all links of the IP layer are congested. The NSFNET network has 46 links and when there is an average rate between the end-to-end nodes of 1.2 Gbps, the IP layer is completely congested. The number of links of the NOBEL's network is 84 links. Full congestion of the IP over WDM topology is reached when there is an average ratio of 1

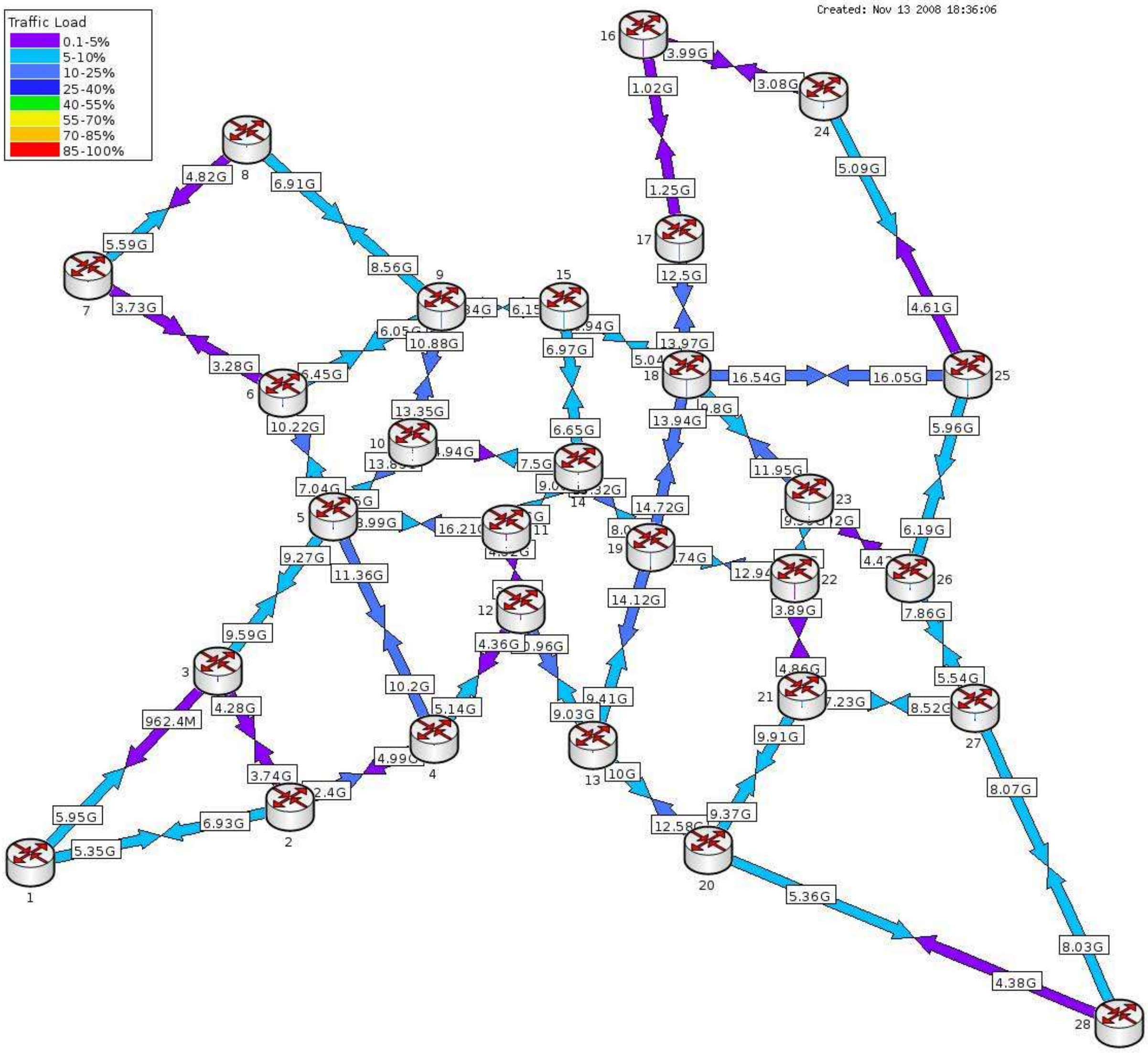
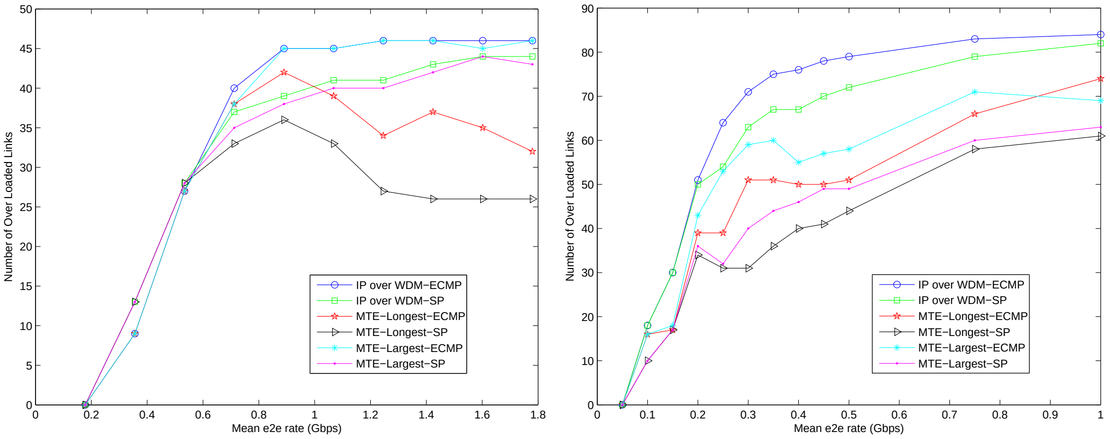


Figure 2.14: Load of the NOBEL reference network with ECMP routing algorithm and MTE algorithm (Longest)

Gbps between destinations. This congestion arrives sooner than in the NSFNET network, because the number of nodes is larger and, consequently, the total traffic in the network. The overload in the IP over WDM topology differs when the routing algorithm changes. ECMP routing should split the traffic among the links and reduce the network congestion, but this has the opposite effect. However, the amount of links congested when the network load increases is similar in both cases.



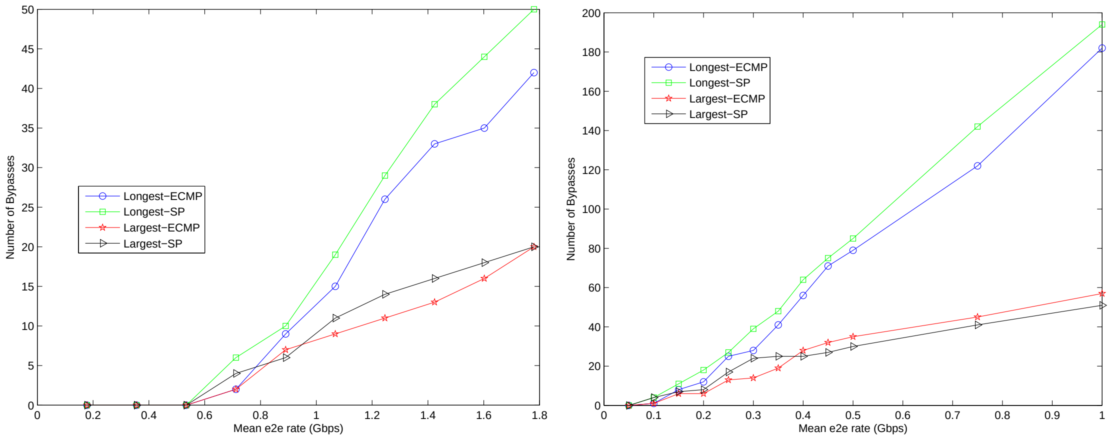
(a) NSFNET Network

(b) NOBEL reference network

Figure 2.15: Number of Overloaded links ($BW > C_{min}$) in each topology

Once the performance of the IP over WDM architecture is outlined, the performance of MTE algorithms needs to be analyzed. The performance improvement of the Largest algorithm in the NSFNET network is negligible. The same number of IP links remain overloaded, with or without this algorithm (see Figure 2.15(a)). This behavior occurs not only with ECMP algorithm, but also with SP. However, its behavior in the NOBEL's network is much better, Largest reduces the overloaded links in a 40% in some cases. The results of the “Longest” algorithm are better than those of the Largest algorithm. It works properly in both topologies with both routing algorithms. Figure 2.16(a) shows the amount of by-passes established for the NSFNET and Figure 2.16(b) depicts the ones for NOBEL's topology. The number of by-passes established with the Longest algorithm is much larger than the number of the Largest algorithm creates. This behavior is related with the algorithm itself. The Longest algorithm established the longest candidates by-passes at the beginning and then it continues with the shorter ones. On the other hand, the Largest algorithm creates the by-passes with the greatest amount of traffic at the beginning, thus, reducing the amount of candidate paths that exceed the minimum bandwidth (C_{min}).

This reaches a more fragmented network configuration, yielding to a more overloaded IP topology, when the average distance in hops is lower. Let us remind that the average distance in number of hops is 2.0889 in the NSFNET and 3.4311 in NOBEL reference network. This is the reason why the performance of the Largest algorithm is better in the NOBEL's network than in the NSFNET network.



(a) NSFNET Network

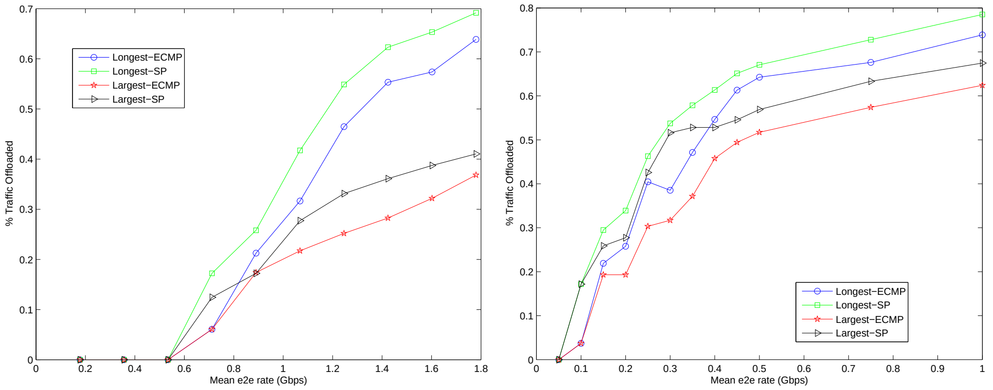
(b) NOBEL reference Network

Figure 2.16: Number of created By-passes

Finally, let us focus on the amount of off-loaded traffic. Figures 2.17(a) and 2.17(b) show the percentage of the traffic that is off-loaded to the optical layer. At low loads, there is no off-loading at all, since the amount of traffic is high enough to be groomed through the end-to-end ligthpaths. As the system becomes overloaded, this percentage of traffic is higher, consequently, reducing the amount of traffic at the IP layer and absorbing the traffic increment just by the WDM layer. The better performance of the "Longest" algorithm is because it is able to send more saturated paths to the optical network, yielding to a higher percentage of extracted traffic in comparison with the Largest algorithm.

2.5.4 Impact on IP Equipment

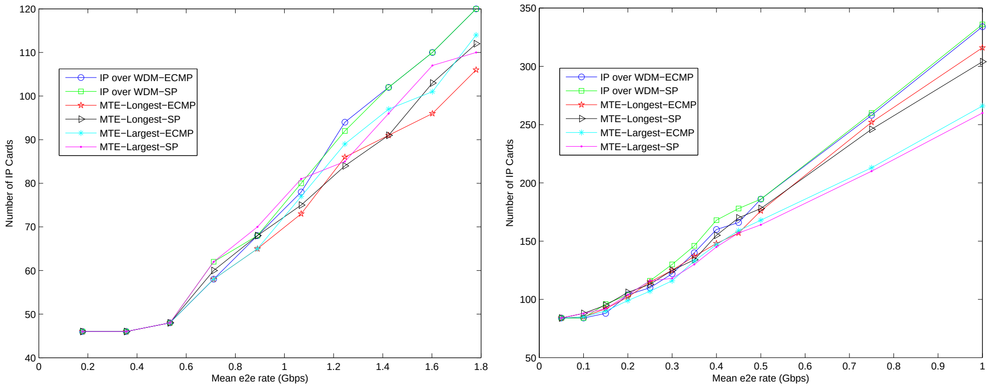
The previous section analyzed the behavior of the MTE algorithms, but this section is focused on the impact of such algorithms on the IP equipment required in each solution. This section just studies the number of IP Network Interface Cards (NIC) because it is the most expensive device in IP over WDM networks. This information is well-known, but cost models such as the one carried out in NOBEL project [42], which is used in [36], or the cost



(a) NSFNET Network (b) NOBEL reference Network

Figure 2.17: Percentage of off-loaded traffic over the total traffic

model to validate the OIS architecture [43, 44] enhance this assumption. Figures 2.18(a) and 2.18(b) show the amount of IP cards required for each solution.



(a) NSFNET Network (b) NOBEL reference Network

Figure 2.18: Number of IP cards required for the IP over WDM solution and MTE algorithms

In light of the results, the MTE algorithms require a similar number of IP cards at low load levels, while they reduce the amount of cards required at high load situations. Such difference is enhanced by the fact that the cost of IP cards is about tens times the cost of other devices. The previous section showed that Largest algorithm does not achieve a great performance in some cases, but its equipment consumption is lower than Longest with the algorithm. This is related to the number of optical by-passes created 2.5.3.

2.6 Multi-layer Traffic Engineering Conclusions

The aim of this chapter was to show that Multi-layer Traffic Engineering algorithms were feasible and they constitute a real improvement for real network operators. The evolution of backbone networks has led to an IP over WDM architecture that allows network operators to reduce the complexity and the cost (CAPEX and OPEX). This new architecture opens new technical problems such as the operational and functional performance of such networks.

Such open issues in IP over WDM networks are outlined throughout this chapter. First, there is an explanation on the evolution of the optical equipment, which permits the dynamic configuration of the optical layers. This functionality brings the need for protocols that perform such an operation automatically. The development of GMPLS and ASON allows this automation, but it opens the problem of the optimal architecture of the control plane. The main configurations for the control plane are presented, including not only the proposals of the standardization organisms, but also some from vendors. There is no close solution for the control plane problem, but this architectural proposal allow us to conclude that from the control plane point of view, MTE mechanism are feasible. The integration of equipment in the data plane is also introduced, but yielding again an open solution.

Finally, the main studies in the MTE field are presented with the addition of a case study, where two MTE algorithms are validated and compared in two well-known network topologies. This case study is just a first step, but it shows that the integration of the information in both layers can help the operator to achieve a better network configuration. The configurations achieved in this study are less congested and require a fewer amount of IP resources. Moreover, no optimization problem is proposed, but “Longest” and Largest algorithms are just heuristic proposals. In light of the results, we can see that a better network performance is inversely proportional to the number of resources used. The Longest algorithm achieves a lower network congestion, while the Largest algorithm uses a less amount of optical by-passes.

Chapter 3

Bayesian decisor of a multi-layer capable router

This chapter defines a set of rules to transmit a Label Switched Path (LSP) either in the IP layer or using a by-pass connection for a single multi-layer capable router. The motivation of this work is presented and the function to construct this Bayesian decisor are explained. We show that this multilayer mechanism trades of the delay perceived by the customers and the utilization of the optical or electronic resources. A set of experiments is carried out to show that the behavior of the decisor fits with the design purposes.

This chapter is organized as follows. Firstly, section 3.1 shows the motivation of the work and the assumptions of the model. Then, section 3.2 covers the mathematical foundations for set of rules with a Bayesian decisor. Section 3.3 provides a set of experiments and numerical examples to show how to reach an optimal decision. Section 3.4 studies the behavior of the Bayesian decisor in a dynamic environment, with its analytical definition and experiments. Finally, section 3.5 outlines a summary of the results obtained and future work.

3.1 Motivation and model assumptions

Core networks are typically equipped with both electronic and optical resources. This means that incoming traffic can be routed in either the optical or electrical domain. Essentially, electronic routing has the well-known advantages of statistical multiplexing and granularity, but is a hard-computational process for high-speed networks and it further introduces queuing delay to packets. On the other hand, data packets switched in the optical domain only experience propagation delay. However, optical resources provide a granularity which is too coarse for typical Internet streams, even if they come from the multiplex of many users. The feasibility of such multi-layer architecture is explained in chapter 2. With the aim of dealing with optical and electronic resources, some router architectures are defined [20, 45]. Figure 3.1 illustrates a multi-layer architecture which is dealing with lambdas and SDH tributaries and it can transmit IP traffic or any other kind of traffic source. Such multi-layer router can by-pass the traffic at SDH, lambda, waveband or fiber levels.

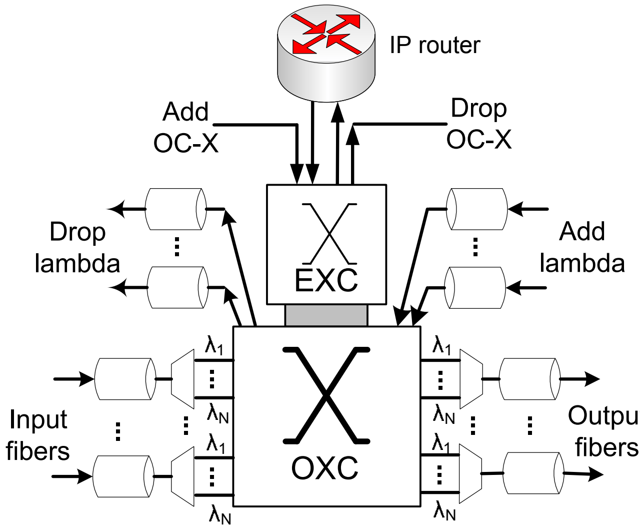


Figure 3.1: Multi-layer capable router scenario

Let us assume that Generalised Multi-Protocol Label Switching (GMPLS) framework is used in this multi-layer router. Figure 3.2 depicts the control plane units in the multi-layer capable router. An exchange of information between both layers os requiered, to be performed by an external entity called “Multi-layer TE Module”, which is located in the management plane, as explained in section 2.2.1. If the network operator uses the peer model, this function of collecting the information is done in the node itself as Figure 3.2 displays. The incoming Label Switched Paths (LSPs) traverse the multi-layer capable router and the “Multi-layer TE Module” has to decide whether to perform optical or electronic switching. If an incoming Label Switched Path (LSP) is routed in the electronic domain, it suffers a queuing delay and a hop-by-hop O/E/O conversion (with subsequent delay), otherwise the router provides an optical by-pass. On the one hand, the availability of buffering in the electronic domain allows for a larger utilization in presence of bursty traffic. In the optical domain it is unfeasible to multiplex several LSPs onto the same wavelength, except in the ingress node, and the optical bandwidth has a negative impact on utilization.

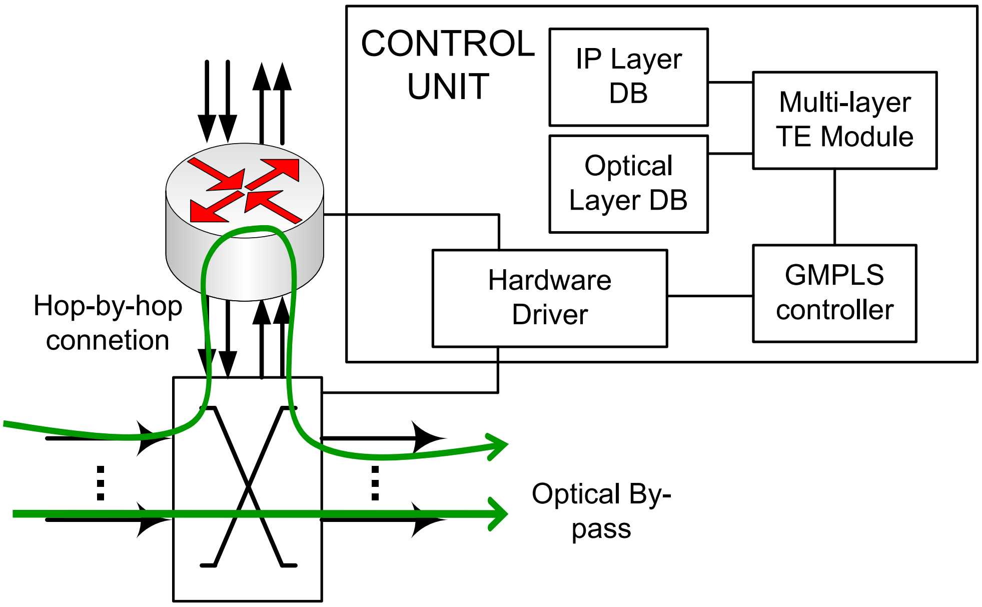


Figure 3.2: Decision make in a multi-layer capable router

Traditionally, network resources utilization and QoS provisioning are studied as separate problems. For instance, the authors in [32] ask for optical connections to the WDM layer using a cost function, which penalized the high and low loaded links. Consequently, the network resource utilization is balanced. In [46], the authors propose an ILP optimization algorithm to minimize the load in the electronic domain using cut-through lightpaths, subject to the network equipment restrictions. Nevertheless, no QoS evaluation, in terms of end-to-end delay, is performed. On the other hand, the authors in [47, 48] propose QoS schemes for IP-over-optical multilayer networks based on the DiffServ concept. They propose to divide traffic into Class of Service (CoS) and map them into end-to-end optical lightpaths or hop-by-hop connections. However, these proposals do not take into account the cost associated to the use of the electronic or the optical resources.

As section 2.6 outlines, resource utilization and network performance are usually opposite targets. The more resources are used (Longest algorithm), the best performance is achieved. A Bayesian decisor formulation trades-off between the resources utilization and the QoS experienced by the traffic flows. The only formulation similar to our approach is found in [49], where an IP over WDM framework is defined but with little insight in real network scenarios.

In the following two aspects of our model are introduced: the utility functions and the traffic model. A utility function measures the relative satisfaction perceived by a customer when a service is provided. Section 3.1.1 introduces the definition of such functions, which help us solve this decision problem. On the other hand, we assume a traffic model for our analysis. Section 3.1.2 defines the traffic model used, its behaviour is studied and an experiment shows that the analytical equation and the simulation fit.

3.1.1 On the use of Utility functions

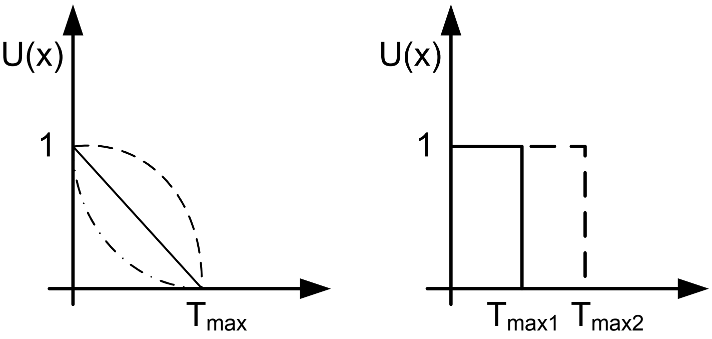
Utility functions are widely used in many fields, from economy to mobile networks, but not in optical networks. Quality of Service (QoS) is somehow related to a utility function. The more performance, the higher utility for users. One of the first definitions of the utility function of Internet Services is performed in [50]. It defines the performance of the network based on "the efficacy or total network utility", called V, which is defined as follows:

$$V \equiv \sum_i U_i(s_i) \tag{3.1}$$

where $U_i(\cdot)$ is the utility function associated to a given service s_i. The most extended service is the Internet, it is the so-called "best-effort" service. Such kind of applications (like web browsing, email, FTP, etc.) do not require strict delay requirements, but if the lower the delay perceived the better the service utility. The same happens when sending an email or transferring a group of files using FTP. However, the average delay is not always a useful (or at least, representative) metric in the evaluation of the Quality of Service experienced by certain applications, especially when quantifying the relative QoS experienced by real-time applications. Let us consider two other utility functions used in the literature for "hard-real time" and "elastic applications" [50, 51].

"Hard real-time" applications are those which tolerate a delay of up to a certain value, say $T_{\max}$, but their performance degrades very significantly when the delay they experience exceed such value (Figure 3.3(b)). Examples of these are: online gaming, back-up services and grid applications. The parameter $T_{\max}$ denotes the tolerated delay threshold for each particular application. The ITU-T recommendation Y.1541 [52] and the 3GPP recommendation S.R0035 [53] define service classes based on thresholds.

Other services consider a more flexible QoS function, since the service is degraded little by little (Figure 3.3(a)). These "elastic" services consider zero delay as the maximum



(a) Elastic (b) Hard real-time

Figure 3.3: Utility functions

possible utility, but the utility slowly reduces with increasing delay. For instance, the ITU-T recommendation G.107 defines the "E model" [54], which explains in detail the voice service degradation as perceived by humans. When the delay is increased in a voice conversation, the utility perceived by the user decays with the delay, but it is useful until a $T_{\max}$ delay is reached. Once this $T_{\max}$ delay is exceeded the conversation becomes inaudible.

3.1.2 Traffic model

Traffic characterization is one of the most important problems and one of the most challenging research issues in computer networks and communications. If the traffic process could be predicted, a lot of open problems would find solution (quality of service, routing optimization, network dimensioning, etc). The state of the art greatly improved after the publication of [55], which found self-similarity and long range dependence in network traces. These features provided a new point of view to study the traffic behaviour.

The authors in [55] analysed Ethernet traces and found the so-called long-range dependence and self-similarity. In this study, the authors analyse traces from the range of seconds to months providing very robust results. These experiments were repeated by other authors (like in [56]), validating the results in [55].

Fractional Brownian Motion Definition

One of the models to study the long-range dependence characteristics is the Fractional Brownian Motion (FBM). Here, this model will be defined to characterize the incoming traffic to a system ([56]). Let A_t denote the number of bits arrived to the system in an interval $[0, t)$. The process A_t is defined for $t \in (-\infty, \infty)$. Let us define the traffic in a time interval $[s, t)$ as $A(s, t) = A_t - A_s$. We assume that the process increments are stationary and that the process is square integrable. The process is called short range dependent if for any $s < t \leq u < v$, the correlation between $A(\alpha s, \alpha t)$ and $A(\alpha u, \alpha v)$ converges to zero, when α approaches to infinity. Otherwise, it is said to be long-range dependent. Traditional models used in telecommunications were short range dependent, but after the results in [55] this assumption changed. Let us define $v(t) = Var(A_t)$, as the variance function of the A_t process. The variance for long-range dependent processes follows the expression:

$$VarA_{\alpha t} = (\alpha t)^p = \alpha^p Var(A_t), \quad t \in (-\infty, \infty) \tag{3.2}$$

where p is a factor strictly between 1 and 2. This implies that $A_{\alpha t}$ and $\alpha^{p/2} A_t$ has the same correlation structure.

A process is strictly self-similar with Hurst parameter (H), if for any $\alpha > 0$, the process Y_t and $\alpha^H Y_t$ have the same finite-dimentional distribution. The simplest second order model is the Gaussian one. Z_t is a normalized FBM, if the following properties are fulfilled [56]:

1. Z_t has stationary increments.
2. $Z_0 = 0$, and $EZ_t = 0$ for all t.
3. $EZ_t^2 = |t|^{2H}$ for all t.
4. Z_t has continuous paths.
5. Z_t is Gaussian.

The fractional Brownian traffic model is defined as follows:

$$A_t = mt + \sqrt{am} Z_t, \quad t \in (-\infty, \infty) \tag{3.3}$$

where Z_t is a normalized FBM. The process is completely characterized by the mean (m), variance coefficient $(a = \sigma^2/m)$ and Hurst parameter $(H \in [1/2, 1))$.

Performance of a FBM model in fluid queue

Once the FBM process has been defined, we can analyse its performance in a fluid based queue. The occupancy of a fluid queue model is defined as follows:

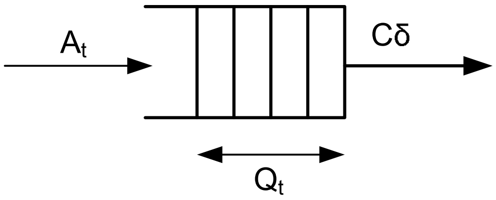


Figure 3.4: Fluid queue model

$$Q_{t+1} = max\left\{Q_t + (A_{t+1} - A_t) - C * \delta, 0\right\}, \quad t \in (-\infty, \infty) \tag{3.4}$$

where C is the queue capacity and δ is the increment in seconds from t to $t+1$. Figure 3.4 represents the fluid queue model used.

If a FBM trace is injected to a fluid based queue, the survival function of its occupation is as follows ([56]):

$$\begin{aligned} P(X > x) &\sim e^{(\frac{x}{r})^s}, \quad x \geq 0 \\ s &= 2 - 2H \\ r &= \frac{1}{C}\left(\frac{2K(H)^2 am}{(C-m)^{2H}}\right)^{\frac{1}{2-2H}} \end{aligned} \tag{3.5}$$

where $K = H^H(1-H)^{(1-H)}$ and m, a and H are the parameters of the incoming traffic to the queue defined as a FBM. Our model does not consider the queue occupation, but the queueing delay that is related with queue occupation: $d = x/C$, where d is the delay in fluid queue. Just for notation purposes, x variable is used as the queuing delay in the following sections.

Simulation example

In order to validate the traffic model, the experiment carried by the authors in [56] using the Bellcore trace was repeated. Table 3.1 shows the estimated parameters for the Bellcore trace. The estimation of the Hurst parameter is still an open issue. In [57], there is a throughout explanation of the challenges in the estimating of the Hurst parameter.

The δ parameter measures the minimal timescale of the fluid model (eq. 3.4). The Bellcore trace information is the arrival packet time and the packet size. To transform this information to a fluid model (bits per time unit), it is necessary to give a value to δ. For

Parameter	m	a	H
Bellcore (1024s)	2279 Kbps	262.8 Kbps	0.78

Table 3.1: Estimated Parameters Bellcore trace

example, figure 3.5 shows the trace bitrate $(A_{t+1} - A_t)$ using a δ of half a second.

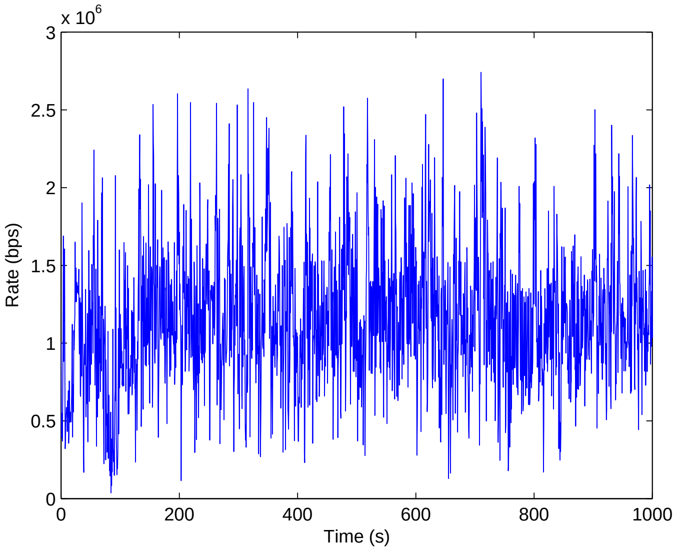


Figure 3.5: Variation of the bitrate (b/s, averaged over intervals of 0.5 s) in the Bellcore trace

After aggregating the traffic to timescales from 2^{-1} to 2^{-8} seconds, we simulated their behaviour into the model in eq. 3.4. With these simulations, the occupancy was obtained and its survival function was computed. Besides, using eq. 3.6 the analytical survival function was calculated. Figure 3.6 shows the survival function of all aggregation timescales and the analytical one. The analytical expression fits the results in the body of the distribution, so the well-known results of [56] are validated and so is the use of this traffic model for simulations. The analytical survival function cuts the simulations, but only in the function tail. The reason is that most of the samples are between 0 and 100 Kb and bigger values are strange events.

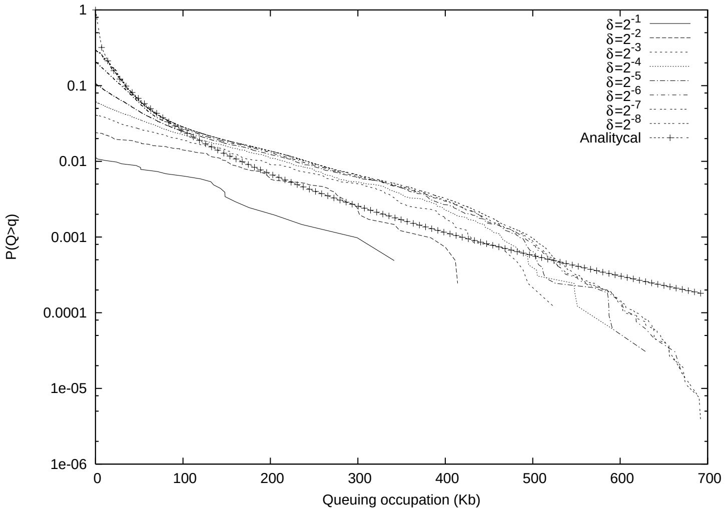


Figure 3.6: Bellcore trace FBM model validation

3.2 Analysis

3.2.1 Problem statement

As previously stated, the aim is to define a mathematically rigorous set of rules that helps such multi-layer capable core routers decide whether to switch a given LSP in the optical domain or in the electronic domain.

At a given time, a multi-layer router handles a number of LSPs. Typically, due to QoS constraints, optical switching is preferred due to the lack of queuing delay. In principle, many LSPs can be multiplexed in the electronic domain, whereas the lightpath bandwidth may be underutilized if LSPs are switched in the optical domain. This can be seen as a capacity planning problem. Given a set of input LSPs, the question is to derive the number of LSPs that should be switched in the electronic domain and the amount of LSPs to be switched in the optical domain (Figure 3.7), on attempts to maximize utility. It is preferred to switch in the electronic domain because the availability of buffering in core nodes allows for a higher utilization, and the remaining optical bandwidth can be used for newly arriving LSPs.

Thus, the router must trade-off these two parameters: queuing delay versus the cost associated to the utilization of optical switching, and needs to have a set of rules predefined to make a decision on how many LSPs should be switched in the optical domain and how many in the electronic domain.

To do so, let N refer to the number of LSPs handled at a given random time by the

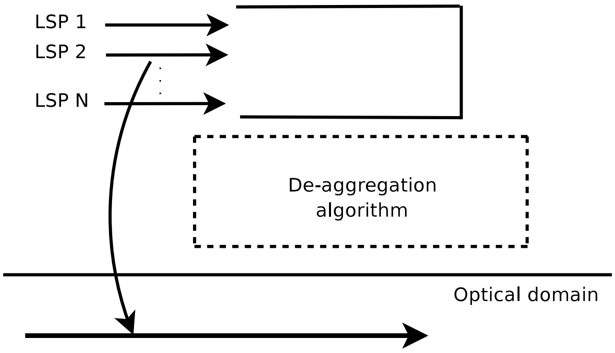


Figure 3.7: Multi-layer decision problem

multi-layer router, and let $L(d_e, x)$ refer to the loss function. The loss function $L(d_e, x)$ denotes the cost or loss of switching e LSPs in the electronic domain (thus, $N - e$ LSPs in the optical domain) with subsequent queuing delay experienced by the packets of the electronically switched LSPs, which is denoted by x (for simplicity, the optically switched LSPs have been assumed to experience zero delay). The term d_e denotes the "decision" of routing e LSPs out of a total of N in the electronic domain, and is defined for some decision space $\Omega = \{d_1, \ldots, d_N\}$. In this light, $L(d_e, x)$ is given by:

$$L(d_e, x) = (C_e(e) + C_o(N - e)) - U(x), \qquad e = 1, \ldots, N, \quad x > 0 \tag{3.6}$$

where $C_e(e)$ and $C_o(N - e)$ refer to the cost associated to routing e LSPs in the electronic domain and $N - e$ in the optical domain respectively; and $U(x)$ refers to the utility associated to a queuing delay of x units of time, experienced by the electronically switched LSPs.

Following [58], the Bayes risk, which is essentially the expectation of the loss function with respect to x, equals:

$$R(d_e) = \mathbb{E}_x L(d_e, x) = (C_e(e) + C_o(N - e)) - \mathbb{E}_x U(x), \qquad e = 1, \ldots, N \tag{3.7}$$

The goal is to obtain the optimal decision d_N^* such that the Bayes risk $R(d_N^*)$ is minimum. In other words:

$$\text{Find } d_N^* \text{ such that } R(d_N^*) = \min_{d_e, e=1,\ldots,N} R(d_e)$$

Following, section 3.2.2 computes the contribution of the utility functions($\mathbb{E}_x U(x)$) to the risk function. This contribution is based on the QoS experienced (in terms of queuing delay) by the electronically switched packets. Section 3.2.3 introduces a metric for quantifying the relative cost of optical switching with respect to electronic switching.

3.2.2 The utility function $U(x)$

As previously stated, the utility function $U(x)$ is defined over the random variable x, which represents the queuing delay experienced by the packets of electronically switched LSPs. The queuing delay shall be assumed to be Weibull distributed, since this has been shown to accurately capture the queuing delay behavior of a router with self-similar input traffic [56, 59, 60]. Consequently, we assume the model introduced in section 3.1.2. Eq. 3.6 is the survival function of a Weibull distribution. However, for notation purposes, let us compute the delay probability density function:

$$\begin{aligned} p(x) &= \frac{s}{r^s}(Cx)^{s-1}\exp\{-\left(\frac{Cx}{r}\right)^s\}, x \geq 0 \qquad (3.8) \\ s &= 2-2H \\ r &= \frac{1}{C}\left(\frac{2K(H)^2ame}{(C-me)^{2H}}\right)^{\frac{1}{2-2H}} \end{aligned}$$

where C is the lightpath capacity, m is the average input traffic and a is a variance coefficient such that $am=\sigma^2$ (with σ^2 being the input traffic variance) and H is the Hurst parameter.

Once $p(x)$ has been defined, the next step is to define a measure of the "utility" associated to routing LSPs in the electronic domain.

Delay based utility

In its simplest way, we can easily evaluate the utility based on the observed delay, that is, $U_{\text{mean}}(x) = -x$. The utility function is thus opposite to the queuing delay x, since the more utility occurs for smaller delays. Thus, computing the Bayes risk defined in eq. 3.7 yields:

$$\mathbb{E}_x[U_{\text{mean}}(x)] = \mathbb{E}_x[-x] = -\int_0^\infty xp(x)dx \qquad (3.9)$$

which equals the average queuing delay experienced by the electronically switched packets. Such value takes the following analytical expression:

$$\begin{aligned}
\mathbb{E}_x[U_{\text{mean}}(x)] &= -r\Gamma\left(1+\frac{1}{s}\right) \\
&= -\frac{1}{C}\left(\frac{2K(H)^2ame}{(C-me)^{2H}}\right)^{\frac{1}{2-2H}}\Gamma\left(\frac{3-2H}{2-2H}\right) \qquad (3.10)
\end{aligned}$$

Hard real-time utility

Hard real-time services require that the delay is under a given threshold ($T_{\max}$) in order to achieve utility. If the delay is greater than $T_{\max}$, they do not receive any utility at all. Consequently, hard real-time utility can be modeled by a step function, and takes the expression:

$$U_{\text{step}}(x) = \begin{cases} 1, & \text{if } x < T_{\max} \\ 0, & \text{otherwise} \end{cases} \qquad (3.11)$$

The Bayes risk requires to compute the average utility:

$$\mathbb{E}_x[U_{\text{step}}(x)] = \int_0^{T_{\max}} p(x)dx = 1 - \int_{T_{\max}}^{\infty} p(x)dx = 1 - P(x > T_{\max}) \qquad (3.12)$$

which, according to eq, 3.9, leads to:

$$\begin{aligned}
\mathbb{E}_x[U_{\text{step}}(x)] &= 1 - e^{(\frac{Cx}{r})^s} \qquad (3.13) \\
&= 1 - exp\left(-\frac{(C-me)^{2H}}{2K(H)^2ame}(Cx)^{2-2H}\right), \quad x > 0 \qquad (3.14)
\end{aligned}$$

Elastic utility

As previously defined in section 3.1.1, the elastic utility gradually decays, till it exceeds a given threshold ($T_{\max}$) and no utility is achieved over such value. The exponential function has been used to describe the degradation of elastic services [51]. Thus, the elastic utility function is modeled as:

$$U_{\exp}(x) = \lambda e^{-\lambda x}, \quad x > 0 \qquad (3.15)$$

where λ refers to decay ratio of the exponential function. Following the definition of $T_{\max}$ above, the value of λ has been chosen such that 90% of the utility lies before $T_{\max}$. That is:

$$\lambda = \frac{1}{T_{\max}\log(1-0.9)} \tag{3.16}$$

Finally, the average elastic utility follows:

$$\mathbb{E}_x[U_{\exp}(x)] = \int_0^\infty \lambda e^{-\lambda x} p(x)dx \tag{3.17}$$

which has no analytical form. However, we can use the Taylor expansion to approximate it, since:

$$\begin{aligned}\mathbb{E}[f(x)] &\approx \int_0^\infty p(x)\left(f(\mathbb{E}[x]) + f'(\mathbb{E}[x])(x-\mathbb{E}[x]) + \frac{1}{2}f''(\mathbb{E}[x])(x-\mathbb{E}[x])^2\right)dx \\ &= f(\mathbb{E}[x]) + \frac{1}{2}f''(\mathbb{E}[x])\sigma_x^2 \end{aligned}\tag{3.18}$$

Thus, :

$$\begin{aligned}\mathbb{E}_x[U_{\exp}(x)] &\approx U_{\exp}(\mathbb{E}_x[x]) + \frac{1}{2}U''_{\exp}(\mathbb{E}_x[x])\sigma_x^2 \\ &\approx \lambda e^{-\lambda\mathbb{E}_x[x]} + \frac{1}{2}\lambda^3 e^{-\lambda\mathbb{E}_x[x]}\sigma_x^2 \end{aligned}\tag{3.19}$$

where $\mathbb{E}_x[x]$ is given by eq. 3.10, and the variance σ_x^2 can be easily derived from eq. 3.9:

$$\begin{aligned}\sigma_x^2 &= r^2\left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right] \\ &= \frac{1}{C^2}\left(\frac{2K(H)^2ame}{(C-me)^{2H}}\right)^{2/(2-2H)}\left[\Gamma\left(\frac{2-H}{1-H}\right) + \Gamma^2\left(\frac{3-2H}{2-2H}\right)\right] \end{aligned}\tag{3.20}$$

3.2.3 The utilization cost of electronic and optical switching

As previously stated, the values of $C_e(e)$ and $C_o(N-e)$ in eq. 3.6 represent the cost associated to switching e LSPs in the electronic domain and $N-e$ in the optical domain. As previously stated, optical resources should be penalized more than the electronic ones in order to maximize link utilization.

For simplicity purposes, we have considered a *linear* cost approach, at which electronic switching is penalized as $C_e(e) = Ke$ for some $K > 0$, and the cost of optical switching is $C_o(N-e) = R_{\text{cost}}K(N-e)$. The value of R_{cost} (generally $R_{\text{cost}} > 1$) denotes the relative optical-electronic cost, that is, the ratio at which the optical cost increases with respect to the electronic cost.

3.3 Experiments and results

3.3.1 Scenario definition

Following, a few numerical examples applied to real case scenarios are shown. The aim is to show practical cases at which the implemented algorithm, at a given core multi-layer switch, decides the number of optically switched LSPs that should be transmitted according to three sets of parameters: (1) QoS parameters, essentially the $T_{\max}$ value introduced above; (2) the relative cost R_{cost} which provides a measure of the utilization cost of switching LSPs in the optical domain with respect to the electronic switching; and, (3) the self-similar characteristics of the incoming flows, represented by the Hurst parameter H. Furthermore, the impact of the LSPs mean and variance modification are studied.

The simulation scenario assumes a 2.5 Gbps core network, which carries a number of $N = 60$ standard VC-3 LSPs (typically 34.358 Mbps each). The values of m, σ and H, which represent the characteristics of the traffic flows, i.e. average traffic load, variability and Hurst parameter, have been chosen as $H = 0.6$ (according to [61]) and m and σ such that $\frac{\sigma}{m} = 0.3$.

Finally, the value of K has been chosen as $K = \frac{1}{N}$, in order to get the electrical cost normalized, i.e. within the range $[0, 1]$.

3.3.2 Basic decisor behaviour

The Bayesian decisor can choose between sending e LSPs using the electronic layer and by-passing the rest of them $(N - e)$ using the optical resources. Since the current amount of LSPs for our case study is 60, $e = 1, \dots, 60$. Figure 3.8 illustrates a case example of the risk function when U_{mean} function is used and $R_{\text{cost}} = 2$. The increasing dashed line is the utility function, the linear dashed line is the cost function and the addition is the risk function (solid line). Let us remark that cost and utility functions are normalized between in the range $[0, 1]$ for this experiment. X axis of the Figure 3.8 shows the possible values of e. If the router would choose $e = 1$, its utility is high (absolute value), but also the associated cost to this choice is high, since most of the traffic would be sent using the optical domain ($N - e = 59$ LSPs). On the other hand, if the decisor would send 60 LSPs in the electronic domain (x-axis), no utility is achieved, since the queue occupation is very high (more than 80%). However, the utilization cost is zero, because just the IP resources are used.

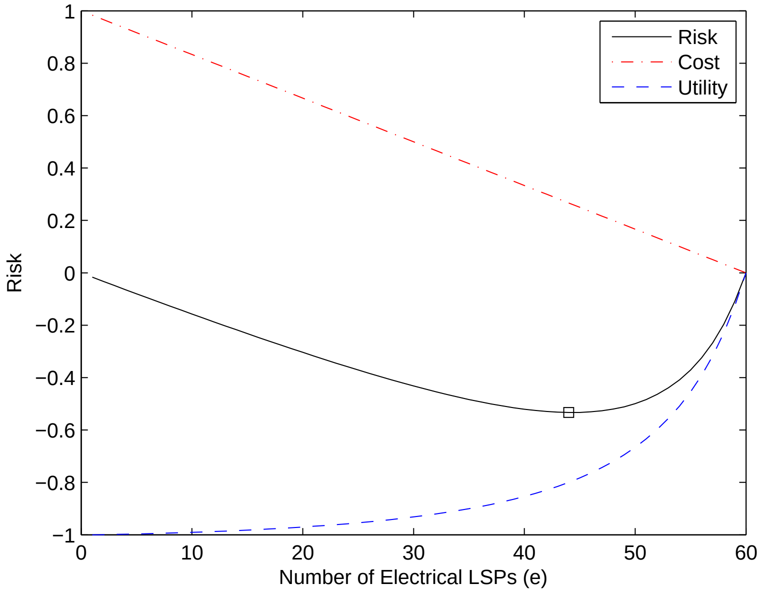


Figure 3.8: Risk function example with $N = 60$, U_{mean} and $R_{\text{cost}} = 2$

As the cost function monotonically decreases while the utility function monotonically increases, there is a minimum point which is the optimal decision. For this example d^*_{60} is 44 LSPs. By minimizing the risk function, the router trades of between queueing delay versus the cost associated to optical switching.

3.3.3 Study of threshold $T_{\max}$

This experiment shows the influence of the choice of $T_{\max}$ in the decision to be made by the multi-layer router with relative optical-electronic cost set to $R_{\text{cost}} = 2$. Figure 3.9 shows this case for several values of $T_{\max}$ assuming the step or hard real-time utility function (left) and the exponential utility function (right). The values of $T_{\max}$ have been chosen to cover a wide range from 1 ms to 100 ms. Clearly, the number of optically switched LSPs should increase with decreasing values of $T_{\max}$, since high QoS constraints require small delays in the packet transmission (thus larger number of optically switched LSPs to reduce latency).

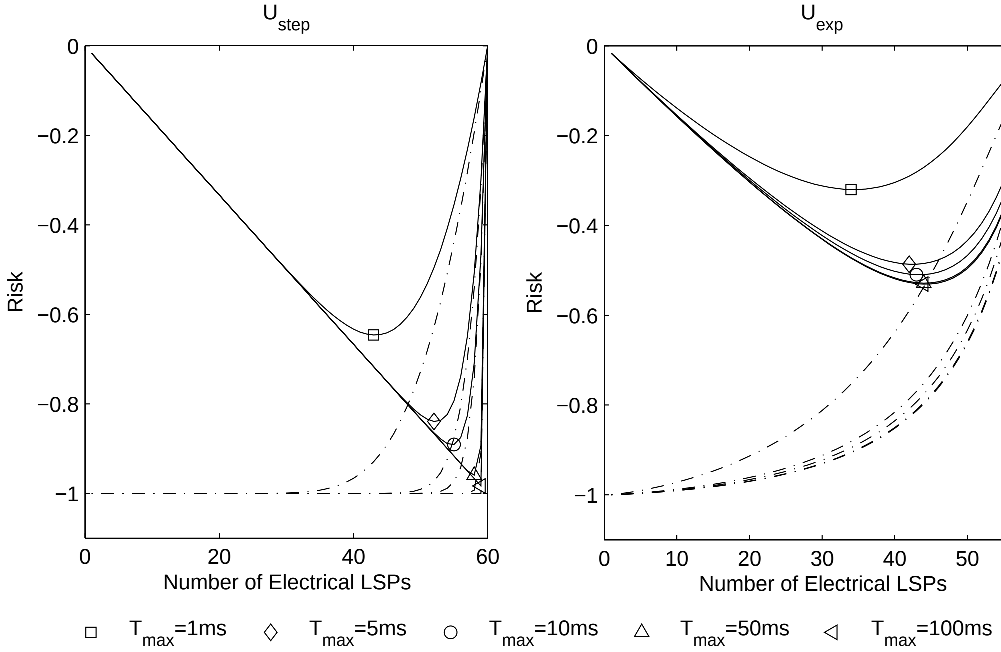


Figure 3.9: Bayes risk curves and optimal decisions (minimum values of risk) for several $T_{\max}$ values assuming hard real-time utility (left) and elastic utility (right) functions. Dashed line = Utility.

Typically, most of the end-to-end delay suffered by applications occur in the access network, and it is widely accepted that the core network should be designed to introduce delay of no more than $1-10\%$ of the total end-to-end delay. For hard real-time applications, which may demand a maximum end-to-end of 100 ms, the core delay is thus in the range of $1-10$ ms. This would require a total number of electronically switched LSPs of $d^*_{60} = d_{43}$ (see □) and $d^*_{60} = d_{55}$ (see ∘) respectively of a total of $N = 60$ LSPs. For the same delay constraints, elastic applications impose a number of electronically switched LSPs of $d^*_{60} = d_{34}$ (see □) and $d^*_{60} = d_{44}$ (see ∘) respectively.

3.3.4 Analysis with different R_{cost} values

This experiment shows the impact of R_{cost}, which refers to the relative cost of optical switching with respect to electronic switching, in the final decision d^*, to be taken by the multi-layer router. Figure 3.10 left shows where the optimal decision lies (minimum cost) for different R_{cost} values considering the case of linear utility function. As shown in the right, the more expensive optical switching is (large values of R_{cost}), the less number of LSPs is switched optically. In other words, for high R_{cost} values, only a small portion of

LSPs is switched in the optical domain. This becomes clear for $R_{\text{cost}} = 4$, at which the first optical LSPs occurs after $i = 40$ electronically switched LSPs. Figure 3.11 shows the evolution of the optimal number of electronically switched LSPs i with respect to N when for the exponential utility function. Its behavior is quite similar to the U_{mean} function with differences in the function slope.

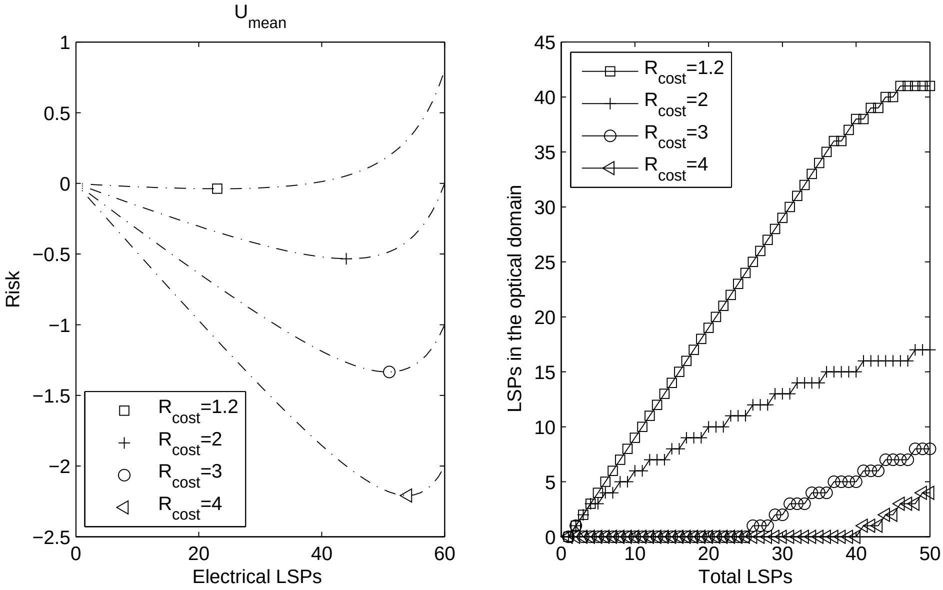


Figure 3.10: Variation of relative cost (R_{cost}) for U_{mean}

Figure 3.12 shows the evolution of the optimal number of electronically switched LSPs i with respect to N for the hard real-time utility function. Essentially, the functions U_{mean} and U_{exp}, as defined in section 3.2.2, show a smooth decrease with respect to delay, whereas U_{step} has an abrupt utility transition at the value $T_{\max}$. Such abrupt transition is further translated to the optimal decision, as shown in the figure.

To sum up, when optical switching becomes too expensive, the R_{cost} is critical in the optimal decision, thus canceling any influence of the QoS parameter $T_{\max}$. In this light, the network operator has a means to decide where the optimal decision lies, trading off the R_{cost} parameter and the QoS values.

3.3.5 Influence of the Hurst parameter H

The previous two numerical examples have assumed a value of $H = 0.6$, as observed in real backbone traces [61]. However, other scenarios may show different values of H and it

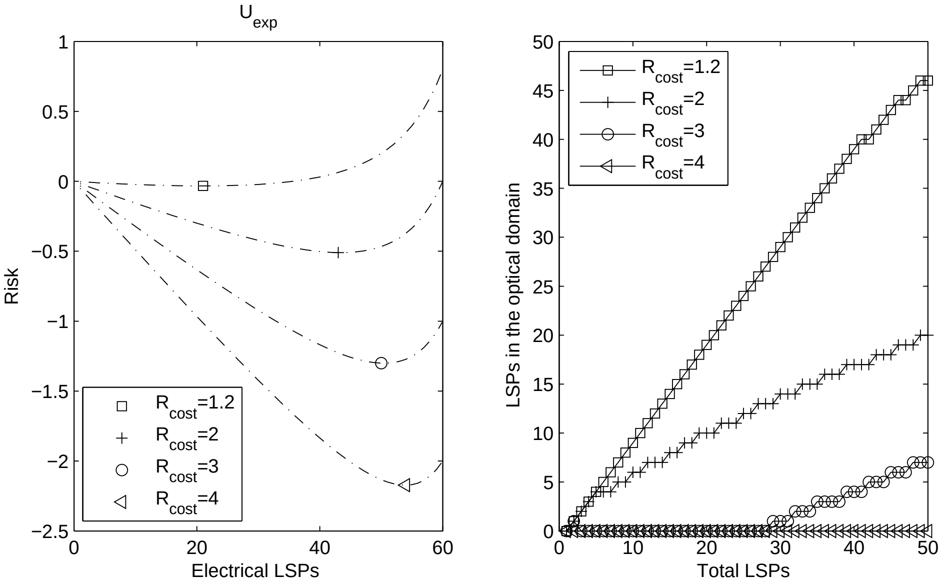


Figure 3.11: Variation of relative cost (R_{cost}) for U_{exp}

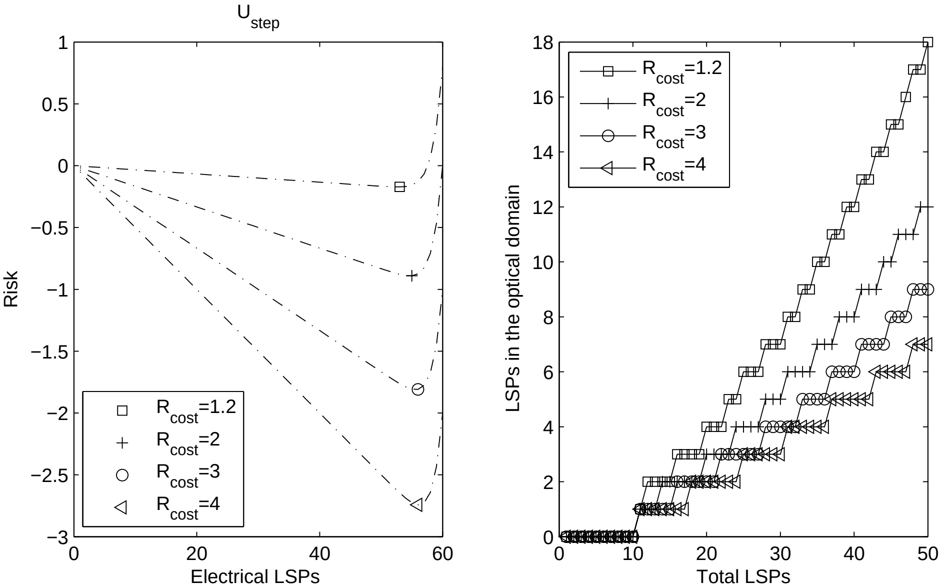


Figure 3.12: Variation of relative cost (R_{cost}) for U_{step}

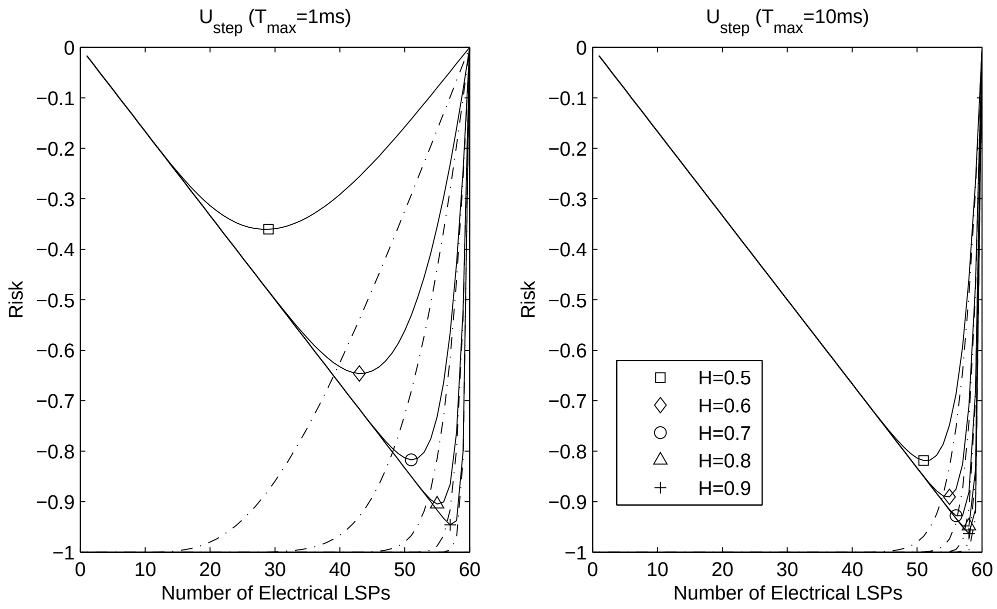


Figure 3.13: Hurst parameter variation

is interesting to study its impact on the bayesian decisor. In this light, Figure 3.13 shows the influence (left) or no-influence (right) of such parameter H in the optimal decision. In spite of the fact that long-range dependence degrades queuing performance generally, at high-delay values, the delay variability is smaller for high values of H (see [56], Figure 3.13).

Thus, the characteristics of the incoming traffic have a higher or lower impact on the bayesian decisor, depending on the QoS parameters. When $T_{\max} \geq 10$ ms, there is little influence of H (fig 3.13 right), but for $T_{\max} = 1$ ms and smaller, the value of H is key since it moves the decision in a wide range of optimal values: from d_{29} in the case of $H = 0.5$ to d_{57} for $H = 0.9$ (fig 3.13 left).

The level curves shown in Figure 3.14 show such behavior for the three utility functions ($T_{\max} = 10$ ms). Each level curve corresponds to a different utility.

Figure 3.14 middle (case of exponential utility function) and left (case of mean utility function) shows an influence with the H value. However, Figure 3.14 right (case of step utility function) should read as no influence with the Hurst parameter (i.e. parallel level curves = optimal decision independent of H value). It is important to remark that such independence behavior with parameter H does not occur if $T_{\max} = 1ms$ is chosen.

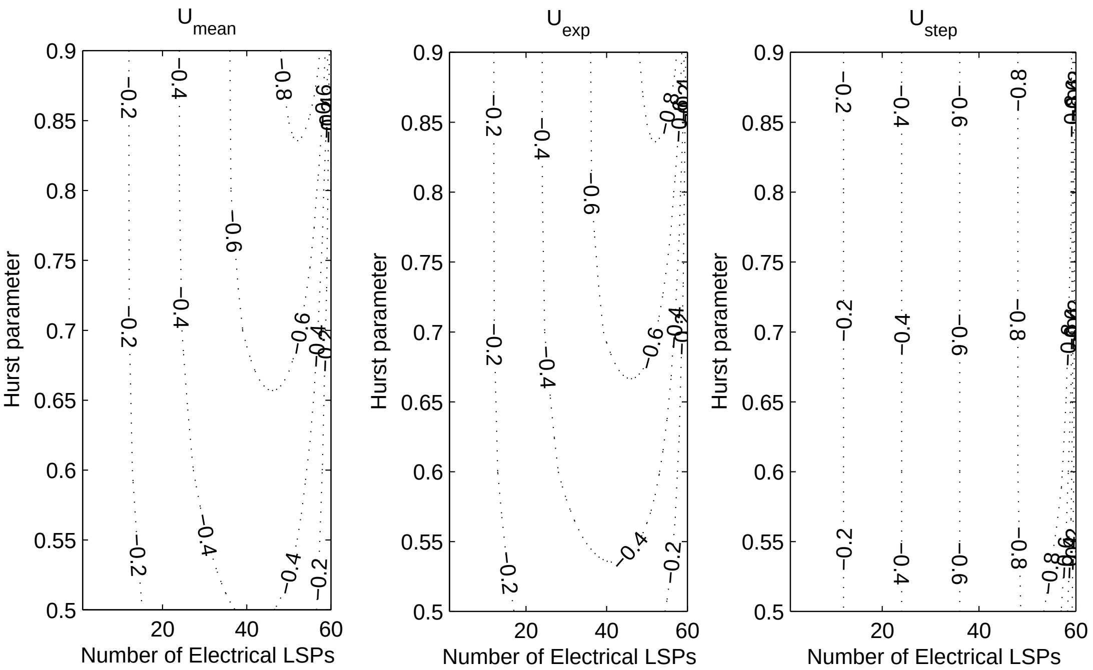


Figure 3.14: Hurst parameter variation (risk level curves)

3.3.6 Impact of the mean and variance of the LSPs

Typically, a network operator agrees a Service Level Agreement with its customers, but it may well happen that less channel capacity is used or that the traffic variation changes. Accordingly, this experiment studies the decisor behavior when the LSPs' mean and variance (values m and σ^2) vary, for different utility functions, which are shown in Figure 3.15 and 3.16 respectively. In the former, the original LSPs mean value is $m = 34.358$ Mbps (VC-3), depicted with $\triangle$. This value has been modified in the range from -10% (30.92 Mbps, depicted as $\Box$) to $+1\%$ (34.7 Mbps, as $\triangleleft$). This range aims to simulate the case of LSPs transmitting at a much lower ratio (-10% to $+1\%$), but rarely exceed 1% of its nominal rate. As shown, there is little influence in the final decision d^*_{60}, especially in the case for the step utility function. Clearly, the change in the LSPs' transmission rate has an impact on the optimal decision, but nevertheless this impact is much smaller than the impact of other system parameters: $T_{\max}$ (Figure 3.9), R_{cost} (Figure 3.10) and H (Figure 3.13).

Finally, for changes in parameter σ^2 (rate variance), the impact is negligible, as shown in fig. 3.16.

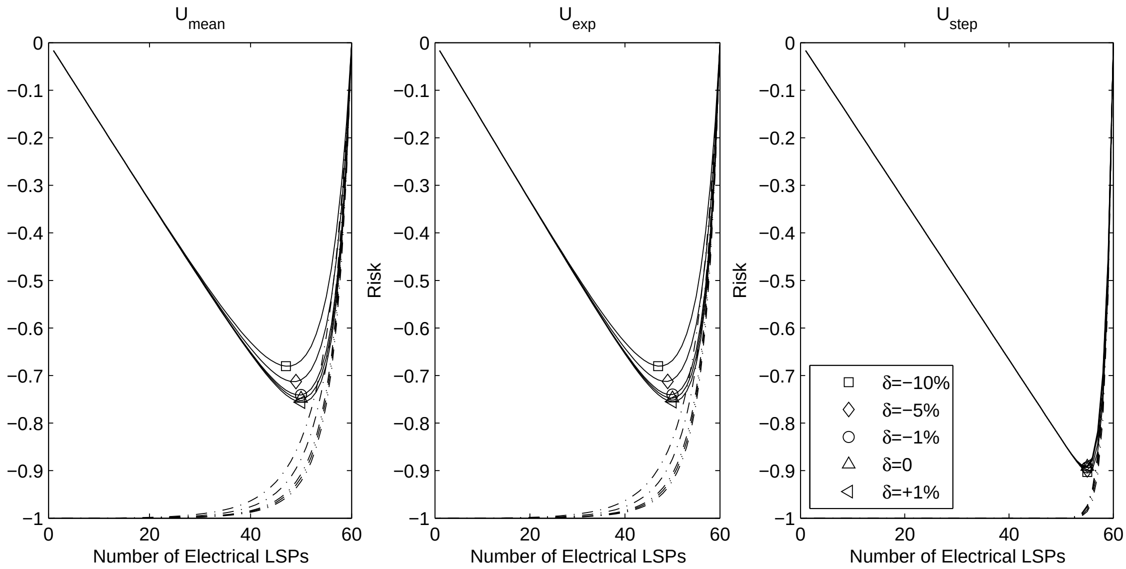


Figure 3.15: Mean LSPs variation

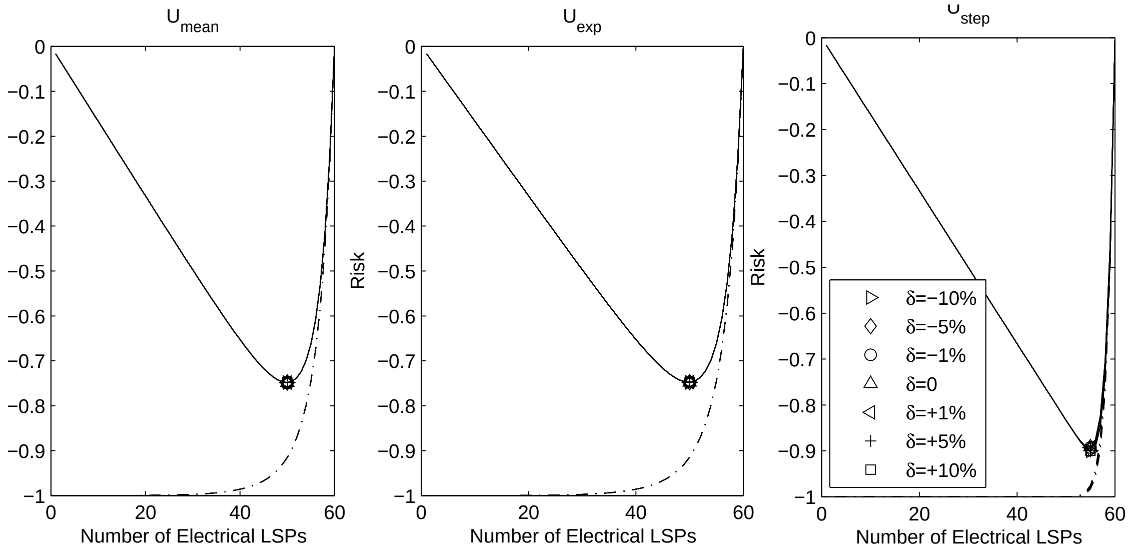


Figure 3.16: Variance LSPs variation

3.4 Dynamic behavior of the risk process

3.4.1 Analysis

This section studies the dynamic behavior of the risk process in a multi-layer router. Let us consider that LSPs arrive following a Poisson process (rate λ) with exponentially distributed duration (mean $1/\eta$). The cost process can be formulated as a Continuous-Time Markov Chain. More specifically, let $\{X(t), t > 0\}$ denote the chain that gives the *total* number of LSPs under service by a multi-layer router at time t, out of which i LSPs are switched in the electronic domain and $X(t)-i$ are being switched in the optical domain. And let N_{max} denote the maximum number of LSPs supported by the multi-layer router simultaneously.

For simplicity, let us consider discrete-time transitions between states of the chain, which we denote by $\{X(n), n = 0, 1, \ldots\}$. Then, the transition probabilities p_{jk} are given by:

$$\begin{aligned} p_{j,j+1} &= \frac{\lambda}{\lambda + j\eta} \\ p_{j,j-1} &= \frac{j\eta}{\lambda + j\eta} \end{aligned} \tag{3.21}$$

for $j = 1, \ldots, N_{max} - 1$ and $p_{01} = p_{N_{max}(N_{max}-1)} = 1$, while $p_{jk} = 0$ for all other values of (j, k). We analyze the process $\{V_j(n), n = 0, 1, 2, \ldots; j = 0, 1, \ldots, N_{max}\}$ which refers to the accumulated cost in n steps of the chain $X(n)$, assuming it departed from state j. Let r_{jk} denote the cost associated to the transition from state j to state k. Let us define $r_{jk} = L(d_k^*, x)$, which corresponds to the decision policy of choosing the optimal decision according to eq. 3.6 for a total number of k LSPs. Accordingly, it follows that:

$$\begin{aligned} V_j(n) &= \sum_{k=1}^{n} p_{jk} \left(r_{jk} + V_k(n-1)\right) \\ j &= 1, 2, \ldots, N_{max} - 1; \quad n = 1, 2, \ldots; \end{aligned} \tag{3.22}$$

and $V_j(0) = 0, V_0(1) = L(d_1^*, x), V_{(N_{max})}(1) = L(d_{N_{max}-1}^*, x)$, which follows directly from the one-step Chapman-Kolmogorov equations. Note that the accumulated cost in n steps from state j is equal to the cost of the one-step transition to state k plus the accumulated cost from such state k in $n-1$ steps. Now, we use eq. 3.21 to obtain:

$$\begin{aligned} V_j(n) &= p_{j(j+1)}(r_{j(j+1)} + V_{j+1}(n-1)) + p_{j(j-1)}(r_{j(j-1)} + V_{j-1}(n-1)) \\ j &= 1, 2, \ldots, N_{max} - 1; \quad n = 1, 2, \ldots; \end{aligned} \tag{3.23}$$

$$V_0(n) = r_{01} + V_1(n-1); \quad n = 1, 2, \ldots; \tag{3.24}$$

$$V_{N_{max}}(n) = r_{N_{max}(N_{max}-1)} + V_{N_{max}-1}(n-1); \quad n = 1, 2, \ldots; \tag{3.25}$$

and $V_j(0) = 0$. Finally, expanding the transition probabilities brings the final recursion formula:

$$\begin{aligned} V_j(n) &= \frac{\lambda}{\lambda + j\eta}(r_{j(j+1)} + V_{j+1}(n-1)) + \frac{j\eta}{\lambda + j\eta}(r_{j(j-1)} + V_{j-1}(n-1)) \\ j &= 1, 2, \ldots, N_{max} - 1; \quad n = 1, 2, \ldots; \end{aligned} \tag{3.26}$$

and $V_j(0) = 0$. It is worth noticing that the expression for $V_0(n)$ and $V_{N_{max}}(n)$ remain the same. Taking expectations in both sides of the equation gives:

$$\begin{aligned} \overline{V_j(n)} &= \frac{\lambda}{\lambda + j\eta}(R(d^*_{j+1}) + \overline{V_{j+1}(n-1)}) + \frac{j\eta}{\lambda + j\eta}(R(d^*_{j-1}) + \overline{V_{j-1}(n-1)}) \\ j &= 1, 2, \ldots, N_{max} - 1; \quad n = 1, 2, \ldots; \end{aligned} \tag{3.27}$$

$$\overline{V_0(n)} = R(d^*_1) + \overline{V_1(n-1)}; \quad n = 1, 2, \ldots \tag{3.28}$$

$$\overline{V_{N_{max}}(n)} = R(d^*_{N_{max}-1}) + \overline{V_{N_{max}-1}(n-1)}; \quad n = 1, 2, \ldots; \tag{3.29}$$

and $\overline{V_j(0)} = 0$ where $R(\cdot)$ is given by eq. 3.7 and $\overline{V_j(n)} = \mathbb{E}(V_j(n))$.

Now, we explicitly calculate the first steps of the recursion formula ($n = 0, 1$), and provide results for a generic $n = 1, 2, ..., 10$ in section 3.4.2. For $n = 0$, as defined above:

$$\overline{V_j(0)} = 0 \qquad j = 0, 1, \ldots, N_{max} \tag{3.30}$$

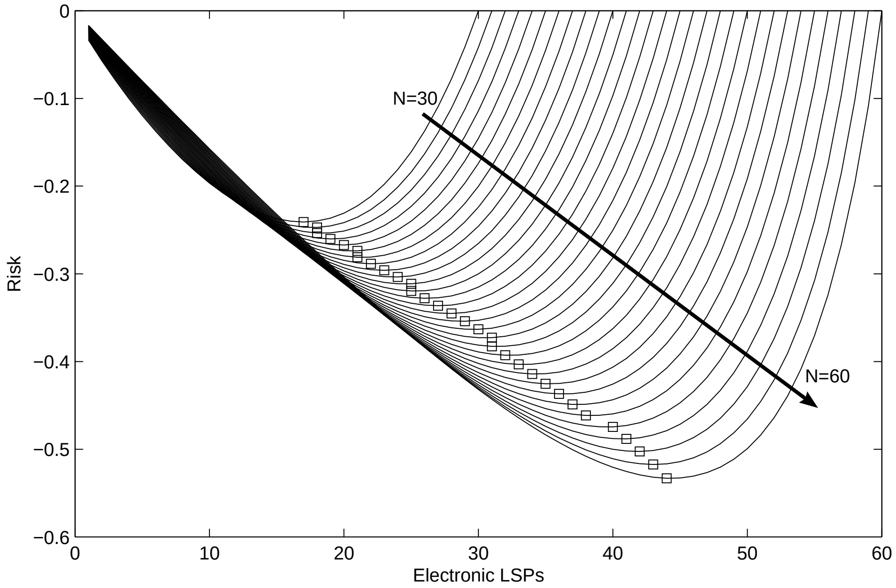


Figure 3.17: Risk versus optimal number of electronic LSPs for a total number of LSPs in the range $30-60$

For $n=1$, using the result for $n=0$, it yields:

$$\overline{V_j(1)} = \frac{\lambda}{\lambda+j\eta}R(d^*_{j+1}) + \frac{j\eta}{\lambda+j\eta}R(d^*_{j-1}) \qquad j=1,\dots N_{max}-1 \tag{3.31}$$

and $V_0(1) = R(d^*_1), V_{(N_{max})}(1) = R(d^*_{N_{max}-1})$. Eq. 3.27 provides the dynamic behavior of the Bayes risk, as we move n steps forward from any state j.

3.4.2 Numerical example

Fig. 3.17 shows the risk curves of a multi-layer router as a function of the total number of LSPs switched. Actually, each curve represents a different (increasing) number of LSPs. The optimal decisions are represented with symbol □. In the figure, a total number of $j=30$ LSPs gives an optimal decision of $d^*_{30}=d_{17}$ electronically switched LSPs. As shown, as the number of LSP arrivals increases, the optimal number of electronically switched LSPs i also increases, thus reducing the risk. For $j=60$, the optimal number of LSPs switched in the electronic domain is $d^*_{60}=d_{44}$. Interestingly, in some cases, the optimal number of electronically switched LSPs remains the same regardless of a small increment in the number of incoming LSPs j (several squares along the same vertical line).

The accumulated risk function $V_j(n)$ as defined in the section above was analytically solved for $n = 0$ and $n = 1$ only. Fig. 3.18 shows the time evolution of $V_j(n)$ for a time horizon from $n = 1$ to $n = 10$, given different initial states ($j \in \{30, 50, 70\}$). The $\rho = \frac{\lambda}{\eta}$ value for the experiment was 50%, although the variation of this parameter was tested and it was not outstanding. $V_j(n)$ is a monotonically decreasing function, since as previous risk curves have shown (for instance Fig. 3.17), the risk of the optimal decision (d_j^*) is negative. Besides, as the decisor works with the loss function (eq. 3.6), it is reasonable that the function decreases. If the decisor had been defined with a profit function, $V_j(n)$ would have been a monotonically increasing function. In other words, a negative loss function implies that there is a (positive) revenue for the operator. The initial state for $V_j(n)$ determines the evolution of the accumulative risk function and its slope is determined by this point. Furthermore, it is worth noticing that there is a quasi-linear behavior for all curves.

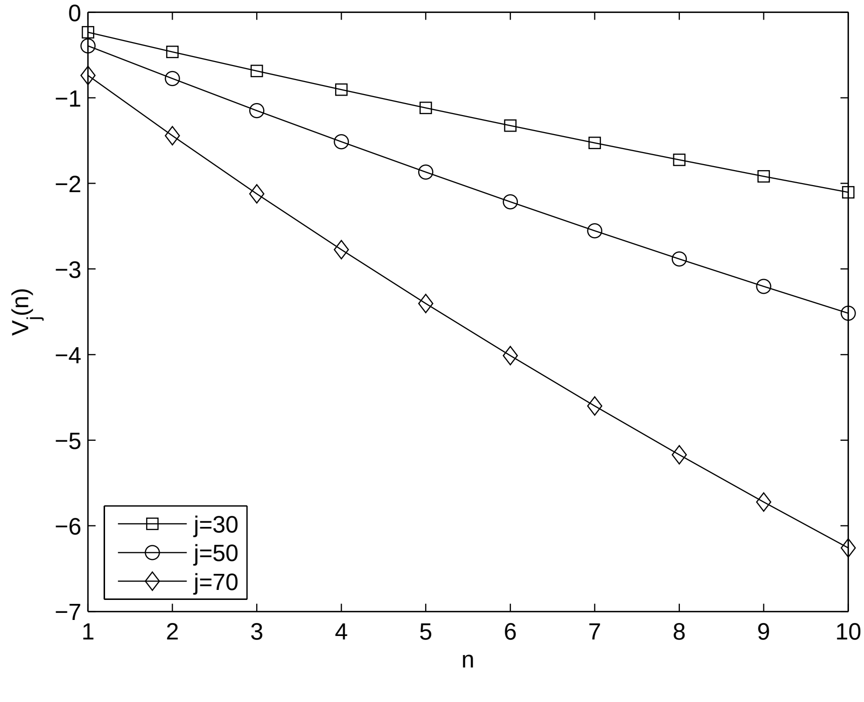


Figure 3.18: Evolution of $V_j(n)$ from different states

3.5 Conclusions

The main contribution of this chapter is that first, it presents a novel methodology, based on the Bayesian decision theory, that helps multi-layer capable routers to make the decision of either optical or electronic switching of incoming LSPs. Such decision is based on technical aspects such as QoS constraints and long-range dependence characteristics of the incoming traffic, but also considers the cost differences of optical and electrical switching. This way permits high flexibility to the network operator to trade off both QoS and resource utilization aspects.

Moreover, the Bayesian decision theory framework is of low complexity, and can easily adapt to changing conditions: QoS guarantees, traffic profiles, resource utilization and network operator preferences.

Finally, this first approach of the algorithm can be implemented in a per node basis by using local and independent parameters (e.g delay thresholds and optical-electronic cost) in each node. However, the local parameters are not enough in the provisioning of end-to-end services.

Chapter 4

End to end service provisioning in Multi-layer networks

This chapter provides an extension of the Bayesian decisor model to an end-to-end scenario. This model is based on chapter 3 framework, so this chapter demonstrates that the compromise between the utility perceived by the customers in terms of delay and the utilization costs of the optical and electronic resources is kept in this multi-hop scenario. The mathematical formulation of such Bayesian decisor is formulated for this new scenario and its behavior is further analyzed for different configurations. Our findings show that the algorithm is capable of adapting decisions according to traffic characteristics, while utilizing only the optical resources when it is required.

This chapter is organized as follows: Section 4.1 presents the multi-hop scenario whereby the Bayesian decision algorithm is to be applied. This section also introduces the risk function as it is defined in Bayesian decision theory, and computes its individual components: cost and utility (QoS perceived) to find the optimal decision. Section 4.2 shows the behavior of the algorithm in such a multi-hop scenario and shows how to adjust the model parameters to trade off QoS and cost. Finally, Section 4.3 concludes this chapter.

4.1 Problem statement: Multi-hop scenario

Let us assume the multi-hop scenario proposed in Figure 4.1, with $M+1$ multi-layer capable nodes. In this scenario, each node j (with $j = 1, \ldots, M$) is offered a number of N_j Label Switch Paths (LSPs) sent to the destination node. Each node j decides the number e_j of LSPs that are electronically switched to the destination node, thus remaining o_j LSPs

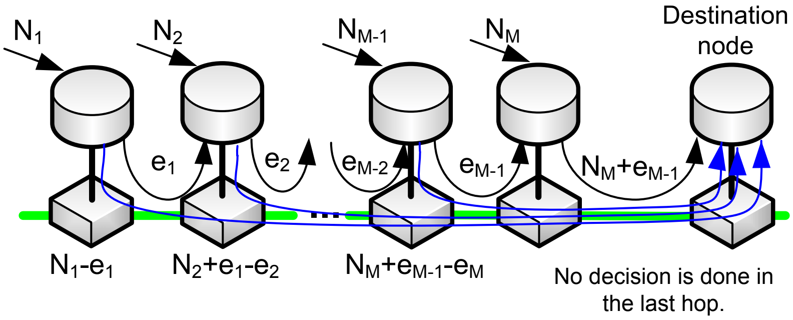


Figure 4.1: Multi-hop scenario under study

to be transmitted all-optically. The electronic transmission of LSPs implies traversing all intermediate nodes, thus suffering O/E and E/O conversions and electronic buffering (hence delay) at each of them. Optical switching implies the creation of an end-to-end (e2e) lightpath (from the source to the destination node) with no O/E conversion or delay experienced. Clearly, optical switching provides better QoS experienced by the LSPs (no delay at intermediate nodes), but requires extra resource consumption (the creation of new lightpaths per e2e optical switching). These two aspects of multi-layer switching (QoS perceived and extra Cost of using resources) must be traded off by the Bayesian decisor to find the optimum number of optically- and electronically-switched LSPs.

For notation purposes, let node 1 offer N_1 LSPs to the multi-layer Bayesian decisor. This decides to switch e_1 LSPs electronically (thus offered to node 2), and $o_1 = N_1 - e_1$ LSPs optically (Figure 4.1). Node 2 therefore must decide the number e_2 of electronically switched LSPs among the total N_2+e_1 offered, thus leaving $o_2 = N_2+e_1-e_2$ LSPs to go all-optically to the destination node. Following this reasoning, node j is offered N_j+e_{j-1} total LSPs and, among them, e_j and $o_j = N_j + e_{j-1} - e_j$ are transmitted through the electronic (hop-by-hop switching) and optical domains (direct lightpath) respectively. Let us remark that an e2e ligthpath in the M-th node uses the same resources than a hop-by-hop service, hence the last hop makes no decision (thus $e_M = N_M + e_{M-1}$).

The next section defines the Risk function, which is based on the cost of using optical and electronic resources and the e2e delay experienced by the electronically switched LSPs.

4.1.1 Risk function definition

Let $\vec{e} = \{e_1, e_2, \ldots, e_{M-1}\}$ denote the decision vector which gives the number of LSPs transmitted over the IP (electronic) layer at each hop j. The queueing delay experienced by electronically-switched LSPs or flows at hop j (that is, delay from node j to node $j+1$) x_j depends on the load offered e_j. Thus, the e2e delay experienced by a given LSP offered at node j is $x_j^{e2e} = \sum_{i=j}^{M} x_i$ since it must traverse the subsequent nodes until destination (with their respective delays). A loss function is stated as:

$$L(\vec{e}, x^{e2e}) \;=\; K_c C_T(\vec{e}) - K_u \sum_{j=1}^{M} U(x_j^{e2e}) \tag{4.1}$$

where $C_T(\vec{e})$ and $U(x_j^{e2e})$ refer to the utilization cost and utility function associated to the decision vector $\vec{e}$. Both quantities are weighted by constants K_c and K_u. With these parameters, we define the Bayesian Risk function, like we did in section 3.2.1:

$$\begin{aligned} R(\vec{e}, x^{e2e}) &= \mathbb{E}_x L(\vec{e}, x^{e2e}) \\ &= K_c C_T(\vec{e}) - K_u \sum_{j=1}^{M} \mathbb{E}_x \left[U(x_j^{e2e}) \right] \\ & \quad\; x^{e2e} \geq 0 \end{aligned} \tag{4.2}$$

Clearly, the goal is to find the optimum decision vector $\vec{e}^{*}$ that minimises the Bayesian Risk given by eq. 4.2. Note that this is a discrete optimization problem in $M-1$ dimensions, which yields the optimal decision vector $\vec{e}^{*} = (d^{*}_{N_1}, d^{*}_{N_2+i_1}, \ldots, d^{*}_{N_M+i_{M-1}})$.

Since the decision vector $\vec{e}^{*}$ gives the number of LSPs switched in each domain (optical and electronic), the Bayesian Risk finds a trade-off between the utility perceived by the traffic sent through the electronic domain and the cost related with the utilization of the optical and electronic resources. The following explains how to compute $C_T(\vec{e})$ and $\mathbb{E}_x \left[U(x_j^{e2e}) \right]$ in a multi-hop scenario.

4.1.2 Cost of using resources

This function accounts the cost C_e of using hop-by-hop connections (electronic switching) and the cost C_o of using end-to-end ligthpaths (optical switching) which, for a decision vector $\vec{e}$, is given by:

$$C_T(\vec{e}) \;=\; C_e(\vec{e}) + R_{\text{cost}} C_o(\vec{e}) \tag{4.3}$$

where R_{cost} is the relative cost of using the optical and electronic resources. In other words, an optical lightpath is R_{cost} times more expensive than the same connection in the electronic layer. Note that R_{cost} is not a monetary cost but a metric that helps network operators decide how valuable their optical resources are with respect to the already deployed IP layer.

Cost computation has been chosen to follow the next design premises:

1. LSPs should be switched in the electronic domain while their utility perceived is correct, hence the cost of using electronic resources is cheaper than that of using optical resources, for the same amount of traffic (routers are already deployed).

2. If an optical bypass is to be set up, the longer it is, the better (less cost), that is, the cost of long connections should be lower than short optical by-pass connections.

Thanks to this cost model, only the necessary e2e optical connections are created, and this occurs when the IP layer do not provide the necessary utility to the traffic.

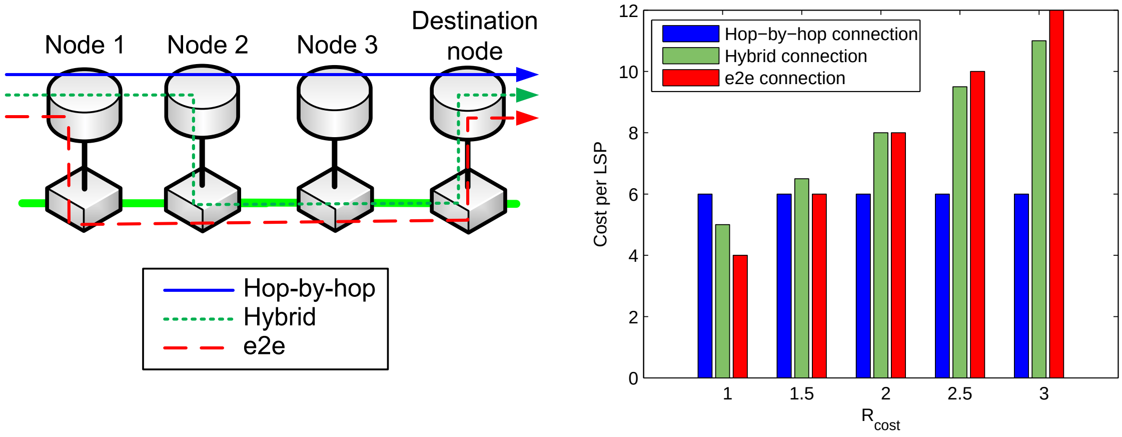
Following these premises, we define the cost of transmitting an LSP optically per hop as $\frac{k+1}{k}$, where k is the length of the optical by-pass (that is, a lightpath created from node j to the destination node is of length $k = M + 1 - j$). Note that this series is strictly decreasing since $\frac{k+1}{k} > \frac{L+1}{l}$, $\forall k < L$, giving a cheaper cost per hop the longer the lightpath is, thus promoting the creation of long e2e by-pass optical connections in the network. It is worth noticing that, in the scenario proposed in Figure 4.1, the longest (and cheapest) lightpath possible is of cost $\frac{M+1}{M}$, and it is the cheapest one since $\frac{M+1}{M} < \frac{k+1}{k}$, $\forall k < M$. In conclusion, the optical cost of sending e LSPs through k hops is $\frac{k+1}{k} \times k \times e = (k+1) \times e$. The definition of the electronic and optical cost functions are as follows:

$$C_e(\vec{e}) = \sum_{j=1}^{M} 2e_j \tag{4.4}$$

$$C_o(\vec{e}) = \left((M+1)(N_1 - e_1) + \sum_{j=2}^{M-1} (M - j + 2)(N_j - e_j + e_{j-1}) \right) \tag{4.5}$$

According to the previous definitions (eq. 4.3, optical resources are R_{cost} times more expensive than electronic in the case of one-hop switching. The one-hop electronic cost is 2 ($k = 1$), and $\frac{(M+1)}{M}$ is the cheapest optical cost per hop in this scenario, since the maximum optical path length is M. According to this, R_{cost} must satisfy $R_{\text{cost}} > \frac{2 \cdot M}{M+1}$ to ensure that the cheapest optical lightpath is more expensive than its electronic counterpart. Otherwise, if optical resource utilization is very cheap and they add no delay to the packets, there is no reason to send the traffic using the IP layer.

Figure 4.2(a) shows a multi-hop scenario with three hops ($M = 3$). In such scenario there are three possible end-to-end paths from node 1 to the destination node. If a LSP is sent through the hop-by-hop connection, the associated cost is due to the electronic cost, since no utilization of the optical resources is done. Its cost would be $2 \times M$. If the end-to-end connection is used, the cost is just optical cost and it is $(M+1) \times R_{\text{cost}} = 4 \times R_{\text{cost}}$. Concerning the hybrid case, the cost is one hop electronic and two hops in the optical domain, so its cost is $2 + M \times R_{\text{cost}} = 2 + 3 \times R_{\text{cost}}$. Figure 4.2(b) depicts the cost of sending one LSP using the hop-by-hop connection, the end-to-end lightpath or a hybrid connection, when $R_{\text{cost}} = \{1, 1.5, 2, 2.5, 3\}$. According to the previous designed rule, R_{cost} should be greater than $\frac{2 \cdot M}{M+1}$, in this case 1.5. This is the reason why the cost per LSP of a hop-by-hop connection is more expensive when $R_{\text{cost}} = 1$ or equal when $R_{\text{cost}} = 1.5$ than the end-to-end path. When $R_{\text{cost}} > 1.5$, the cost per LSP is cheaper in the electronic than in the optical domain. The cost of the hybrid connection is intermediate, since the first hop is done in the IP layer and the rest in the optical domain. Let us remark that the cost is not the only value to make the decision. The LSPs sent through the IP layer suffer a delay, which increase their risk. Therefore, although the cost is lower the traffic can be routed in the optical domain.



(a) Multi-hop scenario with three hops ($M = 3$) and possible end-to-end paths

(b) Cost of routing one LSP in each path

Figure 4.2: R_{cost} designed rule example

Once the definition of the cost function is stated, replacing eq. 4.4 and 4.4 in eq. 4.2, risk function yields:

$$\begin{aligned} R(\vec{e}, x^{e2e}) &= K_c \left[\sum_{j=1}^{M} 2e_j + R_{\text{cost}} \left((M+1)(N_1 - e_1) + \sum_{j=2}^{M-1} (M-j+2)(N_j - e_j + e_{j-1}) \right) \right] \\ & \quad -K_u \left[\mathbb{E}_x \left\{ U(x_{e2e}) \right\} \right], x_{e2e} \geq 0 \end{aligned} \tag{4.6}$$

4.1.3 Utility functions definition

The utility function applied to a decision vector $\vec{e}$ gives a metric for the delay experienced by the electronically-switched LSPs, such that, the more delay experienced by them, the less utility achieved. The electronically-switched LSPs are assumed to experience some degree of delay, since they must traverse several hops with their respective electronic queues. On the other hand, the delay experienced by the optically-switched LSPs is assumed negligible compared to the electronic delay, since optical LSPs are provided a dedicated e2e path. Such an electronic delay is calculated based on the load level of a queue fed with self-similar traffic, as explained in section 3.1.2. Once the e2e electronic delay is obtained, the utility function operates to derive a utility metric following one of these Class of Service (CoS) utility models (see section 3.2.2: average delay, hard real-time and elastic utilities, as follows:

Average delay-based utility (U_{mean})

This utility is defined as: $U_{mean}(x_j^{e2e}) = -x_j^{e2e}$ which, after applying the expectation operator $\mathbb{E}_x$ of eq. 4.2, provides a utility function based on the average e2e delay experienced by the electronically-switched LSPs. This value is computed as:

$$\begin{aligned} \mathbb{E}_x[U_{\text{mean}}(x_k^{e2e})] &= \mathbb{E}_x[-x_k^{e2e}] = -\sum_{j=k}^{M} \mathbb{E}_x[x_j] \\ &= -\sum_{j=k}^{M} \left\{ r\Gamma \left(1 + \frac{1}{s} \right) \right\} \end{aligned} \tag{4.7}$$

As we previously explained, this utility function can be used for best-effort services, whereby great service interactivity provides high utility values, but this utility function does not excessively penalize if such interactivity is low. In this multi-hop scenario, the U_{mean}

function estimates the utility based on the mean service delay. Finally, let us compute the expression of the risk function. $\mathbb{E}_x[U_{mean}(x)]$ computation is explained in 3.2.2 and its expression is the following:

$$\mathbb{E}_x[U_{mean}(x)] = -\frac{1}{C}\left(\frac{2K(H)^2ame}{(C-me)^{2H}}\right)^{1/(2-2H)}\Gamma\left(\frac{3-2H}{2-2H}\right) \tag{4.8}$$

Replacing $\mathbb{E}_x[U_{mean}(x)]$ in eq. 4.6, the expression of the risk function for the U_{mean} utility is:

$$\begin{aligned} R(\vec{e},x^{e2e}) &= K_c\left[\sum_{j=1}^{M} 2e_j + R_{\text{cost}}\left((M+1)(N_1-e_1)+\sum_{j=2}^{M-1}(M-j+2)(N_j-e_j+e_{j-1})\right)\right] \\ &\quad -K_u\left[\sum_{j=1}^{M}\frac{1}{C}\left(\frac{2K(H)^2ame_j}{(C-me_j)^{2H}}\right)^{\frac{1}{(2-2H)}}\Gamma\left(\frac{3-2H}{2-2H}\right)\right] \end{aligned} \tag{4.9}$$

Hard real-time utility (U_{step})

Some applications tolerate very well a certain e2e delay value until a given delay threshold, U_{step} is defined to deal with such scenarios (section 3.2.2):

$$U_{\text{step}}(x_j^{e2e}) = \begin{cases} 1, & \text{if } x_j^{e2e} < T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{4.10}$$

where the threshold $T_{\max}$ depends on the service or application. After applying the expectation operator $\mathbb{E}_x$ to U_{step}, it yields:

$$\mathbb{E}_x[U_{\text{step}}(x_k^{e2e})] = \mathbb{E}_x[U_{\text{step}}(\sum_{j=k}^{M} x_j)] = P(x_k^{e2e} < T_{\max}) \tag{4.11}$$

The calculation of the e2e delay expectation requires the convolution of the queueing delay pdf, which it is not possible to obtain analytically. However, we can approximate the e2e delay (x^{e2e}) by a Gaussian distribution, assuming that the per-hop delays are independent. The moments of such a Gaussian pdf are computed by the Weibull delay assumption (eq. 3.9):

$$P(x_j^{e2e} < T_{\max}) \sim N(\sum_{i=j}^{M}\mu_i, \sqrt{\sum_{i=j}^{M}\sigma_i^2}) = N(\mu_j^{e2e}, \sigma_j^{e2e}) \tag{4.12}$$

Once the pdf of the sum of the variables is computed, it is possible to calculate the percentile given by the threshold ($T_{\max}$). The risk function for the U_{step} utility is:

$$\begin{aligned} R(\vec{e}, x^{e2e}) &= K_c \left[\sum_{j=1}^{M} 2e_j + R_{\text{cost}} \left((M+1)(N_1 - e_1) + \sum_{j=2}^{M-1} (M-j+2)(N_j - e_j + e_{j-1}) \right) \right] \\ & -K_u \sum_{j=1}^{M} P(x_j^{e2e} < T_{\max}) \end{aligned} \qquad (4.13)$$

Elastic utility ($U_{\mathbf{exp}}$)

Other applications, such as voice transmission, experience slow service degradation with increasing delay, until a threshold delay is reached. The elastic utility function fits with the properties of this delay-sensitive applications:

$$U_{\exp}(x_j^{e2e}) = \lambda e^{-\lambda x_j^{e2e}}, \quad x_j^{e2e} \geq 0 \qquad (4.14)$$

where λ refers to decay ratio of the exponential function. This utility function lies somewhere in between the previous two, whereby excessive delays are highly penalized, but not that much as in the hard real-time utility case.

Finally, the value of λ is chosen such that $\alpha = 50\%$ of the total utility lies before a delay threshold $T_{\max}$:

$$\lambda = \frac{1}{T_{\max} \log(1-\alpha)} \qquad (4.15)$$

this value can be obviously adjusted for a given α.

Finally, after assuming the Gaussian approximation of eq. 4.12 for computing e2e delays, the expected utility obtained in this case is given by:

$$\begin{aligned} \mathbb{E}_x[U_{\exp}(x_j^{e2e})] &= \mathbb{E}_x[\lambda e^{-\lambda x_j^{e2e}}] \qquad (4.16) \\ &= \int_{-\infty}^{\infty} \lambda e^{-\lambda x_j^{e2e}} N(\mu_j^{e2e}, \sigma_j^{e2e}) dx_j^{e2e} \end{aligned}$$

This integral can be solved completing the square, achieving the following expression:

$$\mathbb{E}_{x_{e2e}}[U_{\exp}(x_j^{e2e})] = \lambda e^{\frac{\sigma_j^{2\ e2e} \lambda^2 - 2\mu_j^{e2e} \lambda}{2}} \qquad (4.17)$$

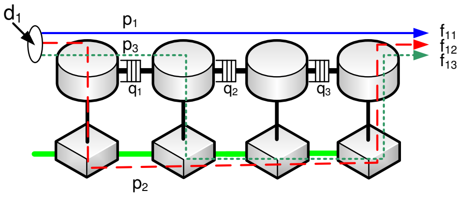


Figure 4.3: Risk Computation Example

Replacing in equation 4.2, the exponential risk function for the end-to-end path is:

$$
\begin{aligned}
R(\vec{e}, x^{e2e}) \;=\;& K_c \left[\sum_{j=1}^{M} 2e_j + R_{\text{cost}} \left((M+1)(N_1 - e_1) + \sum_{j=2}^{M-1} (M-j+2)(N_j - e_j + e_{j-1}) \right) \right] \\
& - K_u \sum_{j=1}^{M} \lambda e^{\frac{\sigma_j^{2\;e2e}\lambda^2 - 2\mu_j^{e2e}\lambda}{2}} \qquad (4.18)
\end{aligned}
$$

4.1.4 Illustrative example

Figure 4.3 depicts an example in which there is a single demand (d_1). Let us assume that d_1 is served using 3 different paths (p_1, p_2 and p_3). The p_1 risk contribution is due to the delay in q_1, q_2 and q_3 and the electronic cost. However, the p_2 risk is due to the cost of the optical resources. The risk added by p_3 consists on electronic and optical cost and the delay in q_1. Let us remind that the electronic cost is compute in each hop and the optical cost depends on the path length as it was defined in section 4.1.2.

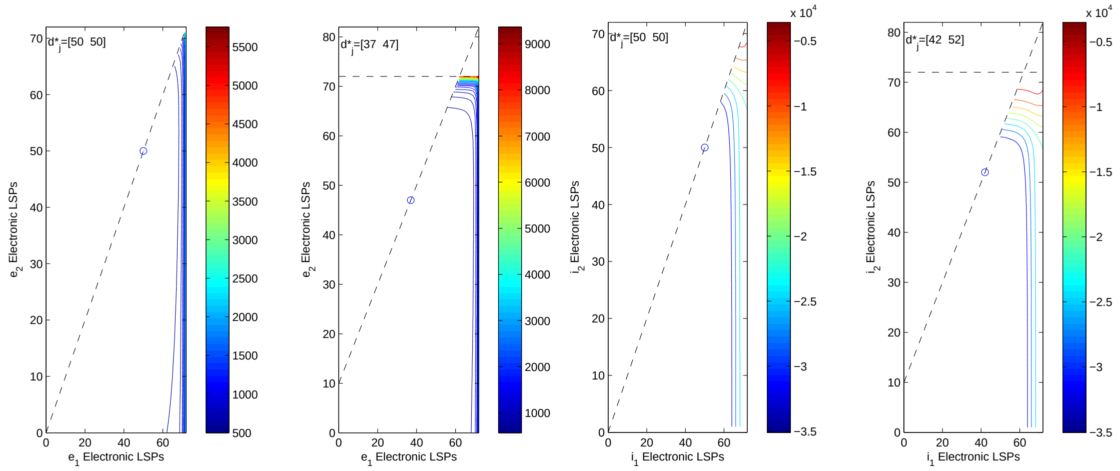
4.2 Numerical results and discussion

Next, the Bayesian decisor is evaluated in the scenario depicted in Figure 4.1 with $M = 3$ hops. We assume the following parameter values: 2.5 Gbps of lightpath capacity fed with standard VC-3 LSPs of 34.358 Mbps bitrate each. The number of incoming flows in the last node is $N_3 = 0$. The bandwidth standard deviation is chosen such that $\frac{\sigma}{m} = 30\%$ and the Hurst parameter selected is $H = 0.6$ (according to [61]). The $T_{\max}$ value is set to $80ms$ for $U_{\exp}$ and $T_{\max} = 5ms$ for U_{step}, since the QoS restrictions are more stringent in the latter case. The value of $R_{\text{cost}} = 2$ by default. K_c and K_u are constants that define the decision when the system operates at maximum network load (that is, $N_{max} = \lfloor C/m \rfloor$). Thanks to these constants it is possible to set the occupation of the optical and electrical link in the worst case. In our numerical experiments, for $N_{max} = \lfloor C/m \rfloor = 72$ incoming LSPs, the hop-by-hop electronic connection transmits 70% of the traffic, that is 50 LSPs. This policy can be adjusted by the network operator as necessary.

4.2.1 Basic behavior of the algorithm

The level curves of the Risk function help us to see how the function changes with the incoming traffic. Figure 4.4(a) (left) shows the U_{mean} level curves for $N_1 = 72$ and no cross traffic at node 2 ($N_2 = 0$). Since this is the normalization working point, the algorithm decides to send 50 LSPs through the IP layer. Figure 4.4(a) (right) illustrates the decision when node 2 injects some cross traffic, more specifically $N_2 = 10$ LSPs. In this situation, the decisor changes its behaviour by sending $o_1 = 35$ LSPs through the optical layer from node 1 to the destination node, which gives $e_1 = 72 - 35 = 37$ LSPs through the electronic domain. These 37 LSPs are added to the $N_2 = 10$ offered at node 2, which are transmitted electronically to the destination node.

It is worth noting that, since node 2 is closer to the destination node than node 1, its QoS restrictions are more permissive than if the same amount of traffic was offered at node 1.



(a) U_{mean} (b) U_{step}

Figure 4.4: Level curves examples without cross-traffic (left) and with cross-traffic (right)

Figure 4.4(b) displays the decision of the U_{step} function when there is no cross-traffic (left) and when there are 10 incoming LSPs in the second node. This happened with U_{mean} function, the decisor reduces the amount of traffic in the electronic layer to minimize the risk function. The level curves of the Risk function for U_{exp} are not included for brevity purposes, but the next experiments examine the behavior of the decision algorithm for such utility function.

4.2.2 Decisor dynamics experiment

Once the basis of the multi-hop algorithm is settled, the next experiments show its behavior when the traffic load is increasing. The experiments range from an empty network setup (no traffic) to congestion. We assume that there is a single lambda available between the first and the destination node. This is the reason why the maximum amount of LSPs is N_{max} in the optical paths.

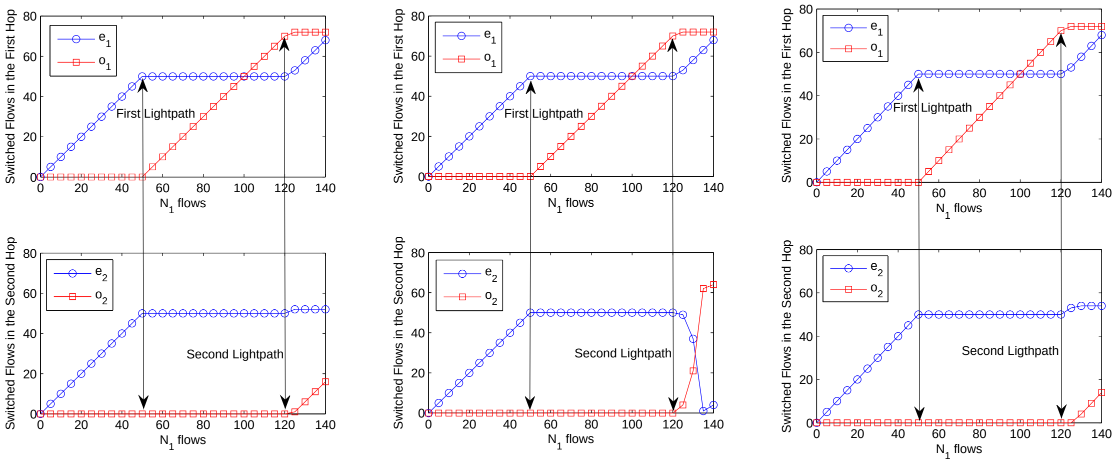
Traffic increment in the first node without cross-traffic

The next experiment shows how the algorithm changes its decision when the first node increases the amount of LSPs offered to the system and there is no cross-traffic ($N_2 = 0$). Figure 4.5(a) shows the number of flows sent through the electronic and optical domains at each hop, for the average delay utility case (U_{mean}). As shown, all traffic flows are sent through the IP layer until the utility given to the flows is smaller than the cost of establishing a new e2e connection, which occurs when $N_1 \geq 50$. At this point, a direct lightpath (first lightpath) from the first node to the destination node is created, as shown in the figure. After this, the network load keeps increasing (more LSPs offered at node 1) and, after some time (when the delay experienced at the second hop is excessive), a second lightpath at node 2 is created for incoming LSPs.

Figure 4.5(b) and 4.5(c) illustrates the same experiment but using the U_{step} and U_{exp} utility functions instead. The U_{exp} utility function behaves very similarly to U_{mean}. However, when the U_{step} is employed, the system forces all electronically-switched LSPs in the second node to be switched over the second lightpath, once this is created. The step utility is shown to be more QoS aware than both the exp and mean utility functions.

First node constant rate and second node load increment

This experiment evaluates the decision when the load offered to the first node is constant ($N_1 = 10$) and the second node sends a variable number N_2 of LSPs. Figures 4.6(a) depicts the amount of traffic sent using the electronic and optical layers in both hops. When the second node gets saturated (QoS degraded), the first node decides to send its 10 LSPs using



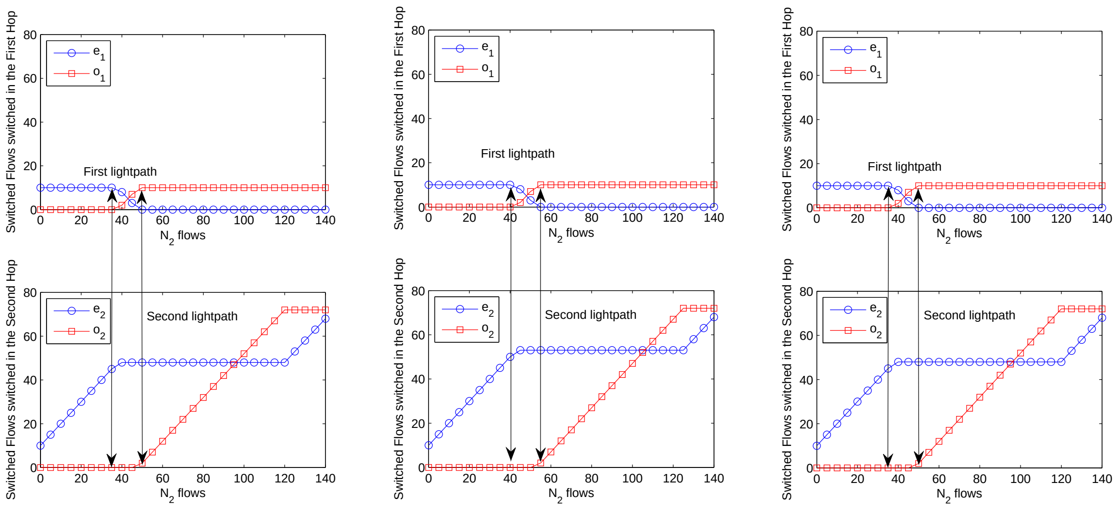
(a) Umean (b) Ustep (c) Uexp

Figure 4.5: LSPs sent through the electronic and optical layers at nodes 1 and 2 when the load in the first node increases

a direct e2e lightpath (first lightpath). It is worth noticing that a lightpath is created at the first node, rather than at the second node. This behaviour occurs thanks to the cost function, which favors the creation of long ligthpaths. However, since the traffic offered at node two N_2 keeps increasing, the bayesian decisor establishes a second e2e ligthpath at the second node. As it is depicted in Figures 4.6(b) and 4.6(c) this behavior is common to all utility functions.

4.2.3 Influence of the utilization cost (R_{cost})

Table 4.1 shows the optimal decision for different values of R_{cost} and the three utility functions. Remark that R_{cost} satisfies the condition: $R_{\text{cost}} > \frac{2*M}{M+1} = \frac{3}{2}$ $(M = 3)$ to make the cheapest (longest) lightpath more expensive than its electronic counterpart (further information in section 4.1.2). The results show that the more expensive the optical resources are (large values of R_{cost}), the fewer LSPs are routed using the optical domain, as expected. When U_{mean} and U_{exp} are used, the value of R_{cost} indeed decides the number of LSPs switched through each domain. For example, with U_{exp}, $e_1^* = 32$ LSPs are switched over the electronic layer for $R_{\text{cost}} = 1.6$, while for $R_{\text{cost}} = 3$, we have $e_1^* = 58$. On the other hand, the results obtained for the U_{step} function are different than for U_{mean} and U_{exp}. In this case, the decision does not vary significantly with respect to R_{cost} (Table 4.1), since the decision is mostly determined by the QoS parameters.



(a) Umean

(b) Ustep

(c) Uexp

Figure 4.6: LSPs sent through the electronic and optical layer when the load in the second node increases

		$N_1 = 60, N_2 = 0$			$N_1 = 60, N_2 = 10$		
		U_{mean}	U_{step}	U_{exp}	U_{mean}	U_{step}	U_{exp}
$R_{\text{cost}} = 1.6$	e_1^*	33	50	32	17	41	16
	e_2^*	33	50	32	27	51	26
$R_{\text{cost}} = 2$	e_1^*	50	50	50	37	42	37
	e_2^*	50	50	50	47	52	47
$R_{\text{cost}} = 3$	e_1^*	58	51	58	54	49	54
	e_2^*	58	51	58	55	52	55

Table 4.1: Optimal decisions with the variation of the R_{cost} parameter

Let us compare this results with the behavior of the decisor in a single-hop scenario. Section 3.3.4 shows that the R_{cost} parameter fixes the working point in a single-hop scenario. However, the QoS restrictions applied by U_{step} function makes them more important than the resources cost due to the nature of "hard-real time" applications. This behavior fits better with the desired performance in a multi-hop scenario. However, section 3.3.4 shows that the R_{cost} parameter has a lower influence on U_{step} utility than on the U_{mean} or U_{exp} functions. We can say that this R_{cost} influence avoidance remains in the utility function itself.

4.2.4 Study of delay QoS threshold (T_{max})

This section presents the decision results for changing T_{max} for offered traffic $N_1 = 60$ and $N_2 = 10$ fixed. As previously stated in Section 4.1.1, the QoS parameter (T_{max}) is only introduced for the elastic (U_{exp}) and hard-real time (U_{step}) utility functions. Therefore, U_{mean} is not studied in this section.

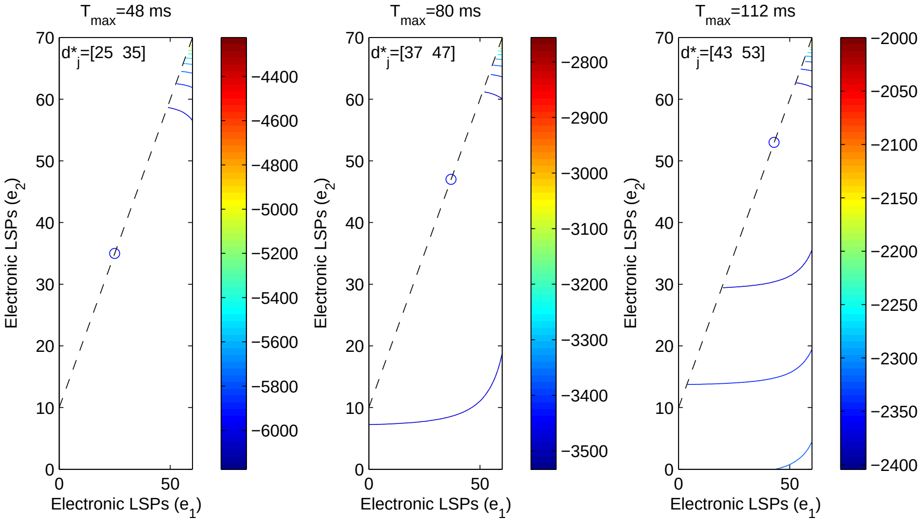


Figure 4.7: Variation of the T_{max} parameter (U_{exp})

Figure 4.7 illustrates the optimal decision for U_{exp} when T_{max} varies from $48ms$ to $112ms$, which is a 40% of variation from $80ms$. In light of the results, we observe that the network operator can tune the number of LSPs to be sent through the optical layer by changing T_{max} value. If flows are subject to coarser QoS constraints, the Bayesian decisor sends more LSPs over the electronic layer.

For the U_{step} function (Figure 4.8), the results are the following: for $T_{\text{max}} = 3ms$ is

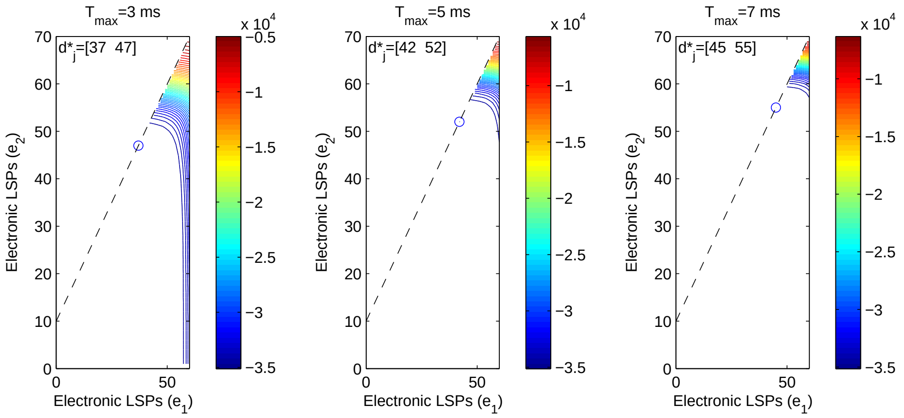


Figure 4.8: Variation of the $T_{\max}$ parameter (U_{step})

$\vec{e} = \{37, 47\}$, for $T_{\max} = 5ms$ is $\vec{e} = \{42, 52\}$ and for $T_{\max} = 7ms$ is $\vec{e} = \{45, 55\}$. The variation from $3ms$ to $7ms$ is a 40% from $5ms$ to make a fair comparison. Table 4.1 showed that U_{step} is not very sensible to R_{cost}, but it is to $T_{\max}$ (the QoS parameter). The reason is that this parameter is related to the e2e QoS performance experienced by the LSPs.

4.2.5 Hurst parameter

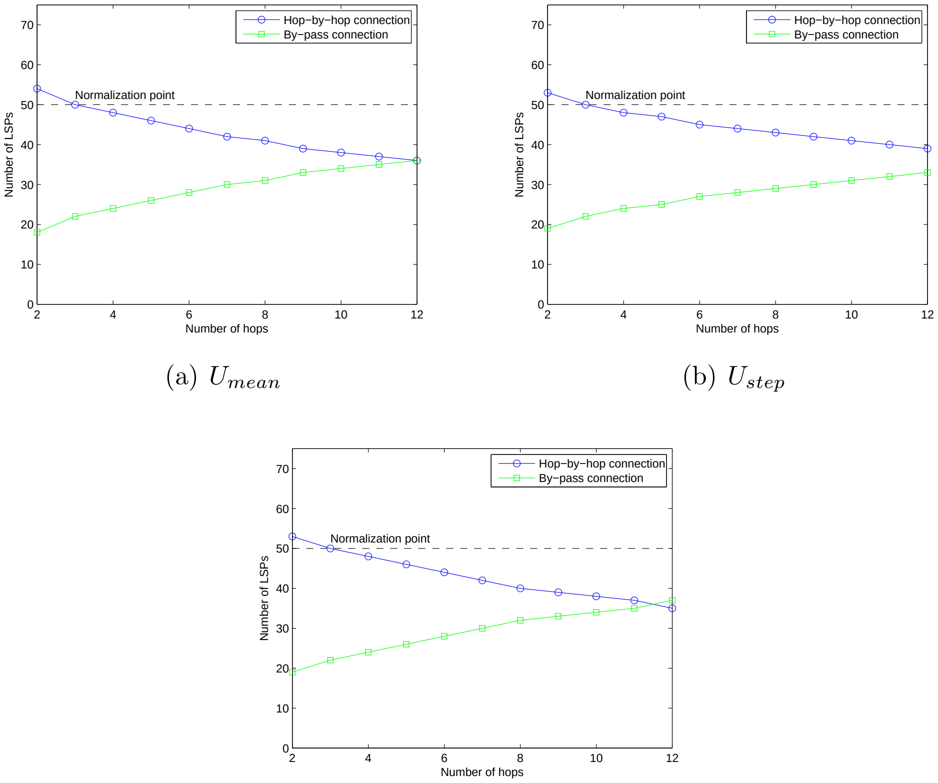
Traffic self-similarity is a well-known property of the Internet traffic. However, depending on the aggregation level and the network topology, the Hurst parameter changes [61]. Therefore, it is advisable to revise its impact on the multi-hop decisor. Table 4.2 shows the influence of the H parameter in the optimal decision. As shown in the table, a substantial increase of H always affects the optimal decision, but specially when U_{step} is used. This is expected since the value of H affects the variability of delay and the utility function U_{step} is more delay-sensitive than the other two.

		$N_1 = 60, N_2 = 0$			$N_1 = 60, N_2 = 10$		
		U_{mean}	U_{step}	U_{exp}	U_{mean}	U_{step}	U_{exp}
$H = 0.55$	e_1^*	46	43	44	37	35	36
	e_2^*	46	43	44	47	45	46
$H = 0.6$	e_1^*	50	50	50	42	42	42
	e_2^*	50	50	50	52	52	52
$H = 0.65$	e_1^*	54	56	54	45	47	45
	e_2^*	54	56	54	55	57	55

Table 4.2: Optimal decisions with the Hurst variation

4.2.6 On the influence of the path length in the decision

Previous experiments yield with a $M = 3$ hops scenario. The network delay varies when the path length is increased, so the decisor sends less traffic at the IP layer when the number of hops increase. This experiment sends $N_1 = 72$ LSPs, while the remaining nodes do not inject traffic. Figure 4.9(a) shows the variation of the decision when the path length increase. It is plotted the amount of traffic sent with the hop-by-hop and the by-pass connection. When $M = 3$, the amount of LSPs is the defined amount by normalization. Section 4.2 defines the amount of LSPs sent using the hop-by-hop connection to 50 LSPs.



(a) U_{mean}

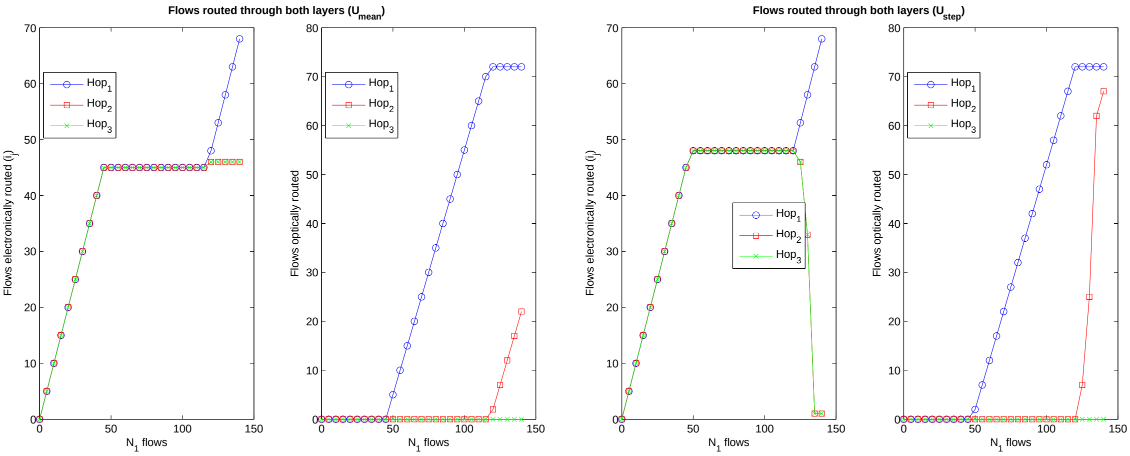
(b) U_{step}

(c) U_{exp}

Figure 4.9: Variation of the decision when the path length changes

Figure 4.9(b) and 4.9(c) depicts the results for U_{step} and U_{exp}. The decision in the normalization point remains at the same point. When the path length becomes longer, the amount of traffic send using the electronic layer is reduced, while the remaining traffic is transmitted via the by-pass connection. All utilities transfer almost the same amount of traffic at the IP and optical layer when the number of hops is 12.

Figure 4.10(a) depicts the decision taken in each hop when the amount of traffic N_1 is increased in a $M = 4$ scenario. We can see that as there is no cross-traffic the decision $(e_i \ , o_i)$ is the same along the path. The amount of traffic is lower than 50 (normalization point). The same experiment is shown in Figure 4.10(b) but with U_{step} function. U_{exp} results are not displayed here, but they are similar to U_{mean} results.



(a) U_{mean} (b) U_{step}

Figure 4.10: LSPs sent in each hop when the load in the first node increases (left) electronic layer (e_i) (right) optical layer (o_i)

4.3 Conclusions

This chapter builds over the previous single-node model a functions definition to deal with the utilization of the electronic and optical layers in a multi-hop scenario with multi-layer capable routers. Essentially, this includes the definition of a multi-hop Bayesian decisor which decides the amount of traffic routed through the optical and electronic domains, and its behavior is explained. The experiments show that thanks to the $T_{\max}$ and R_{cost} parameters, the network operator is provided with a means to define QoS and Relative optical/electronic cost aware metrics, which can be further applied to define traffic engineering mechanisms.

The path-oriented approach of this mechanism is an advantage, since it can yield into a distribution path risk-aware algorithm. However, the first step is to evaluate the Bayesian decisor in a full network topology, on attempts to define a full risk-oriented planning mechanism.

Chapter 5

Evaluation of the Bayesian decisor in real networks

This chapter studies the behavior of the Bayesian decisor in a full network topology. Previous chapters have shown the benefits of this model in a single-node (chapter 3) and a multi-hop scenario (chapter 4). The next step is to carry out a deeper analysis of the performance behavior of the Bayesian algorithm in a full network topology. The experiments will reveal a similar compromise between QoS and resource utilization as in the previous cases.

Once again, the Bayesian algorithm tries to find the minimum in a Bayesian Risk function which gives the number of LSPs that goes through the optical and electronic domains. This minimum is found in two ways, following a complete optimization approach and a heuristic-based algorithm, which are both shown to output the same solution on small, medium and large-size topologies. These experiments show the potential of the Bayesian approach to deal with the problem of trading-off between the resources and the QoS provisioning problem.

The remainder of this chapter is organized as follows: section 5.1 redefines the Bayesian optimization problem for a full network topology. Then, section 5.2 proposes the two approaches to solve the MTE problem. Section 5.3 compares both solutions and shows the performance of the algorithm in full network topologies. Section 5.4 explains in detail the behavior of the algorithm. Finally, section 5.5 concludes with remarks of the algorithm.

5.1 Risk definition in the full network topology

The Bayesian decisor deals with the problem illustrated in Figure 5.1. As stated in section 2.3, MTE problems are defined by five blocks: (1) traffic demand, (2) network equipment, (3) objective, (4) MTE module and (5) network configuration. For our problem, (1) we assume a static traffic matrix which is explained in detail for each experiment in section 5.4. The network equipment are multi-layer capable routers (2), which minimize the risk function (3) using the Bayesian decisor (4). Finally, the amount of traffic sent through each connection is given, as well as the required light-paths between the nodes (5).

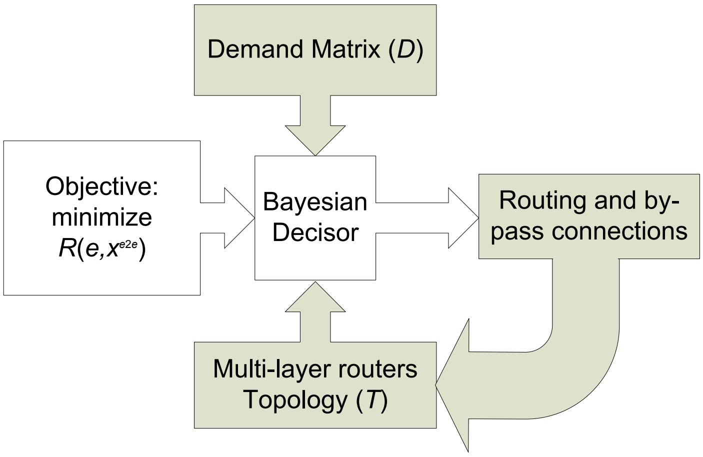
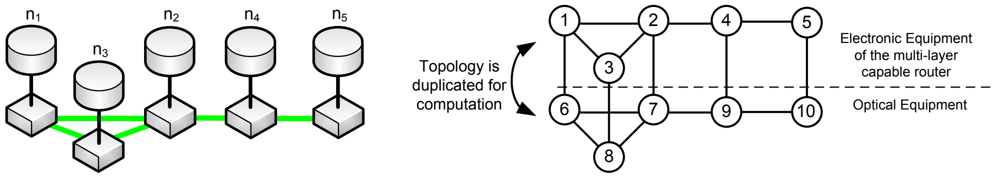

Demand Matrix (*D*)
Objective: minimize $R(e,x^{e2e})$
Bayesian Decisor
Routing and by-pass connections
Multi-layer routers Topology (*T*)

Figure 5.1: Blocks Definition Bayesian Decisor Problem

Following this, section 5.1.1 describes how the network topology is defined and section 5.1.2 expands the definition of the risk function to a full network.

5.1.1 Topology Definition

Let us consider a network topology $T = [V; L]$ consisting of a vertex set V and an arc set L (see [41] or section 2.5.1). It is assumed that each vertex represents a multi-layer capable router. Each arc $(x, y) \in L$ has associated a non-negative real number $c(x, y)$, which refers to its capacity. Let us define D as the traffic demand matrix, where each value associated to the position (i, j) is the traffic demand from node i to node j. Figure 5.2(a) illustrates a network with five multi-layer capable routers. Each node is a multi-layer capable router, so its logical topology has not only the IP but also the optical layer resources. Let us define the logical topology as $T' = [V'; L']$. Figure 5.2(b) shows the topology duplicated which is used for the risk function computation.


n_1
n_3
n_2
n_4
n_5
Topology is duplicated for computation
1
2
3
4
5
6
7
8
9
10
Electronic Equipment of the multi-layer capable router
Optical Equipment

(a) Physical topology (T)

(b) Logical topology (T')

Figure 5.2: Five node network

Over such topology (T'), we assume that a set of paths $p = 1, 2, \ldots, P_{dp}$ is defined. The aim of this work is to give an open framework to solve the multi-layer switching problem, but not to define which paths are the more suitable for computation. The network administrator can define such set of paths based on administrative motivations, and a routing algorithm can be run to give this information or all possible paths among the nodes can be defined in the set p [62].

A portion of such paths are hop-by-hop connections and the rest of them uses end-to-end ligthpaths, so some demands (D) are routed through end-to-end ligthpaths, while the rest of them are sent using hop-by-hop electronic connections. Let us identify the optically served demands as the matrix O and the electronically ones as E.

5.1.2 Redefining the Risk function

Section 4.1.1 defined the risk function (eq. 4.2) for the multi-hop scenario and the contribution of the cost and utility functions. In the multi-hop scenario it is very clear the relationship between the number of optically routed flows ($\vec{o}$) and the number of incoming flows $\vec{N}$ and the electronically routed flows $\vec{e}$ (see Figure 4.1). However, in a whole network this computation is not straightforward, because end-to-end paths may share a segment in the network. Therefore, to clarify notation and avoid confusion the O matrix is included in the risk function definition. The computation of the utility contribution in the full network is performed on a per-path basis. Each path computes its utility which is added to the global risk function. The reason for this is that the utility depends on the end-to-end delay (x_p^{e2e}), so it is necessary to calculate the risk function in that path-oriented way. The risk function for the full network is defined as follows:

$$R(E,O) = K_c C_T(E,O) - K_u \sum_{d=1}^{D} \sum_{p=1}^{P_d} \mathbb{E}_x \left[U(x_p^{e2e}) \right], x_p^{e2e} \geq 0 \tag{5.1}$$

Note that the delay in the electronic layer (x_p^{e2e}) just depends on the number of LSPs in the electronic domain (E). It is possible that a given path traverses three nodes using the IP layer and then establishes an optical by-pass to reach its destination. For instance, Figure 5.2(a) shows the five-node network topology example. Let us assume that there is only one demand (h_1) from node 1 to node 5. To satisfy such a demand, several paths can be defined: $p_1 = \{n_1, n_2, n_4, n_5\}$, $p_2 = \{n_1, n_6, n_7, n_9, n_{10}, n_5\}$, $p_3 = \{n_1, n_2, n_7, n_9, n_{10}, n_5\}$, $p_4 = \{n_1, n_3, n_2, n_4, n_5\}$, $p_5 = \{n_1, n_6, n_8, n_7, n_9, n_{10}, n_5\}$, $p_6 = \{n_1, n_3, n_8, n_7, n_9, n_{10}, n_5\}$ and $p_7 = \{n_1, n_3, n_2, n_7, n_9, n_{10}, n_5\}$. Let us remind that the resources in the last hop are the same using the IP and the optical domains. Therefore, a hypothetical path $p_8 = \{n_1, n_2, n_4, n_9, n_{10}, n_5\}$ would use the same resources as p_1, so it is p_1. Figure 5.3 illustrates some of these paths.

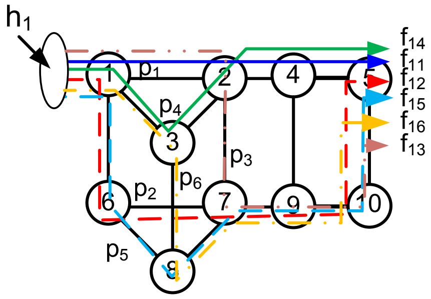


Figure 5.3: Path definition example

Let us explain in detail the contribution of the paths to the risk function. The contribution of p_1 is due to the electronic delay observed in queues q_{1-2}, q_{2-3}, q_{3-4} and q_{4-5} and the electronic cost. However, the risk of p_2 is only due to the cost of the optical resources. The risk added by the hybrid path p_3 consists on electronic and optical cost and the delay in q_{1-2}. Let us remind that the electronic cost is computed on each hop but the optical cost depends on the path length as it was defined in section 4.1.2.

5.2 Problem solution

This section proposes two solutions to this problem. The first of them based on defining the objective as the typical optimization problem that finds a solution using any of the well-known methods in the literature: Steepest-descent, Newton-Raphson, etc. Due to computation time and the optimization constraints that solvers add to the risk function, an heuristic algorithms is also proposed.

5.2.1 Optimization Problem

Given a demand matrix D, we want to allocate such demands over a topology T' using the set of paths p that minimizes the risk in the full network. This is a classical optimization problem and is formulated in Algorithm 2.

All assumptions of the optimization problem in Algorithm 2 are straightforward, only the upper limit to the number of flows needs to be clarified. Thanks to this restriction, it is assumed that there is only one free wavelength in the optical domain for bypassing nodes. It is possible to change such a restriction and allow all nodes to ask for as many wavelengths as necessary. The aim of this restriction is to show the behavior of this algorithm when resources are limited. Section 5.4 shows these results.

To solve this optimization problem we have used the "Active-set" algorithm of the Matlab optimization toolbox, which allows to compute the minimum without finding an analytical expression of the risk function gradient. Such a method is useful for convex functions like the U_{mean} function, but it is not suitable for the rest of utility functions. Moreover, the computation time of the optimization in large networks is excessive.

5.2.2 Heuristic Algorithm

The heuristic algorithm allocates every traffic demand, as if they randomly arrived to the system, sending them over the path that minimizes the risk among the total possible paths. Once one LSP is allocated, the next incoming LSP is evaluated the same way: it is sent lowest and the algorithm makes its decision accordingly using the path with the lower risk and so on. Note that the demands find the network with the previous LSPs allocated. The motivation of this solution is that this Bayesian decisor sends incoming LSPs based on the resources occupation and it can dynamically change its decision depending on the network state. The behavior is not changing randomly, so the differences between the decisions gradually change when the traffic increases. These assumptions are verified with

Algorithm 2 Optimization Problem

Indices:

- $d = 1, 2, \ldots, D$ demands
- $p = 1, 2, \ldots, P_{dp}$ paths
- $l = 1, 2, \ldots, L$ links

Constants:

- $\delta_{ldp} = 1$ if link l belongs to path p for the demand d; 0, otherwise
- $R(\cdot)$ risk function
- h_d volume of demand d
- c_l capacity of the link l
- N_{max} maximum number of LSPs in a wavelength

Variables:

- f_{dp} number of flows allocated to path p of demand d. Let us call f to the vector with all flow allocations ($f = [f_1, f_2, ..., f_{P_{dp}}]$)

Objective:

- Find $\;f^*$ such that $\;R(f^*) = \min R(f)$

Constraints:

- $0 \geq f_{dp} \geq N_{max}$
- $\sum_p f_{dp} = h_d$
- $\sum_d \sum_p \delta_{ldp} f_{dp} \leq c_l$

the dynamic behavior experiments carried out in section 4.2.2. Therefore, if the system allocates the traffic using the global network information, the minimum solution should be found. Section 5.3.2 validates this assumption with a comparison between the optimization and the heuristic solution . The pseudo-code is displayed in the Algorithm 3.

Algorithm 3 Heuristic Algorithm

```
while $traffic\_to\_allocate > 0$ do
  $d \leftarrow Generate\_random\_arrival$
  for $i = 1$ $to$ $p$ do
    $x_i$ $update\_load(p_i)$
    $R_{pi} = Compute\_System\_Risk(x_i)$
  end for
  for $i = 1$ $to$ $p$ do
    $p^* \leftarrow min(\vec{R_p})$
    $x^* \leftarrow update\_decision(p^*)$
    if $x^* \leq cl$ & $Ax^* < b$ then
      $x \leftarrow x^*$
      $break$
    end if
  end for
end while
```

5.3 Performance evaluation in a backbone topology

5.3.1 Scenario definition

This section presents some results concerning the behavior of the Bayesian decisor algorithm in a multi-layer network. The simulation scenario assumes that each wavelength is of 2.5 Gbps capacity, and that incoming LSPs are standard VC-3 LSPs (typically 34.358 Mbps each). Each node is connected by two wavelengths: one for hop-by-hop switching and the second wavelength for the end-to-end connections. The Hurst parameter has been chosen as $H = 0.6$ (according to [61]) and $\frac{\sigma}{m} = 0.3$. The constants K_c and K_u define the decision when the system operates at maximum network load (that is, when $N_{max} = \lfloor C/m \rfloor$). Thanks to these constants it is possible to set the occupation of the optical and electrical link in the worst-case scenario. In our numerical experiments, for $N_{max} = \lfloor C/m \rfloor = 72$ incoming LSPs, the hop-by-hop electronic connection transmits 70% of the total traffic, that is 50 LSPs. This policy can be adjusted by the network operator as necessary.

The K-shortest path algorithm is computed to provide the set of paths to serve the demands, using the number of hops as metric, and with maximum number of shortest paths $K = 4$. The set of all possible paths include hop-by-hop paths and end-to-end optical connections. Figure 5.4 shows the topology of the 15-node NSFNET network.

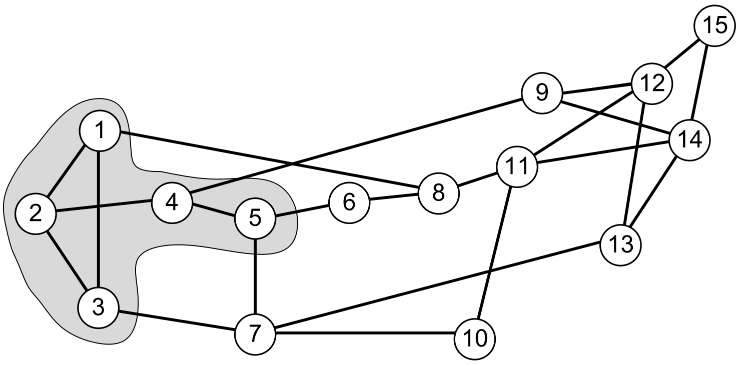
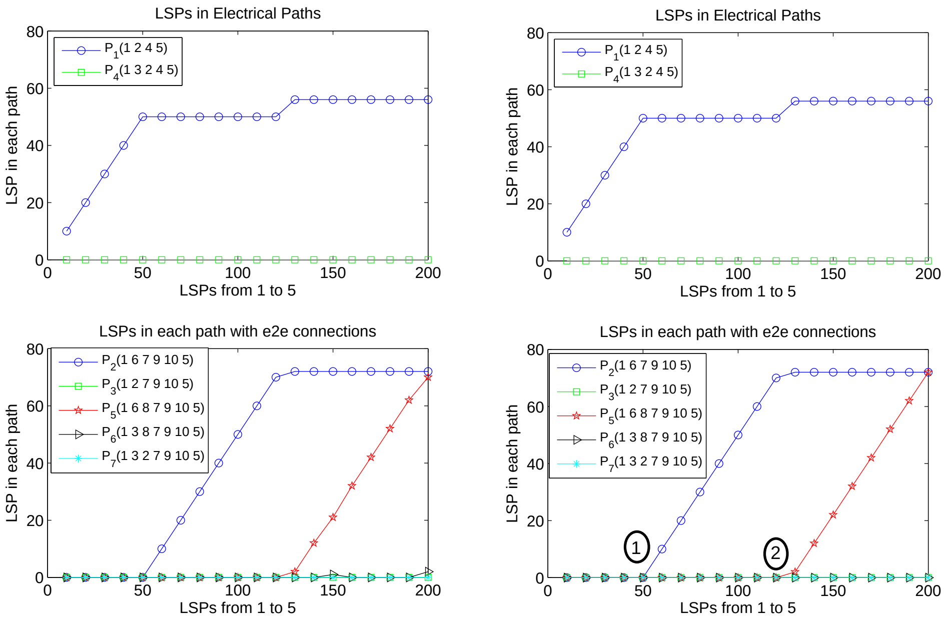


Figure 5.4: 15-node NSFNET topology

5.3.2 Validation of the algorithms

Section 5.2 has defined the heuristic algorithm and the optimization problem. The optimization problem in a convex function (like the U_{mean} function) assures that a global minimum is found. Figure 5.5 shows an experiment which is solved using the heuristic algorithm and the optimization algorithm. Let us assume only the nodes from 1 to 5 in the NSFNET network (Figure 5.4). This set of nodes coincides with the five-node network (Figure 5.2(a)). This experiment assumes that the traffic demand (h_1) from node 1 to node 5 increases with time an amount of LSPs ranging from 10 to 200 LSPs. The decisor starts sending the traffic using the IP resources (p_1), because this cheapest path in terms of risk. When the amount of traffic at the IP path (p_1) reaches 50 LSPs, an end-to-end lightpath is established (p_2). This is the status point 1 displayed in Figure 5.5. Figure 5.6(a) shows the paths used when the status point 1 is reached. At this point, there are 50 LSPs in the path p_1, while path p_2 starts transmitting LSPs. Let us remind that the amount of 50 LSPs is predetermined by the normalization constants. When the optical end-to-end lightpaths (p_2) is full ($N_{max} = 72$ LSPs), the system arrives to the status point 2 (Figure 5.5). At this point, a third path (p_5) is established, when not only the IP but also the optical end-to-end lightpath (p_2) gets saturated. In this situation there are three paths being used to transmit traffic from node 1 to the node 5, as depicted in Figure 5.6(b). These results coincide with the results of the previous chapter 4.



(a) Optimization Problem (b) Heuristic Algorithm

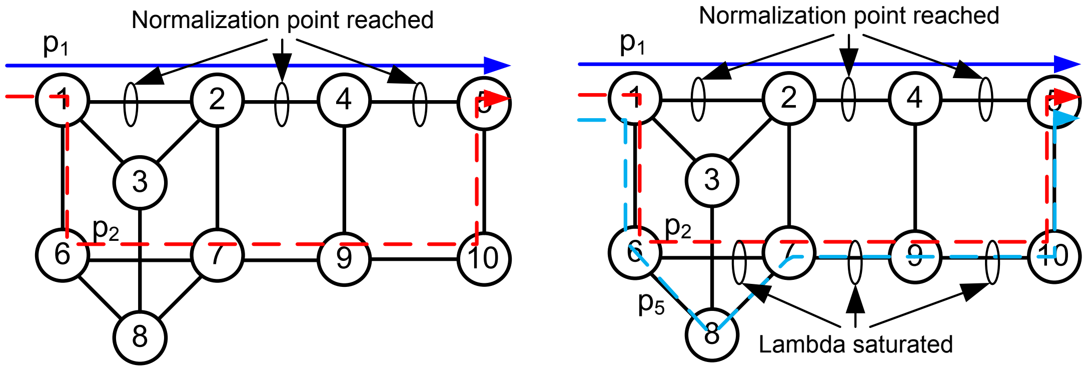
Figure 5.5: Decision when the traffic is increasing from node 1 to 5 (Umean)

In light of the above results, both optimization and heuristic approaches shows the same behavior, so the heuristic algorithm finds the global minimum like the optimization algorithm does. The following considers the heuristic algorithm in all the simulations to show the results.

5.3.3 Performance with the resource utilization cost variation (R_{cost})

The following experiment evaluates the LSPs allocation in the NSFNET network (Figure 5.4) when varying the R_{cost} value. To carry out the experiment, we have assumed a random uniform VC-3 demand matrix, which is scaled from [0,9] to [0,108].

Figure 5.7 shows the amount of paths used by the U_{mean} utility when R_{cost} varies and the traffic load is increased. As we can see (Figure 5.7 top-left) the amount of electronics paths changes depending on the R_{cost} parameter used. Moreover, the amount of paths remains similar while the traffic load increases. This means that the Bayesian decisor uses the IP paths first whenever possible and only uses the optical lightpaths when the IP layer is saturated. This behavior fits with the idea of first using the already deployed IP layer if



(a) Working point 1 (see Figure 5.5)

(b) Working point 2 (see Figure 5.5)

Figure 5.6: Paths used when the traffic load increases

possible, since it is cheaper in terms of risk. The same paths are used given a R_{cost} value, but they do not transport the same amount of traffic. Figure 5.7 (bottom-left) shows the mean occupation of the IP layer links. The U_{mean} function gradually fills in the paths. At the beginning, when the traffic load is still low, the links occupation is low too, but, progressively, this mean occupation becomes higher. On key feature of the Bayesian decisor is that it does not saturate the IP layer. This remains at medium load levels, but they are never set congested (it does provide QoS).

However, although the number of electronic end-to-end paths remains constant with the traffic increment, the amount of optical by-passes increases with the traffic load (Figure 5.7 top-center). At the beginning, just a few optical end-to-end connections are used, but when the IP layer becomes saturated and can not accept more LSPs, new lightpaths are required to provide a proper service to the traffic demands. This behavior remains similar compared with the previous results in simpler networks or in the multi-hop scenario. It is worth noticing that the number of partly optical and electronic paths referred to as hybrid paths is very small. These paths are used when the hop-by-hop and direct end-to-end connections are saturated. Furthermore, as the traffic matrix is uniform the amount of traffic between one-hop destinations saturates the links when the traffic becomes high. Such reasons imply that its number is minimum. The occupation of the hybrid paths is not displayed, since they used the lambdas of the electronic and the optical layer. Therefore, neither its occupation is the occupation of the IP layer, nor the occupation of the by-pass.

Once the U_{mean} function has been is analyzed, the performance of the U_{step} case is studied next. Its results are displayed in Figure 5.8. Like the U_{mean} function, the amount of paths used at the IP layer are constant when the traffic load increases. Moreover, the

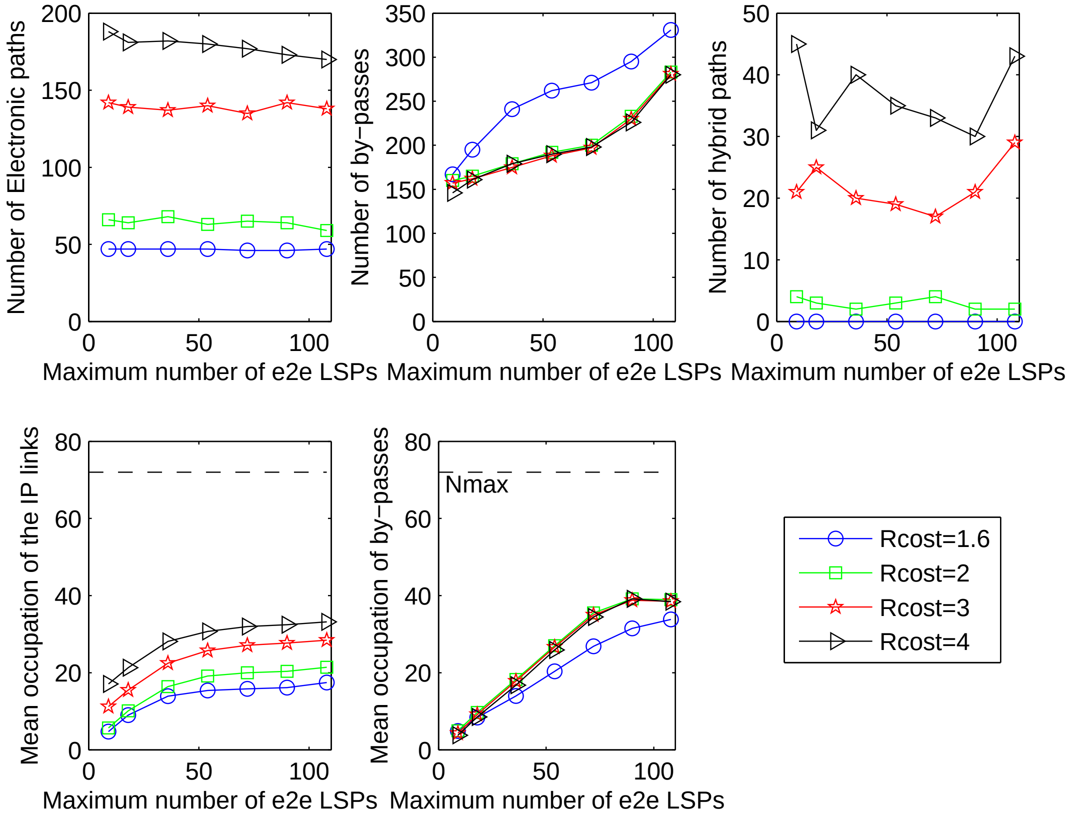


Figure 5.7: Number of paths used in each domain and their mean occupation when R_{cost} varies (U_{mean})

amount of electronic paths is almost constant when the R_{cost} parameter varies. The number of IP layer connections for the U_{step} case does not change as much as for the U_{mean} case, only about 20 LSPs, whereas the U_{mean} case this difference was about 150 LSPs. The amount changes, but just 20 paths, while the variation of number of paths with the U_{mean} function is 150 LSPs. This behavior fits again with the numerical results explained in chapter 4. The U_{step} function sets a given working point depending on the QoS rather than the other parameters. Again the amount of optical paths increases with the traffic, but the amount of LSPs remains similar for all R_{cost} values. The number of hybrid by-passes is higher in the U_{step} case. Section 4.2.2 showed that when there is cross-traffic in the network and the U_{step} function is used, the decisor sends the incoming LSPs through the IP layer until they reach a congested node. At this node, the traffic is optically switched to the destination node. This behavior fits with the increment of the number of hybrid connections.

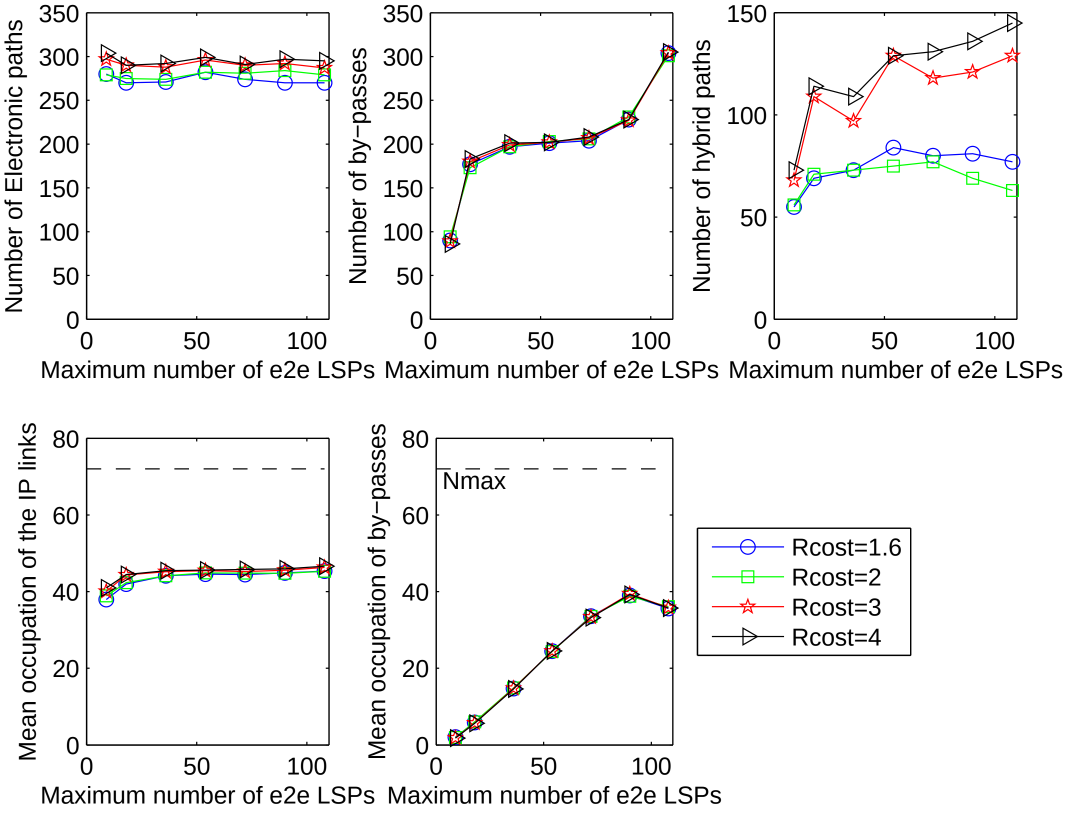


Figure 5.8: Number of paths used in each domain and their mean occupation when R_{cost} varies (U_{step})

Concerning the path load, the occupation at the IP layer clearly remains constant in all cases in order to provide the desired QoS (see Figure 5.8 bottom). On the other hand, the optical links' occupation increases when the amount of traffic becomes greater, since the traffic is transmitted optically, since the IP layer is congested. Again the occupation does not depend on the R_{cost} parameter for this utility function.

The results for the U_{exp} function are shown in Figure 5.9. Its behavior is similar to the U_{mean} function, but the difference between $R_{\text{cost}} = 1.6$ and 2 are smaller. Both cases use the same amount of LSPs at the electronic domain. The number of optical and hybrid connections show the same behavior than at the U_{mean} case. The main difference between both utility functions is the link occupation. While the U_{mean} function takes into account the mean delay, the objective of the U_{exp} function is to provide a QoS to every service, thus filling the links more gradually.

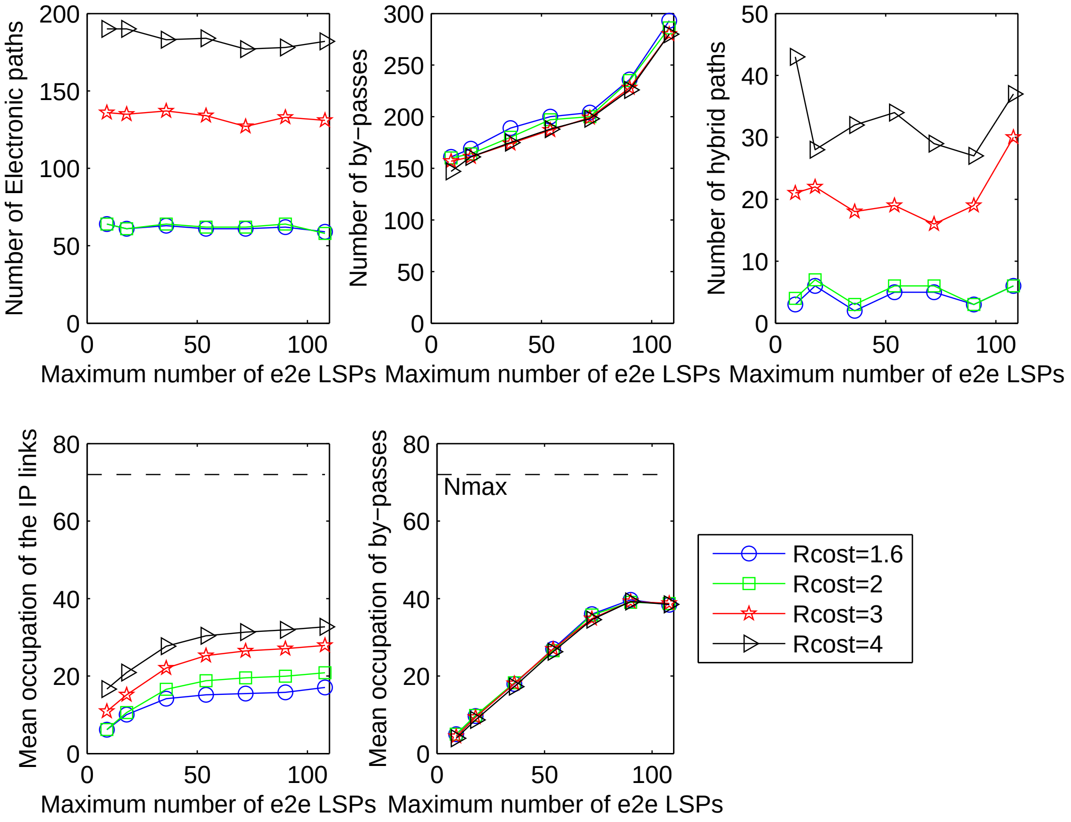


Figure 5.9: Number of paths used in each domain and their mean occupation when R_{cost} varies (U_{exp})

Let us review the path length in the three cases. Table 5.1 summarizes the mean length of the paths in each domain for all simulations. The mean is just shown, because there is almost no variation of the average path length with the traffic load increment. We may notice that the electronic paths are shorter than the optical or hybrid paths. This is because the one-hop demands can just send the traffic in the electronic domain, there is no choice of by-passing intermediate nodes. For instance, when R_{cost} is 1.6, only the one-hop demands are established at the electronic domain, while the others demands use the optical domain.

	Electronic paths			Optical paths			Hybrid paths		
	U_{mean}	U_{step}	U_{exp}	U_{mean}	U_{step}	U_{exp}	U_{mean}	U_{step}	U_{exp}
$R_{\text{cost}} = 1.6$	1.0152	2.4570	1.7021	2.7620	2.7701	2.7479	–	3.3487	4.0238
$R_{\text{cost}} = 2$	1.2816	2.4973	1.7043	2.7324	2.7661	2.7362	3.4286	3.3247	4.0408
$R_{\text{cost}} = 3$	1.9278	2.5556	2.0630	2.7226	2.7724	2.7206	3.1972	3.3994	3.2214
$R_{\text{cost}} = 4$	2.1184	2.5865	2.2470	2.7287	2.8160	2.7227	3.1689	3.4820	3.1868

Table 5.1: Path length with the variation of the R_{cost} parameter

When the R_{cost} parameter is 1.6, the electronic paths becomes shorter than when its value is 3. The reason is that the optical resources cost is cheaper, thus enhancing their utilization. Moreover, when $R_{\text{cost}} = 1.6$, the design rule of $R_{\text{cost}} > \frac{2 \cdot M}{M+1}$ is not fulfilled for all path lengths, becoming the end-to-end ligthpaths cheaper than the hop-by-hop connections. The length of the optical resources does not change with the R_{cost} variation. The optical paths length is fix by the network topology, but the end-to-end connections are used at lower or higher rates depending on the R_{cost} parameter. As previous results depicted, the utilization of the electronic layer is higher for the U_{step} case and, consequently, the average length of the electronic paths is higher than at the U_{mean} or U_{exp} cases. The length of the hybrid paths highly varies with the traffic load, but it is caused because of the low amount of hybrid connections. If the utilization of new hybrid paths highly impact on the average length.

5.3.4 Performance with the QoS parameters variation (T_{max})

The following experiment shows the results when the T_{max} parameter changes. Figure 5.10 illustrates the results for the U_{step} function. The amount of traffic sent using the IP layer changes based on the QoS requirements. When the delay is assured, the U_{step} function finds the optimal working point, but when these requirements are not fulfill new end-to-end ligthpaths are established offloading the IP layer traffic. The amount of electronic paths used is around 270 in all simulations and the mean link occupation does not vary when the traffic load increases. When the QoS requirements are more demanding, the path occupation is lower (see Figure 5.10 bottom-left). Table 5.2 shows the path length in each experiment, but the average path length does not change for the U_{step} case.

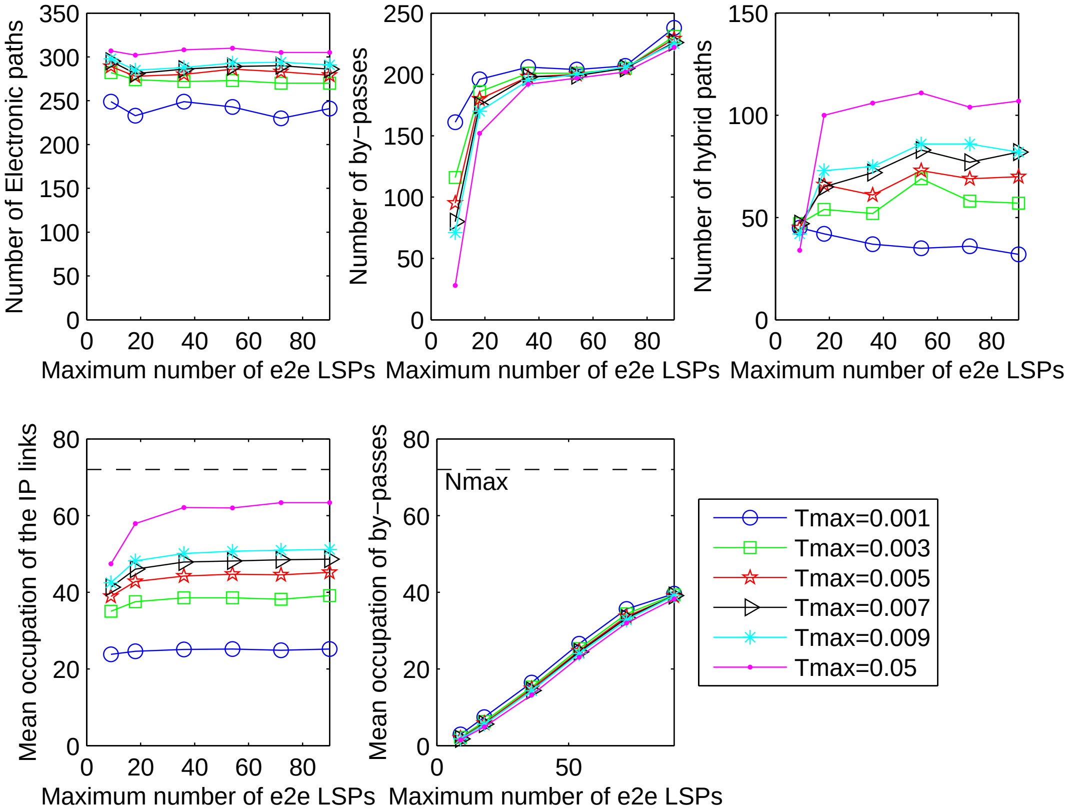


Figure 5.10: Number of paths used in each domain and their mean occupation when $T_{\max}$ varies (U_{step})

The U_{exp} function highly changes the amount of used paths when the QoS features changes. When $T_{max} = 0.8s$, more than one hundreds paths are used at the IP layer, but when $T_{max} = 16ms$ such amount is just around 60. Table 5.2 shows that the electronic path length is similar for the 16, 48 and 80ms cases. This behavior is similar to the U_{step} case, where the optimal paths are used and no new ones are established. The 16, 48 and 80ms cases differs on the mean occupation of the electronic paths (see Figure 5.11 bottom-left). When the QoS constraints are relax, the system increases the number of electronic paths and its occupation (see Figure 5.11 top and bottom-left).

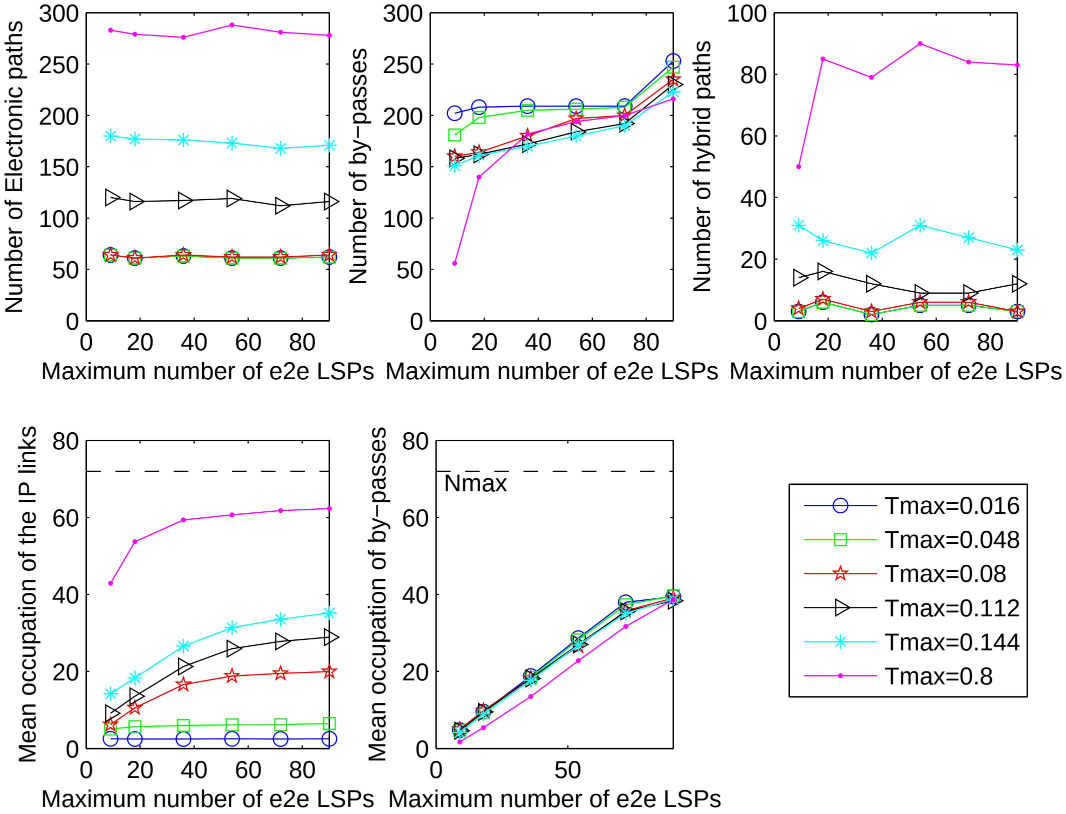


Figure 5.11: Number of paths used in each domain and their mean occupation when $T_{\max}$ varies (U_{exp})

The amount of optical connections highly changes when there is a low load situation, but when the network becomes saturated, the same amount of optical connections are used. However, the mean occupation of them is the same when the T_{max} parameter varies, except for the $800ms$ case, where the occupation is lower. The average optical length is similar for all cases (see Table 5.2).

	Electronic paths		Optical paths		Hybrid paths	
	U_{step}	U_{exp}	U_{step}	U_{exp}	U_{step}	U_{exp}
$T_{\max} - 80\%$	2.4255	1.7147	2.7853	2.8208	3.2383	3.9722
$T_{\max} - 40\%$	2.4855	1.7147	2.7757	2.7679	3.2939	3.9722
$T_{\max}$	2.5049	1.7182	2.7753	2.6922	3.3021	3.9921
$T_{\max} + 40\%$	2.5197	1.9184	2.7642	2.6699	3.3249	3.4444
$T_{\max} + 80\%$	2.5286	2.1498	2.7633	2.6574	3.3459	3.2652
$T_{\max} \times 10$	2.5666	2.4680	2.7534	2.7246	3.4086	3.4444

Table 5.2: Path length with the variation of the $T_{\max}$ parameter

5.4 Detailed explanation of the Bayesian decisor behavior

In the previous section, we validated the performance of the Bayesian decisor algorithm in a full network. Previous experiments have given a general idea of how the algorithm uses the network. Now, a detailed explanation of the algorithm is done, showing its capabilities to carry out Multi-layer Traffic Engineering. The Bayesian decisor is based on some parameters related with the resources cost (R_{cost}), QoS constraints ($T_{\max}$) and the features of the incoming traffic (mean rate, self-similarity, ...). The decisor is designed to change its decision based on the variation of such input parameters. The traffic features can change during a given period of time or the QoS restrictions could be changed for a certain premium customer. This section evaluates the impact of such variations in the decisor.

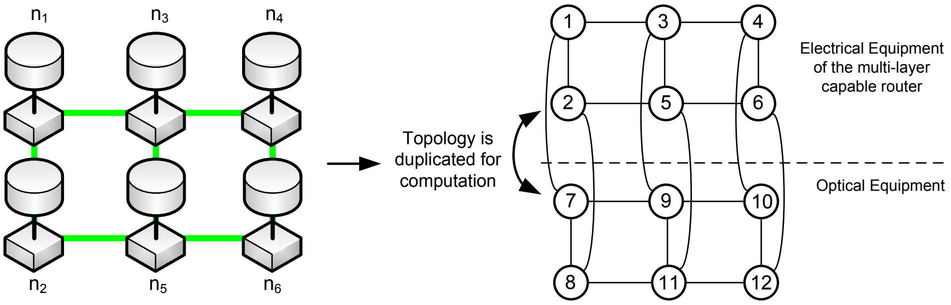

n1
n3
n4
n2
n5
n6
Topology is duplicated for computation
1
3
4
2
5
6
7
9
10
8
11
12
Electrical Equipment of the multi-layer capable router
Optical Equipment

Figure 5.12: Eight Network Topology

Figure 5.12 shows the “Eight” network which is used for the following experiments, in order to show the performance in multi-path situations. The LSPs parameters are set to the default values defined at the beginning of section 5.3.1.

5.4.1 Utilization cost parameter variation (R_{cost})

The R_{cost} parameter is the relative cost of the optical service in comparison with the electronic ones. The cost of N LSPs routed using a hop-by-hop connection is $K_c \times k \times 2 \times N$, while if they are switched as an end-to-end connection, this cost is $K_c \times R_{\text{cost}} \times k + 1 \times N$. Let us remind that $R_{\text{cost}} > \frac{2 \cdot M}{M+1}$, where M is the maximum number of hops in the network. The longest path on the Eight network is $M = 3$ (Figure 5.12), so $R_{\text{cost}} > 1.5$. Consequently, Figure 5.13 shows the decisor behavior when R_{cost} is set to 1.6, 2 and 3 and the traffic from node 1 to 6 is increasing from 10 to 200 LSPs.

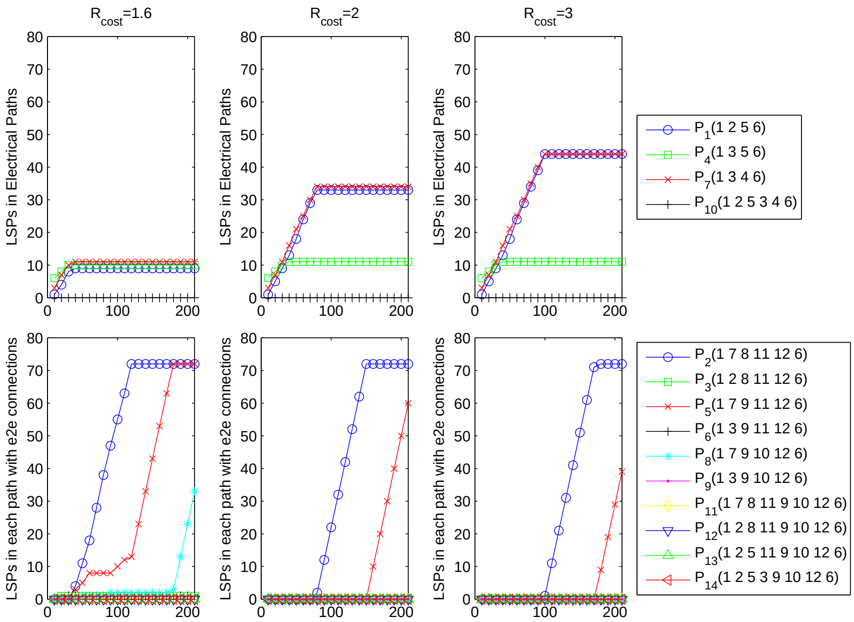
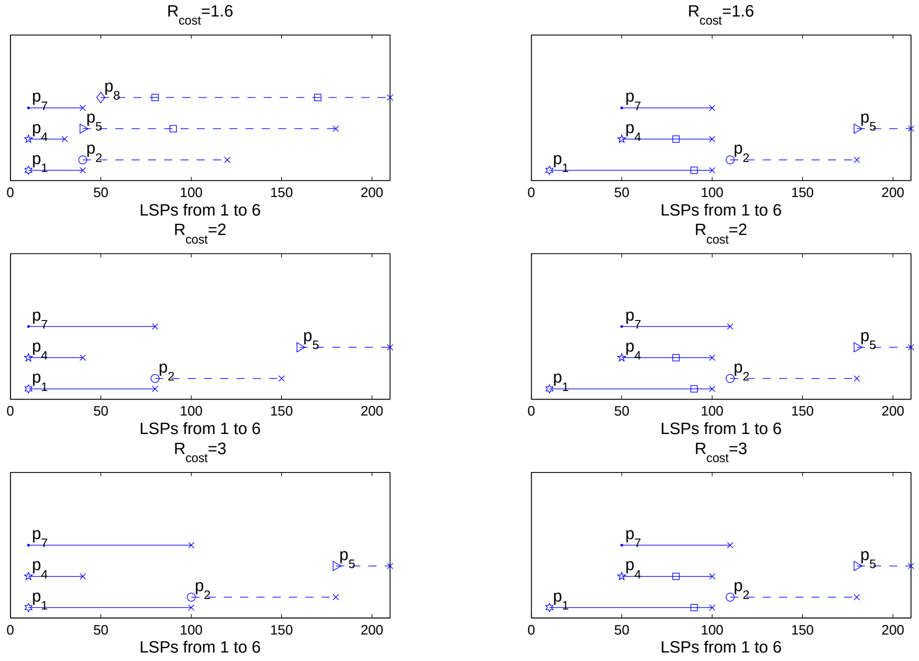


Figure 5.13: Decisor elections when R_{cost} parameter varies (U_{mean})

The number of LSPs using hop-by-hop connections (p_1 and p_7) increases when the optical resources are more expensive (Figure 5.13 right). There is a third hop-by-hop path which is used p_4. This path uses the third shortest path from node 1 to 6 (1-3-4-6). When there is no traffic in the network the link between node 3 and 4 is unused. Consequently, the utility is high, thus reducing the total risk. Figure 5.14(a) depicts when each path starts to transmit traffic (with the name of the path p_i), when there is a step increment (□) and when the path no longer increment its traffic (×). The solid lines are the electronic connections, while the dashed lines are for the optical connections. Figure 5.14(a) shows that at the beginning, the three shortest paths share the demands from node 1 to 6. The first path that stops transmitting more connections is p_4, whose route is overlapped by p_1

and p_7 routes. When p_4 does not accept any more traffic, the end-to-end connections (p_2 and p_5) are used for $R_{\text{cost}} = 1.6$. Nevertheless, when R_{cost} is greater, p_2 and p_5 paths are not used until p_1 and p_7 are loaded. In light of the results, we can see that when the optical resources cost is lower the lightpaths are used sooner and there are more lightpath requests. When $R_{\text{cost}} = 1.6$, an extra by-pass path is used (p_8). Moreover, when $R_{\text{cost}} = 1.6$, the demands are burstly added to the optical paths. We can see that the paths p_5 and p_8 are not linearly filled, but with steps. Figure 5.14(a) indicates this situation with □ symbol. The results for the U_{exp} function are not displayed, because they are similar to the ones of the U_{mean} case, avoiding the step increments at the optical paths.



(a) U_{mean} (b) U_{step}

Figure 5.14: Paths time-line when R_{cost} parameter varies

Figure 5.15 illustrates the results for the U_{step} function. The behavior of the shortest paths is different than in the U_{mean} case. The decisor first just sends the traffic in p_1, when this path is full, it uses the third shortest path p_7. Again p_7 is used just to transmit and small amount of traffic, since 1-3 and 5-6 links are heavily loaded by p_1 and p_4. and finally p_2 and p_4. For the U_{mean} case the electronic paths are filled linearly. However, for the U_{step} case the electronic path p_1 accepts traffic until 43 LSPs are transmitted and, then, p_4 is used. However, when the LSPs routed through p_7 reaches that amount of p_1, they start to share the incoming demands. This behavior is clearly seen in Figure 5.14(a) thanks to the □ symbol in the path p_1. Concerning the optical paths, when the IP layer reaches congestion levels, p_2 is established. When it can not accept more demands, p_5 is used as a second lightpath. The positions where the electronic and optical paths are created for the U_{step} case are very similar for all R_{cost} values (Figure 5.14(a)). In light of the results, the U_{step} function is not so dependent on the variation of R_{cost}. The reason again is that the QoS constraints reduces the cost impact. The decision changes but only in some LSPs (Figure 5.15).

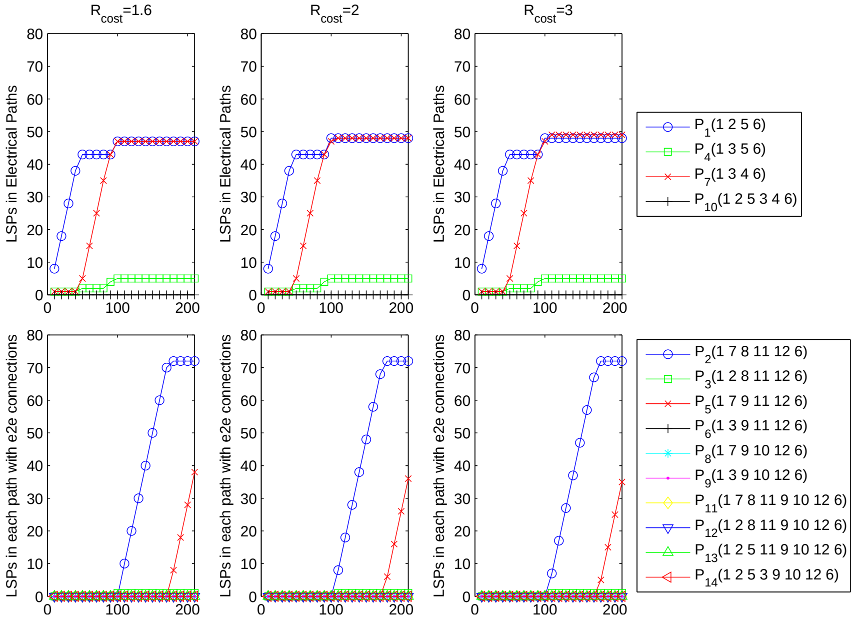


Figure 5.15: Decisor elections when R_{cost} parameter varies (U_{step})

5.4.2 QoS parameter variation ($T_{\max}$)

The $T_{\max}$ threshold determines the QoS expected by the elastic and real-time services. Hard real-time services achieve no utility when they exceed the $T_{\max}$ value, while exponential utility decay (λ) is defined so the 50% of the utility is achieved before the threshold (section 4.1.3). When timing constraints are tighter, the decisor requires more optical resources to serve the LSPs. Consequently, if the load was kept in the electronic domain, the utility achieved by the LSPs would be nill. This logical behavior is achieved by the decisor in Figure 5.16 and Figure 5.18. These figures show the decision changes when $T_{\max}$

threshold is varied a 40% up and down for the U_{step} and U_{exp} functions. The U_{mean} function does not have any QoS restriction, so it is not studied in this section.

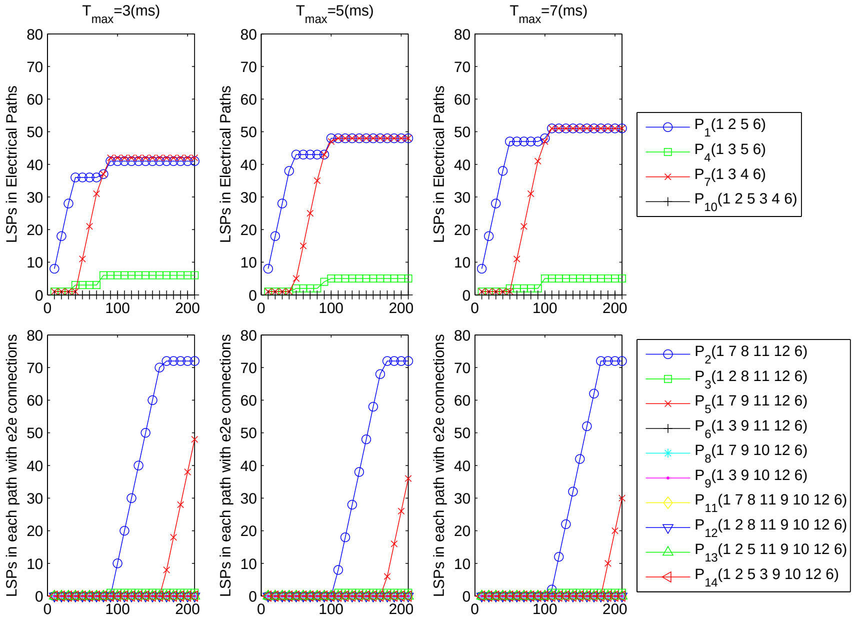
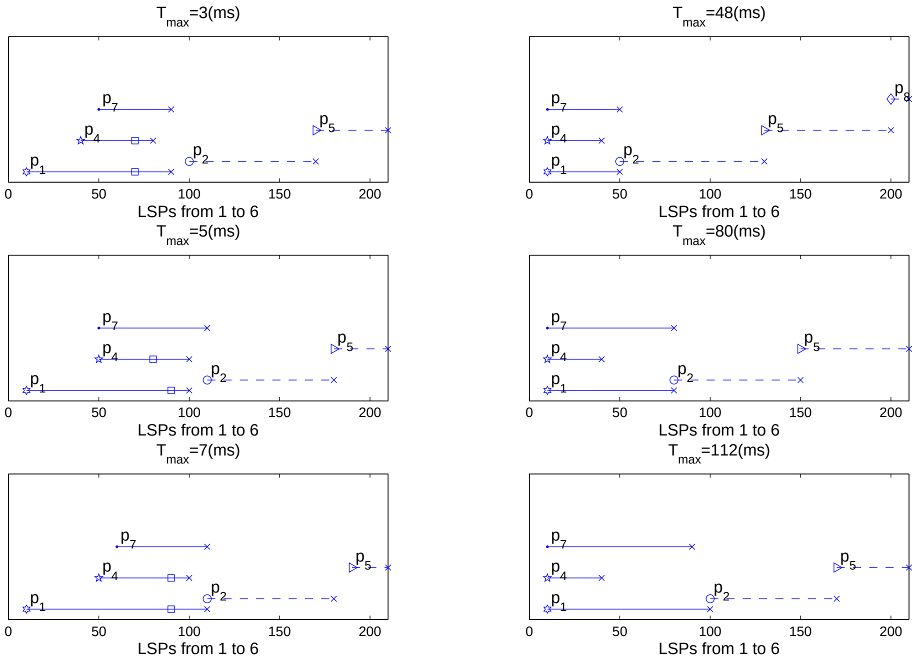


Figure 5.16: Decisor elections when $T_{\max}$ parameter varies (U_{step})

In light of the results, we can see that as the QoS constraints are critical for the U_{step} function: the LSPs receive utility or not. Therefore, depending on the QoS values a maximum number of LSPs in a link will give utility to such LSPs. If this number is exceeded no utility is received. Consequently, the number of LSPs in the primary path (p_1 and p_4) is constant. For the first threshold ($T_{\max} = 3ms$), 41 LSPs is the optimal number of circuits in the electronic layer, while the higher one ($7ms$) sets this value to 51 LSPs. Again the paths are used at similar load levels in all simulations (Figure 5.17(a)). Depending on the QoS parameter, the electronic layer becomes useless sooner or later and, consequently, the hop-by-hop paths (p_1, p_4 and p_7) stops accepting traffic.

The U_{exp} function is more flexible and the $T_{\max}$ parameter variation changes the decision. Instead of a variation of 10 LSPs in p_1 and p_4, the decision for the U_{exp} case differs on 20 LSPs from the minimum ($T_{\max} = 48ms$) to the maximum ($112ms$). This means that the utility perceived by the flows is not so QoS dependent. The behavior showed by Figure 5.18 is similar to the decision made in the U_{step} case (see Figure 5.13). However, the impact of $T_{\max}$ threshold is higher for the U_{exp} case. According to the results in Figure 5.17(b), we can see that the starting utilization point of the first optical path (p_2) varies from 50 to 100, depending on the QoS parameters.



(a) U_{step}

(b) U_{exp}

Figure 5.17: Paths time-line when $T_{\max}$ parameter varies

Finally, let us remark that this behavior is in accordance with the results of the section 5.3.4. The U_{step} function provides an optimal working point for the incoming traffic and it can not vary the optimal decision because of the binary nature. On the other hand, the U_{exp} function allows to vary the decision based on the QoS parameters.

5.5 Conclusions

This chapter validates the results of the Bayesian decisor in a complete network. The problem is defined as an optimization problem and also with an heuristic algorithm. Both methods are compared providing us a mechanism to solve the optimal flow allocation in risk terms. Firstly, the algorithm is studied in a complete network, showing global performance results. We have assessed that the algorithm does not change its properties when the system is more complex, but it is able to dynamically share the optical and electronic resources.

The main conclusion is that thanks to the decisor parameters, the operator can tune the

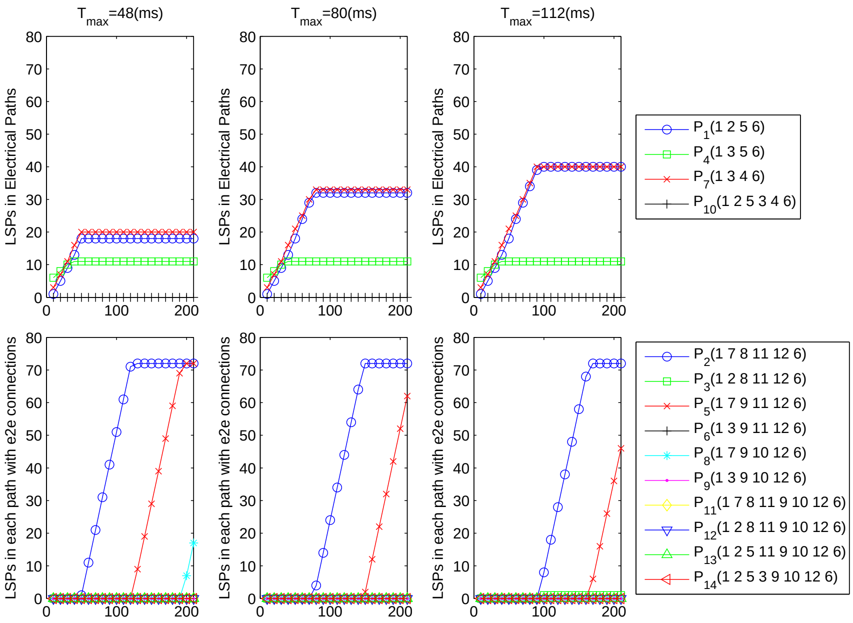


Figure 5.18: Decisor elections when $T_{\max}$ parameter varies (U_{exp})

network utilization, in order to provide an adequate QoS to the customers. The different utility functions can help the operator to deal with multiple services.

Chapter 6

Analysis of the Bayesian decisor in a multi-domain scenario

This chapter studies the extension of the Bayesian decisor to multi-domain networks. The interaction between multiple domains is possible thanks to the definition of ASON and GMPLS and the Path Computation Element (PCE) architecture. Such advances allow the establishment of end-to-end optical and electronic connections. Previous chapters dealt with the total network delay information, but when different operators connect their networks, global information may not be available to all domains. Consequently, the delay model assumed in the previous chapters must be changed. This chapter reviews some end-to-end delay monitoring algorithms found in the literature and show applicability to single-domain scenarios. Once the signaling between domains and the monitoring algorithms are validated, we show that using the Bayesian decisor methodology it is possible to adapt our Multi-layer Traffic Engineering (MTE) problem for multi-domain environments.

The remainder of this chapter is as follows: section 6.1 introduces the signaling approaches for information exchange in multi-domain networks. Section 6.2 gives an overview of the delay monitoring algorithms and it evaluates their performance for characterizing the end-to-end delay. Section 6.3 redefines the Bayesian decisor for a multi-domain scenario and shows a few numerical cases. Finally, section 6.4 outlines the main contributions of this chapter.

6.1 Control plane information exchanged in multi-domain networks

The deployment of intelligent Next Generation Transport Networks (NGTN) opens a new problem: how such intelligent networks have to exchange information among the neighbor network domains administrated by other operators. Each operator's network is known as an Autonomous System (AS). Such AS's or domains can be defined not only based on ownership reasons, but also based on geographical or administrative reasons. Figure 6.1 depicts a multi-domain scenario with four AS's. Every AS has border routers that are connected via inter-domain links with other domains. However, not every router must be connected with other domains, but it can be an interior router that it is just connected to routers in its own domain (see AS 3 and 4 in Figure 6.1). Operator A is an example of multiple AS's administrated by the same operator.

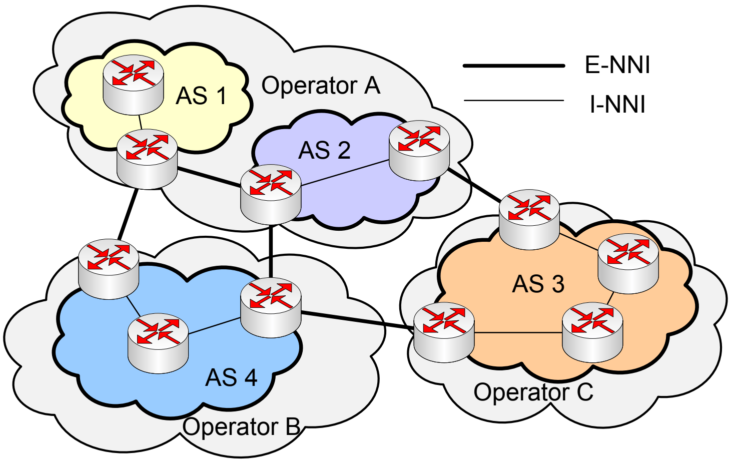
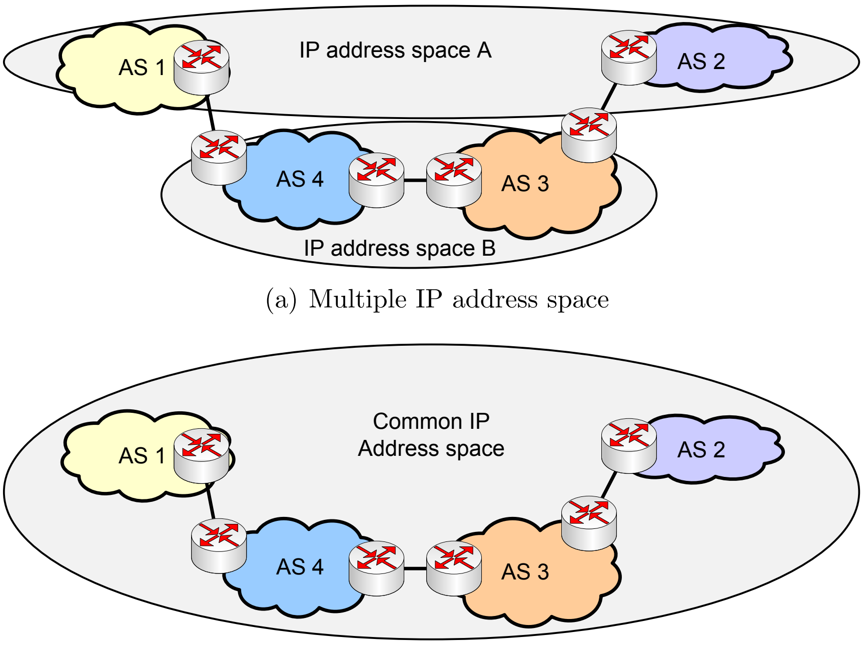


Figure 6.1: Multi-domain Scenario: Autonomous Systems description

According to ASON's nomenclature [8], the inter-domain interface is called External Network-Network Interface (E-NNI), while the intra-domain interface is called Internal Network-Network Interface (I-NNI). In order to manage the connections among the domains, it is essential to exchange information in the Network-Network Interface (NNI) interfaces. There are two main functionalities that must be supported by NNI interfaces: routing and signaling. The following sections give an overview of the main solutions for routing and signaling in multi-domain networks. Let us remark that research projects, like DRAGON [63], define architectures to deal with control plane issues in multi-domain environments and they have carried out test-beds in multi-domain and multi-layer networks.

6.1.1 Routing in multi-domain networks

Intra-domain routing and signaling is provided by GMPLS framework [7] (see section 2.2.1). Inter-domain information exchange is more complicated since there are competing operators which may decide not to exchange complete information concerning its resources and network state for security and privacy reasons. However, some control information is required, so the domains can find where the rest of the network destinations are located. The Border Gateway Protocol (BGP) is the most common protocol for this inter-routing issues [64]. Usually, BGP is not the intra-domain routing protocol, but OSPF or IS-IS. Consequently, such intra-domain protocols collect the information of the domain and they send it to the border routers, which exchange this information using BGP. Figure 6.2 depicts how the IP addresses may be configured in multi-domain networks. Depending on the technologies of the domains, the IP address set is shared or separated.



(a) Multiple IP address space

(b) Common IP address space

Figure 6.2: IP address space in multi-domain networks

Figure 6.2(a) illustrates two separated IP address domains, where AS 1 shares its IP address set with AS 2 and AS 3 and AS 4 share their own IP address space. If AS 3 and AS 4 are optical networks that just provide connectivity to upper client domains (AS 1 and AS 2), this separation of IP domains is useful to decrease the IP space address size and complexity. This view also fits with the overlay model of ASON. On the other hand, Figure 6.2(b) shows a share IP address among all domains. Traffic Engineering (TE) information for BGP is supported in recent RFCs [65, 66]. Such extensions were rejected in the past by the IETF, because of scalability issues, but this problem has been partially solved by BGP [67].

6.1.2 Signaling in multi-domain networks

ASON allows multiple domains to support multiple services (packets, TDM, circuits) and multiple vendors. Moreover, it is not mandatory to know the signaling protocols of all domains, but just to support E-NNI interfaces. Consequently, a multi-session procedure travel through multiple E-NNI (and UNI) interfaces in each domain to establish a given connection domain per domain. If the end-to-end domains support RSVP-TE [15], a single RSVP session can be established among all interfaces. Figure 6.3 depicts the interaction between the domains when GMPLS is supported by all domains, or when there are different intra-domains signaling protocols. The authors in [67] carry out a test-bed to show the interoperability between the domains of ASON and GMPLS in six different locations. There are open issues in such topic, but this test-bed show that multi-domain operation is a reality. The authors in [68] not only achieve the interconnection between domains but also between layers.

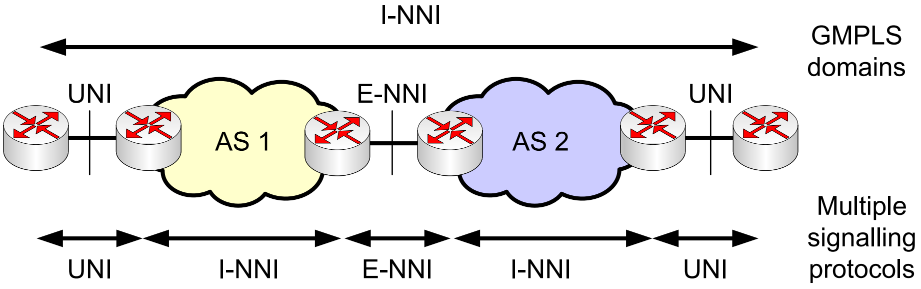


Figure 6.3: End-to-end signaling in multi-domain networks

A second standard for multi-domain signaling is Path Computation Element (PCE) [69]. The PCE architecture delocalizes the path computation from the source router to the PCE that computes the path for the source router. A PCE can collaborate with other PCEs to improve the path computation without exchanging TE information between domains, thus solving the privacy issues. Figure 6.4 depicts the path establishment process in a PCE based architecture. The Path Computation Client (PCC) makes a request to the PCE for a new path establishment. Such PCC entities are network equipment like routers.

A PCE sends information to its neighbor PCE and so on. The message format of these requests and answers is defined in the PCE Communication Protocol (PCECP) [69]. The authors in [70] propose a service plane over the PCE architecture to support the automatic establishment of services and the definition of a business model for such architecture.

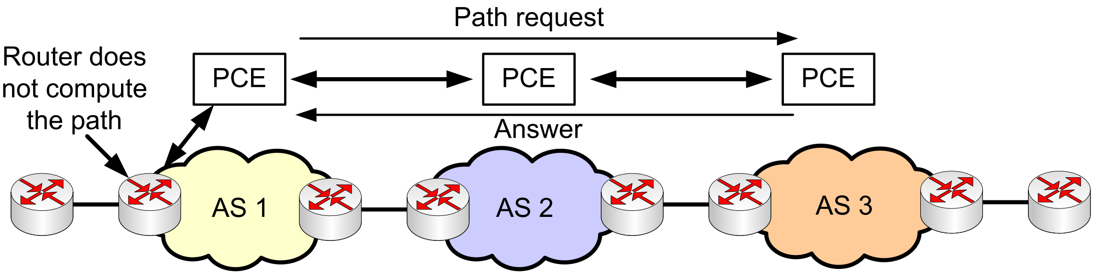


Figure 6.4: Path establishment in a PCE based architecture

There are two main approaches to the path establishment in multi-domain networks: per-domain and PCE-based [71]. Per-domain establishment calculates the paths in limited visibility environments, although if the systems support GMPLS signaling, an end-to-end computation is possible (see Figure 6.3). On the other hand PCE path computation is done by the each PCE in every domain, allowing an increased visibility [72].

6.1.3 Advances on multi-domain/multi-layer architectures

Multi-domain/multi-layer architectures are still a new research topic. According to the authors in [72]: "Despite standards progress, the overall area of multi-domain (multilayer) optical networking has not seen significant research focus". Nevertheless, the research community is improving some adjacent problems such as the information exchange in multi-layer (see chapter 2) and multi-domain networks. Let us remark that when we talk about multi-layer networks in this work, we are focus on IP over WDM architectures.

Some works have studied the routing mechanisms for optical networks, see for instance in [73, 74], which use a hierarchical approach to solve the multi-domain routing problem. The former work [73] is focused on the multi-domain routing in ASON networks, while the latter [74] analyzes the multi-domain routing in optical networks with a more detailed node description. Both approaches are based on node abstraction, so the information of a set of nodes is reduced and managed by the superior level in the hierarchy. This abstraction mechanism improves the scalability of the ASON architecture [73].

The first work in the multi-domain/multi-layer networks was done in [75]. The authors proposed a GMPLS compatible inter-domain routing algorithm for multi-layer and multi-domain networks. They assume all possible inter and intra-domain paths and they compare the performance of their algorithm with the case where only either electronic or optical connections are available. They use two metrics in the routing process which are: Bit Error Rate (BER) and number of hops. A hybrid metric, which is based on blocking probability terms, is also applied. They find that the multi-layer algorithm improves the performance achieved by single-layer algorithms. A approach based on a multi-segment framework is proposed in [76]. The authors define a multi-segment optical framework to solve the end-to-end provisioning of optical networks. They address the problem of multi-granularity (or multi-layer) transforming the topology information in one graph per type of service. Once these graphs are defined, they are interconnected with the multi-segment network representation. Although the approach is a multi-segment description they propose three routing algorithms: end-to-end (E2E), Concatenated Shortest path Routing (CSR) and hierarchical routing. E2E routing can be used if the control plane of the domains is compatible. This is similar to the GMPLS end-to-end approach introduced in section 6.1.2. The CSR routing process is done “segment-by-segment”. Hierarchical routing is done following a SP algorithm in each hierarchical domain. Previous works [75, 76] are based on SONET/WDM networks. The authors in [77] study the convergence of the Ethernet switches and the OXCs to support Ethernet over WDM. They propose a node architecture, which is interconnected using GMPLS and that can operate in ASON networks.

Not only the routing but also the management of such multi-domain/multi-layer framework needs to be studied. The authors in [78] introduce a web-service architecture for provisioning end-to-end optical connections. They add a service layer to the management system to allow such functionality. This idea of a service layer is also used by the authors in [70], but the latter is not centered in optical networks, but in the PCE path establishment. Web services are used also in the DRAGON [63] project to propose an architecture to deal with multi-service, multi-layer and multi-domain hybrid networks [79].

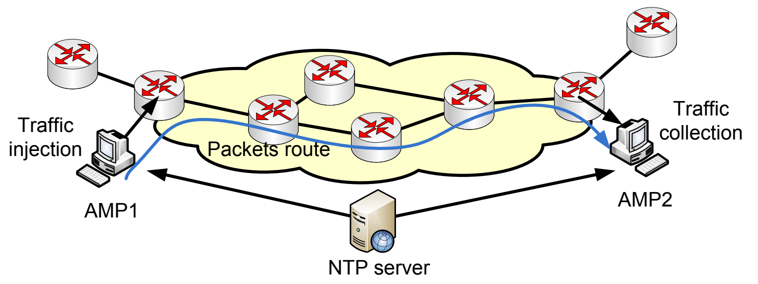
To sum up, let us remark that multi-domain/multi-layer networks is still a new research area where some issues are still open ([80, 67, 72]). However, the research community has done progress in this area which make feasible the establishment of both LSPs and optical lightpaths through multiple domains. PCE approach is proposed as a feasible architecture for problems where there is limited visibility, which is typically the case in multi-domain/multi-layer networks [81].

6.2 Algorithms to characterize end-to-end delay

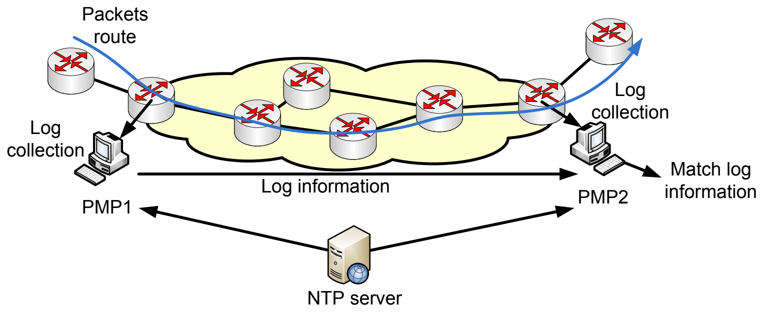
QoS measurements is a very active research field. There are several works focused on finding an optimal architecture to monitor the traffic inside a given network. However, such acquisition process needs a structured methodology [82]. Most of the efforts have been done in the IP, TCP and UDP protocols. There are two standardization organisms involved in the definition of QoS measurements metrics: the International Telecommunication Union (ITU) and the IETF IP Performance Metrics (IPPM) Working Group.

As chapter 3 defines, this work uses delay as the main QoS metric. One-Way Delay (OWD) is defined in [83] as the time elapsed since a packet leaves the source and it reaches the destination. Such metric and definition fits with the delay that we are using in our model. The ITU similarly defines IP Packet Transfer Delay (IPTD). Such packet monitoring process requires the timestamping of the packets at different Measurement Point (MP). In order to precisely acquire the real delay, it is mandatory to synchronize the measurement points. Measurement points are not located in the same location, so they use different clocks, which have small time variations that modifies the real network delay. Synchronization is not a straightforward process if the time scales are below micro seconds. There are two methods for time synchronization: software and hardware methods. Software synchronization is done using protocols like Network Time Protocol (NTP). The precision of these protocols is lower than hardware synchronization, but they are accurate enough for some applications. Hardware synchronization is usually done via Global Positioning System (GPS). When the information arrives to the monitoring equipment the timestamp is stored directly. There are mixed hardware software methods like Precision Time Protocol (PTP).

There are two main types of monitoring [84]: active and passive. Active monitoring estimates the network performance injecting synthetic traffic on it. On the other hand, passive measurements just collect information of the traffic that it is inside the network. Figure 6.5(a) depicts an example of an active measurement scenario, where there are two Active Measurement Point (AMP). AMP 1 injects traffic with a timestamp to the network and AMP 2 collects such packets and computes the OWD from AMP 1 to AMP 2. Figure 6.5(b) shows a passive measurement scenario. Passive Measurement Point (PMP) 1 collects information of packets that are traversing the network and PMP 2 does the same action. An information exchange is required to match the information collected in both monitoring points. The entity that receives this information (PMP 2 in Figure 6.5(b)) has to process the log of the remote monitoring points.



(a) Active



(b) Pasive

Figure 6.5: Measurement scenarios

The monitoring process is carried out to achieve some aggregate measurements of the network delay. It is not feasible to store the delay of all packets in a network and process them on real-time. The aim of monitoring protocols is to provide network administrator aggregated statistics of the traffic traversing the network. The most common measurements are the first moments: mean and variance. Besides, in order to avoid large delays in the network, percentiles are estimated to monitor how many times a given delay threshold is exceeded. If a more detail information of the delays is required, delays histograms can also be computed. They are known as One-Way Delay Distribution (OWDD) or IP Delay Variation Distribution (IPDVD). Such histograms provide a graphical hint of the probability density function of the delay.

Next sections evaluates four approaches to characterize end-to-end delay: mean delay estimation, percentile estimation, Fast-Fourier transform method [1] and the Weibull mixture model [60]. The latter two algorithms provide a detailed description of the delay, since they provide an estimation of the probability density function of the delay.

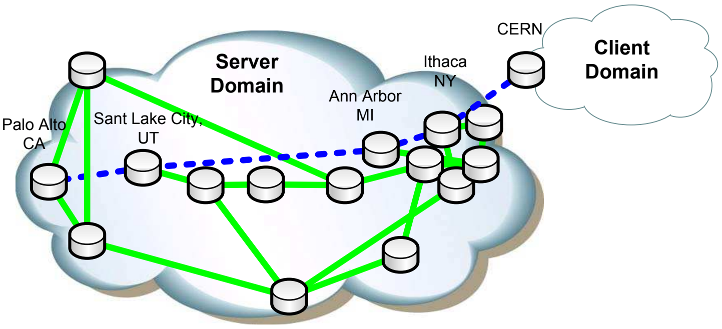


Figure 6.6: Network scenario - NSFNet T1 Backbone

6.2.1 Scenario definition

To validate the algorithms mentioned in this section, we use the scenario shown in Figure 6.7. Let us assume that NSFNet is a server domain, and that a delay estimation from Palo Alto's node to CERN's node is demanded by a client domain attached to the server domain. The path comprises five hops, using the shortest path criterion, as it can be seen in Figure 6.6.

In the above scenario, an end-to-end traffic flow (λ) is mixed with cross traffic (λ_x) at each hop, coming from other nodes in the network. The amount of cross traffic is assumed to be 90% of the link traffic. In what follows, a high load (70%) and low load (10%) network scenarios are analyzed.

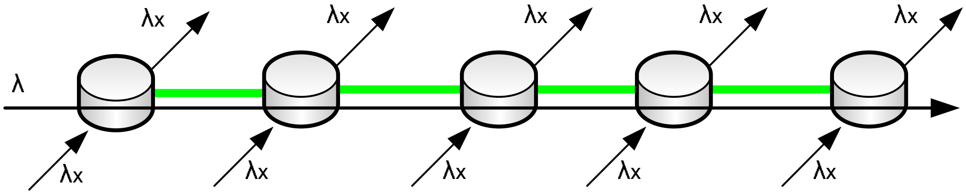


Figure 6.7: A path with an end-to-end flow and cross traffic

Let us further assume that the service time (μ_s) for a packet is either deterministic or Gaussian. By service time we mean the transmission time and the processing time in the router that includes the routing table search. If the processing time is low and the packet sizes are constant then the service time can be deemed as deterministic. On the other hand, if the routing table is large enough, one may approximate the search time by a Gaussian random variable, by virtue of the central limit theorem applied to a random number of lookups in a table.

The packet arrival process is assumed to be Poisson. Besides, a real trace has been included in the simulator to validate the end-to-end results with real traffic. The trace has been taken from an OC-48 link in the US and it can be found in [85]. The interarrival time and packet length histogram can be found in Figure 6.8.

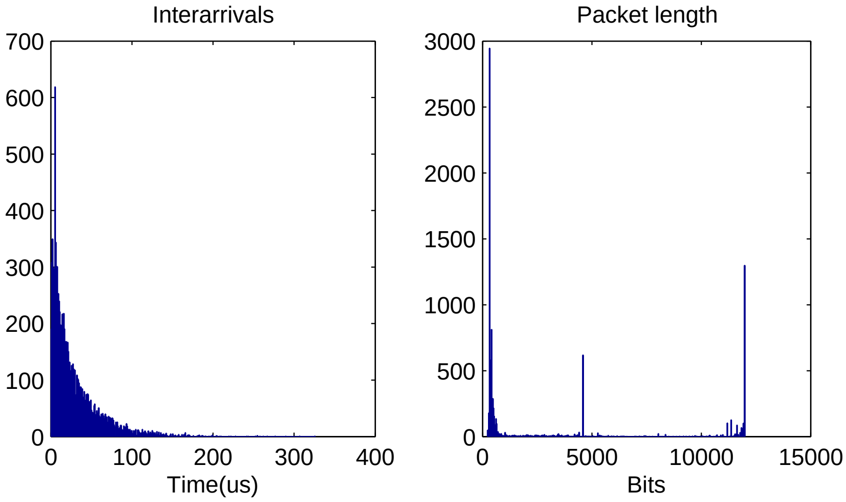


Figure 6.8: Histograms of the CAIDA trace interarrival times and packet lengths.

The incoming traffic interarrivals in the real traffic fix the demands rate and the mean packet length. To simulate the nodes in Figure 6.7, we derive C_{link} as:

$$\rho = \frac{\lambda}{\mu_s} = \frac{\lambda + \lambda_x}{\frac{8\mathbb{E}(B)}{C_{link}}} \Rightarrow C_{link} = \frac{(\lambda + \lambda_x)\rho}{8\mathbb{E}(B)} \tag{6.1}$$

where λ is the trace interarrival mean time, λ_x is the interarrival mean time for the cross-traffic, $\mathbb{E}(B)$ is the trace mean packet length and ρ is the desired queue occupation. Figure 6.9 illustrates the simulated nodes structure.

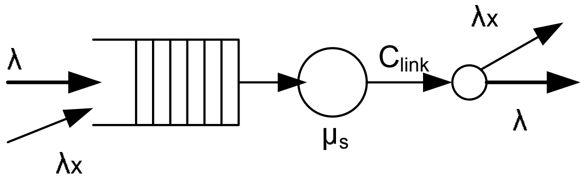


Figure 6.9: Structure of the simulated nodes

In the next section, we consider the following scenarios: *M/D/1* (Poisson arrivals and deterministic service time), *M/G/1* (Poisson arrivals and Gaussian service time). On the other hand, the *real trace* scenarios will be denoted *G/D/1* and *G/G/1*, for the deterministic and Gaussian service times respectively. In all the scenarios considered, the cross arrival process is Poisson.

6.2.2 Mean delay estimation

Mean delay estimation is the easiest way to achieve an end-to-end delay estimation. To this end, each router in the path keeps a counter that is updated with the delay experienced by each packet traversing it. Actually, let $\hat{D}_{n-1}$ refer to the estimated mean delay by the time the $(n-1)-th$ packet leaves the router. Then, $\hat{D}_n$ can be calculated as follows:

$$\hat{D}_n = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n-1} d_i + d_n}{n} = \frac{(n-1)}{n}\left(\hat{D}_{n-1} + \frac{d_n}{(n-1)}\right) = \frac{\hat{D}_{n-1} \times (n-1) + d_n}{n} \tag{6.2}$$

Where d_i stands for the delay experienced by packet number i. The mean end-to-end delay estimation can be computed as the sum of the individual mean delays at each of the routers in the path, namely

$$\hat{D}_{e2e} = \sum_{i=1}^{K} \hat{D}_i \tag{6.3}$$

Note that the above expression holds regardless of any possible correlations in the estimated mean delays, since the expectation is a linear operator.

Algorithm validation Table 3 shows the mean delay estimated $(\hat{D}_{e2e})$ using eqs. 6.2 and 6.3 at each node and the real mean delay $(\overline{D}_{e2e})$ as seen by the simulated end-to-end flow only, for different scenarios. The relative error between the real mean and the estimated one is also provided. The aim is to analyze to which extent the mean per-flow end-to-end delay, i.e. computed from the packets belonging to a specific end-to-end flow, differs from the overall end-to-end mean delay estimation, considering all the packets that go across the routers.

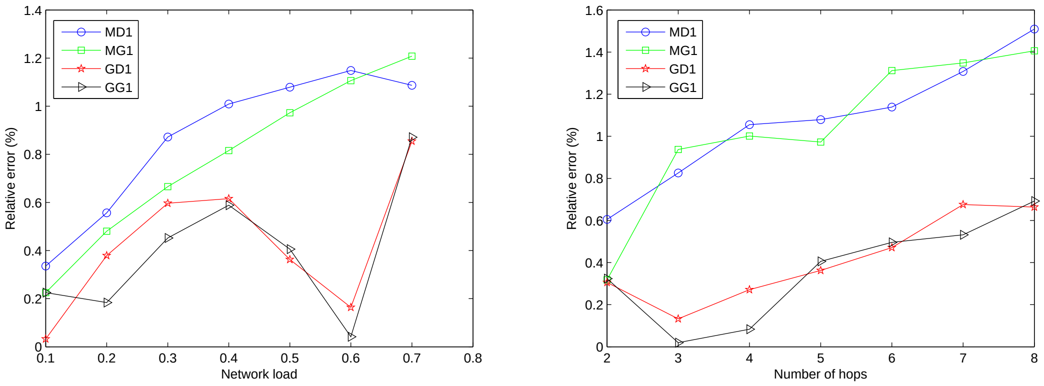
As we can see in Table 6.1, the mean estimation seems to be a very good estimation in all cases, but in the M/D/1 and M/G/1 scenarios the estimation is worst than in the real trace estimation. However, the error is around 1% in the worse case. Figure 6.10(a) shows the relative error variation when the network load changes. According to it, we can see that the error increases when the network load increases too. Nevertheless, in the real trace case this pattern is not observed. A drop in the relative error is noticed when the

Scenario	Network load	$\hat{D}_{e2e}$ (t.u.)	$\overline{D}_{e2e}$ (t.u.)	Relative Error
$M/D/1$	10%	1.531986	1.526863	0.34%
	70%	21.964775	21.72866	1.09%
$M/G/1$	10%	1.545197	1.541724	0.23%
	70%	21.960957	21.698799	1.21%
$G/D/1$	10%	1.532441	1.531941	0.03%
	70%	22.014622	22.20438	0.85%
$G/G/1$	10%	1.545197	1.548682	0.22%
	70%	22.049511	22.24325	0.87%

Table 6.1: Mean delay estimation

network load is 50-60%. An important conclusion is that the relative error is larger in the simulations with poissonian traffic than in the ones with real traffic.

The slight difference between the end-to-end delay can be due to the implicit sampling in the stochastic process that describes the end-to-end delay at packet departures at every hop. As the load increases, the sampling rate decreases for the same flow, and the relative error increases consequently.



(a) Network load variation ($K = 5$)

(b) Number of hops ($\rho = 50\%$)

Figure 6.10: Relative Error of the mean estimation

The number of hops in this scenario was five, but it is important to analyze the performance of the mean delay estimation when this number varies. When the number of nodes increases the relative error also increases (Figure 6.10(b)), and this constitutes a limitation for the mean delay estimation. As in the previous experiment, the performance for Poissonian traffic is worse than for the real trace.

The mean delay estimation approach is a poor QoS parameter because the network operator is not aware of the delay distribution. Therefore, it becomes a gross performance parameter. Alternatively, the following sections present delay estimators which provide further detail about the delay distribution.

6.2.3 Percentile estimation

The mean delay provides no information about delays away from the mean which may be suffered by the end-to-end flow. Precisely, the delay percentile provides the probability that the end-to-end delay falls below a certain threshhold. Each router in the path is in charge of estimating a certain delay percentile, which is sent back to the network edges in order to estimate the desired end-to-end delay percentile. Let us define D_i as the experienced delay in the ith node, let D_{e2e} be the e2e packet delay, and let us consider a path with K intermediate routers. Then, we can reason out as following:

$$D_1 \leq \frac{d_{QoS}}{K} \cap D_2 \leq \frac{d_{QoS}}{K} \cap \ldots \cap D_K \leq \frac{d_{QoS}}{K} \Rightarrow D_{e2e} \leq d_{QoS} \tag{6.4}$$

On the other hand, let $P(D_i > d_{QoS})$ refer to the probability of exceeding d_{QoS} in the $i-th$ node. Consequently, $P(D_{e2e} > d_{QoS})$, the probability of exceeding d_{QoS} end-to-end can be lower bounded as follows:

$$P(D_1 \leq \frac{d_{QoS}}{K}) \times P(D_2 \leq \frac{d_{QoS}}{K} \times \ldots \times P(D_K \leq \frac{d_{QoS}}{K}) \leq P(D_{e2e} \leq d_{QoS}) \tag{6.5}$$

$$P(D_{e2e} \leq d_{QoS}) \geq \prod_{i=1}^{K} P(D_i \leq \frac{d_{QoS}}{K}) \tag{6.6}$$

To compute the percentile probability on a hop-by-hop basis it is only required to calculate the sum:

$$\frac{\sum_{i=1}^{n} I_j(D_i > \frac{d_{QoS}}{K})}{n} \tag{6.7}$$

where $I_j(x)$ is the indicator function that is equal to one whenever the event x is true and 0 otherwise. On the other hand, n indicates the number of packets with which the percentile probability calculation is performed. Therefore, the router is required to keep track of two counters only, one for the indicator function and another one for the number of packets that went across the router.

Algorithm validation In this section we analyze the performance of the proposed percentile estimation. Actually, note that $\prod_{i=1}^{K} P(D_i \leq d_{QoS})$ is only a lower bound, and, thus, we wish to analyze how close this bound is to the real value. Table 6.2 shows the percentile estimation for an M/D/1 scenario with a network load of 10%. The results are really good because the delay at each hop is close to a constant. In this scenario the percentile estimation error is almost neglilible.

Delay threshold (t.u.)	1.00	10.00	100.00
Node 0	0	0.99997	0.99997
Node 1	0	0.99995	0.99995
Node 2	0	0.99999	0.99999
Node 3	0	0.99994	0.99994
Node 4	0	0.99998	0.99998
Estimated	0	0.9999	0.9999
End-to-end	0	0.99999	0.99999

Table 6.2: M/D/1 ($\rho = 10\%$) Percentile estimation

However, when the delay threshold is near the pdf median, the percentile estimation is not as accurate. Table 6.3 shows the results for a G/G/1 scenario. For the 10 units of time threshold, the lower bound is not close to the real end-to-end $P(D < d_{QoS})$. On the other hand, we note that the larger the number of hops the less accurate the lower bound becomes.

Delay threhold (t.u.)	1.00	10.00	100.00
Node 0	0	0.60903	0.99998
Node 1	0	0.60856	0.99998
Node 2	0	0.60396	0.99999
Node 3	0	0.60524	0.99999
Node 4	0	0.60563	0.99999
Estimated	0	0.08205	0.99993
End-to-end	0	0.43713	0.99998

Table 6.3: G/G/1 ($\rho = 50\%$) Percentile estimation

Similar results have been achieved with the other networks scenarios, at all network loads. In those scenarios with not very variable end-to-end delay, the percentile estimation works perfectly but in that case, the mean delay also provides an accurate estimation and it is easier to compute. For the other scenarios, the percentile estimation is a very rough estimation, so we need more sophisticated algorithms to obtain a reliable end-to-end estimation delay.

6.2.4 Fourier transform based algorithm

The end-to-end delay is a random variable that can be denoted by D_{e2e}. Let K denote the number of hops from source to destination. The per-packet delay in the i-th node is a random variable D_i, with probability density function (pdf) $f_{Di}(t)$, $t \geq 0$, and $i = 1, \ldots, K$. The total packet delay is equal to the sum of random variables . Thus, the pdf of D_{e2e} results from the convolution of individual pdfs:

$$f_{De2e} = f_{Di} \times \ldots \times f_{DK}$$

To compute the above convolution it is easier to resort to the frequency domain by means of the Fast Fourier Transform (FFT). It is well known that the FFT of a convolution is the product of the FFTs of the functions involved in the convolution. Even though this is more elaborate than the estimation of the mean or percentile, the characterization accuracy is greatly improved. The algorithm proceeds with the following steps as explained in [1]:

1. *Estimation:* the individual delay pdf $f_{Di}(t)$, $i = 1, \ldots, K$, is estimated using measurements at each node.

2. *Sampling:* each pdf is sampled to obtain a good approximation. If the pdf is bandwidth limited, the Nyquist criterion can be applied to obtain the sampling rate.

3. *Discrete Fourier Transform (DFT) calculation:* the DFT of the sampled pdf is calculated, using N coefficients.

4. *Low-Pass filtering:* as not all coefficients have the same importance, the less significant ones are cut off.

5. *Sending information to the other nodes:* the previously computed coefficients are transmitted to the other network nodes.

6. *Convolution:* The DFT of the convolution is obtained using the DFT coefficients and it is inverted to obtain the end-to-end delay.

These steps are schematically shown in Figure 6.11. Such procedure is to be followed by all the nodes in the path from source to destination.

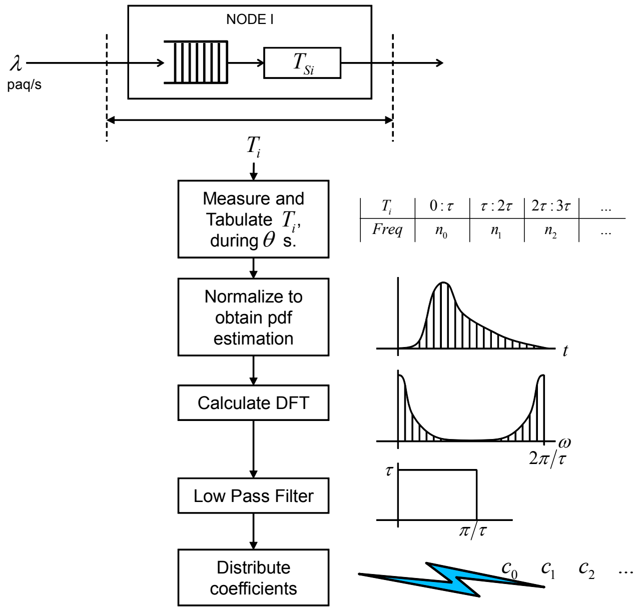

NODE I
λ
paq/s
T_{Si}
T_i
Measure and Tabulate T_i, during θ s.
Normalize to obtain pdf estimation
Calculate DFT
Low Pass Filter
Distribute coefficients
T_i | $0:\tau$ | $\tau:2\tau$ | $2\tau:3\tau$ | ...
Freq | n_0 | n_1 | n_2 | ...
t
ω
$2\pi/\tau$
τ
π/τ
c_0 c_1 c_2 ...

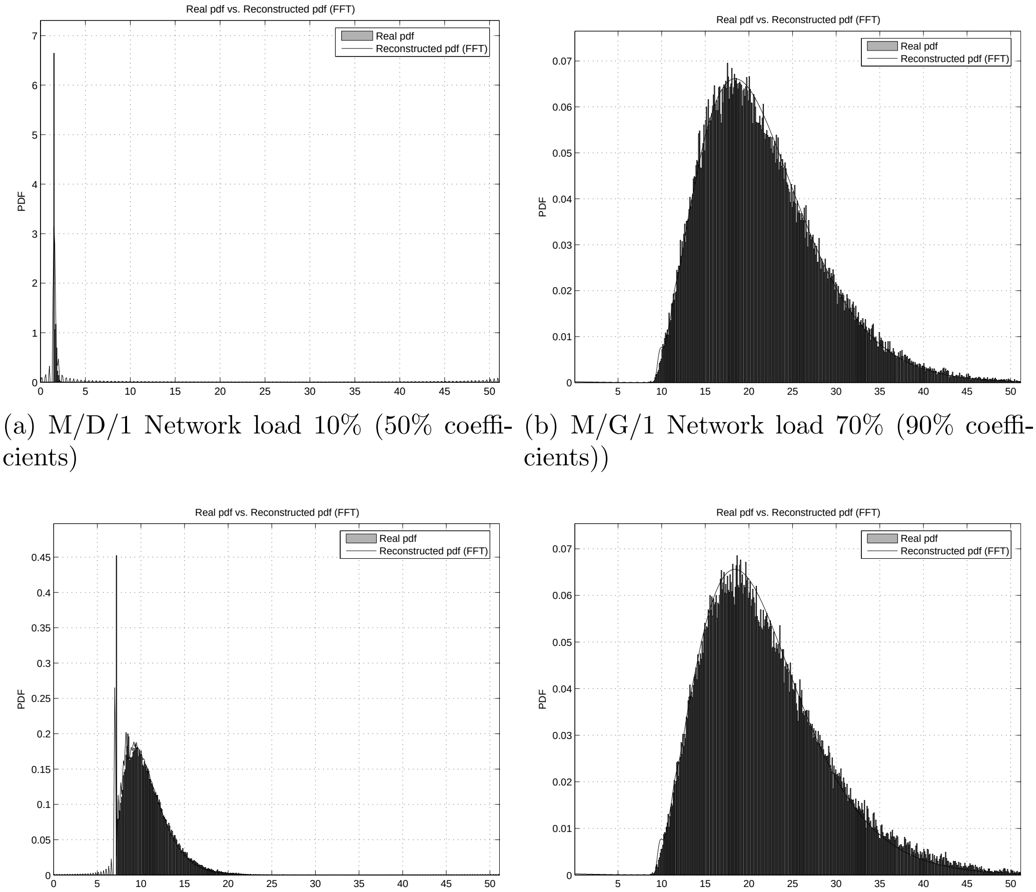
Figure 6.11: LOT algorithm steps to compute coefficients (by courtesy of [1])

The end-to-end delay distribution can be obtained by inverting the end-to-end FFT, which is equal to the product of the corresponding FFTs at the individual routers.

Algorithm validation The main objective of this algorithm is to achieve an approximation of the delay pdf at each node. Figure 17 shows the estimated pdf versus the real pdf. Both pdfs slightly differ due to the limited number of coefficients that was used to reconstruct the pdf from the FFT. However, the pdf is well estimated using this method and relatively good results can be achieved using only the most significant coefficients.

Determining how many FFT coefficients are necessary to properly estimate the end-to-end delay is crucial for the algorithm. To this end, a chi-square test was used with a significance level of 1%. Figure 6.13(a) shows the average number of FFT coefficients that make the reconstructed pdf pass the chi-square goodness-of-fit test to the real pdf. The magnitude in the x-axis is the path load. At high loads, when more accuracy is required, the number of FFT coefficients decrease. Thus, the FFT algorithm shows very good scalability and robustness features.

If the service time is non-deterministic, the delay pdf becomes smoother and the high frequency component is less significant. Therefore, we note an improvement in the number of coefficients, as shown in Figure 6.13(b). While 256 coefficients were required for the



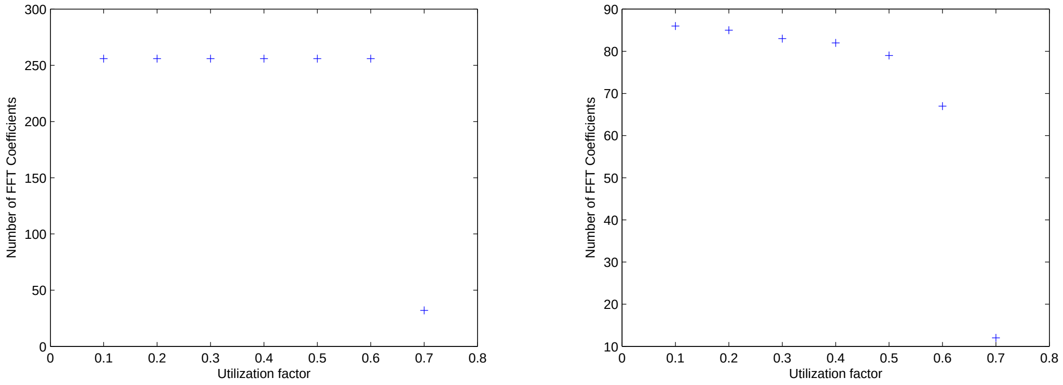
(a) M/D/1 Network load 10% (50% coefficients)

(b) M/G/1 Network load 70% (90% coefficients))

(c) G/D/1 Network load 50% (70% coefficients)

(d) G/G/1 Network load 70% (90% coefficients)

Figure 6.12: Histogram and pdf estimated by Fourier method



(a) M/D/1

(b) M/G/1

Figure 6.13: FFT coefficients required to pass the chi-square test (significance level 1%)

M/D/1 system in most cases, we can now operate with only 86 coefficients in the worst case. Thus, the end-to-end delay can be computed using this method with not so much signaling load impact.

6.2.5 Weibull mixture algorithm

The Weibull mixture model was introduced in [60] as an algorithm to calculate the end-to-end delay using real data donated by RIPE NCC. This algorithm assumes that the end-to-end delay can be described as a finite sum of Weibull distributions (mixture). The Weibull pdf is given by:

$$p(x|r,s) = \frac{sx^{s-1}}{r^s} e^{-x/r}, \quad x \geq 0, \quad r,s > 0 \tag{6.8}$$

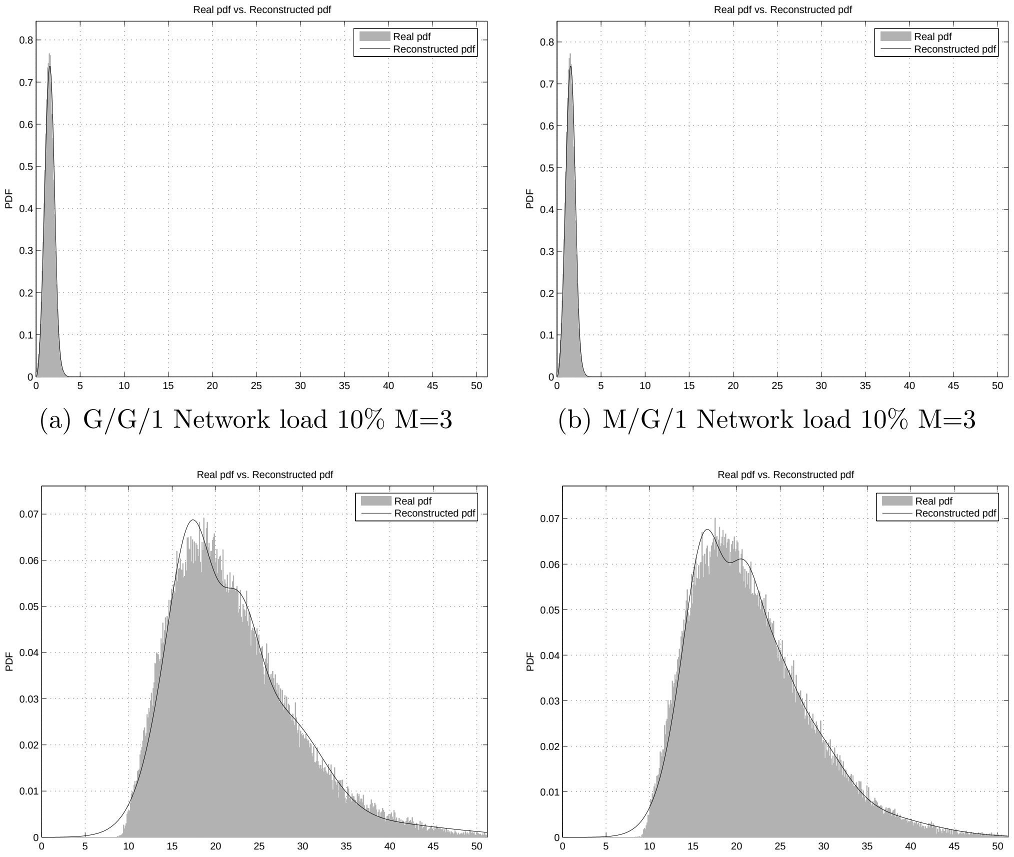
The parameters r and s constitute the scale and shape parameters for the distribution. The parameter r is related to the distribution peak and s is related to the tail behavior. Since one Weibull distribution is not enough for accurately estimating the end-to-end delay pdf, a mixture is employed. The mixture makes use of M Weibull pdfs to compute the e2e delay, each of the M Weibull pdfs are scaled using weights (q_j).

$$p(x|q,r,s) = \sum_{j=1}^{M} q_j p(x|r,s) \tag{6.9}$$

where $\sum_{j=1}^{M} q_j = 1$.

The e2e delay is characterized by the set of parameters (q,r,s), making a grand total of $3*M$ characterizing parameters. To obtain such parameters, the expectation maximization algorithm is employed, as explained in [86].

Algorithm validation This model is well validated in [60] since the dataset used consisted of 70000 one-way delay measurements, from 35 monitored points around the world, donated by the RIPE NCC institution. The clock accuracy is of a few hundreds nanoseconds so the dataset provided very accurate measurements. The main problem of this algorithm it that it finds severe difficulties for estimating nearly constant delay pdfs. Thus, it can not be used neither for the M/D/1 nor for the G/D/1 scenario. Nevertheless, from [60], we observe that the delay profile for a real network is closer to the Gaussian scenarios. Therefore, we will show the results for the M/G/1 and G/G/1 scenario only. Figure 6.14 shows the estimated pdf using the Weibull mixture algorithm. We observe that the estimated pdf fits the real one remarkably.



(a) G/G/1 Network load 10% M=3

(b) M/G/1 Network load 10% M=3

(c) G/G/1 Network load 70% M=4

(d) M/G/1 Network load 70% M=5

Figure 6.14: Histogram and pdf estimated by the mixtures of Weibull distributions method

As a conclusion, we note that the Weibull mixtures algorithm provides very good accuracy, with very little impact on the signaling load.

6.2.6 Performance comparison of the algorithms

We consider delay as the primary metric for QoS in this work. Previous sections validate the performance of some algorithms to monitor the traffic in a single domain. This algorithms allow us to monitor the information of an end-to-end path, thus allowing the delay characterization of a given path. Let us summarize the main advantages and drawbacks of each method:

- **Moments:** *Mean:* This is by far the simplest estimate that we may consider. However, it only provides information about the network stationarity. If the utilization factor is smaller than 100% the observed delay will be bounded, and the mean will remain within a range of a given stationary value. *Variance:* The variance provides information about the delay variability and it serves to fully describe the end-to-end delay for two-moment distributions, e.g. a Gaussian distribution. If the number of hops is large enough then the end-to-end delay converges in distribution to the Gaussian law. A mean and variance characterization may suffice for a variety of applications. As section 6.3 shows, these parameters are enough for the definition of the Bayesian decisor in a multi-domain scenario.

- **Delay percentiles:** The delay percentile $P(D > x)$ being D the end-to-end delay and x a given value provides a tight upper bound for the end-to-end delay which can be used for delay-sensitive applications. However, note that $D = D_1 + \ldots + D_n$, with D_i being the delay at router i, $i = 1, \ldots, N$. Hence, the end-to-end delay percentile cannot be obtained from the individual end-to-end percentiles. Nevertheless, a lower bound can be obtained by considering that $P(D > x) > P(D_i > x)$, $i = 1, \ldots, N$.

- **FFT coefficients of the probability density function:** The Fourier transform of the pdf may be approximated by the FFT of a discretized pdf (normalized histogram). Then, the end-to-end delay pdf may be estimated by performing the product of the individual FFT coefficients and then inverting the FFT.

- **Weibull mixture algorithm:** The Weibull mixture algorithm provides an approximation of the discretized pdf (like FFT algorithm) just with three parameters (q,r,s) for each Weibull used in the model.

Table 6.4 summarizes the pros and cons of each delay estimator.

6.3 Risk function definition for multi-domain networks

6.3.1 Problem statement

In multilayer networks, a network operator provides Label Switched Paths (LSPs) that could be routed either through the IP (electronic domain) or the optical domain depending on the traffic QoS requirements and on resource utilization factors. In this scenario, a multilayer algorithm chooses the proper network layer for a connection, depending on the maximum delay required by the client and the expected delay for this connection at the optical and electronic domains. Complementary to the multilayer analysis in the

Delay estimator	Size (bytes)	Complexity	Accuracy
Moments	Very small (32 bytes)	Very low (several counters to be updated upon packet arrivals)	It is only accurate if the traffic is sufficiently regular (small variation)
Delay Percentiles	Small (several 32-bytes, words depending on the number of percentiles to be submitted)	Medium (the router is required to calculate a delay histogram)	Medium (it is only a lower bound to the actual delay)
FFT coefficients	Fair (as large as 512 32-bytes words, it may be too large to be supported by current signaling systems)	High (the router is required to calculate a delay histogram and, then, perform the FFT calculation)	High (it provides the end-to-end delay histogram)
Weibull mixture algorithm	Small (three 32-bytes, words for each Weibull used in the model	High (the router computes the histogram and then the EM algorithm	High (it provides the end-to-end delay histogram)

Table 6.4: Comparison of each algorithm's size, complexity and accuracy

previous chapters, we now focus on how to operate in multi-domain scenarios. The main difference between a single domain or a multi-domain scenario is that there is no full infomation about the topology of the domain, but just partial visibility privded by border routers [71]. Consequently, the information about the delay in each domain is not detailed, but it is widely accepted that aggregated information is exchanged between the network providers [87]. Previous work on IP networks like [88, 89, 90] validate that it is possible to aggregate the network topology and exchange QoS information among nodes. In the section 6.2.4, some algorithms are validated to monitor the traffic in a single domain. These algorithms allow us to get the information into a single domain (Figure 6.15). An entity in the control plane takes the information of each node and the end-to-end delay can be computed using any of the previously validated algorithms. There is some research concerning the scalability of the exchange information mechanisms like in [91, 92], but these issues are out of the scope of this work.

Such QoS information can be exchanged between the domains, so it is possible to decide whether to route an LSP over the optical or the electronic domain. Figure 6.16 depicts the same scenario than in Figure 6.1, but with the vision that a node in AS 3 has of the network. As our model is based on delay, the border routers announce the delay to reach

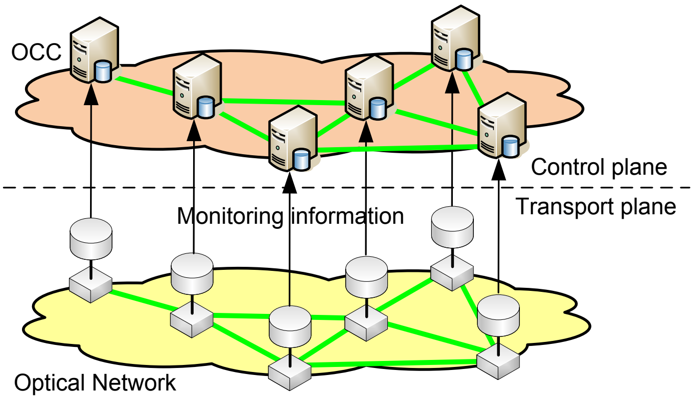


Figure 6.15: Monitoring information in a single-domain

the other border routers in its domain. Consequently, the border routers may know the end-to-end delay of the paths that connects every router in the network.

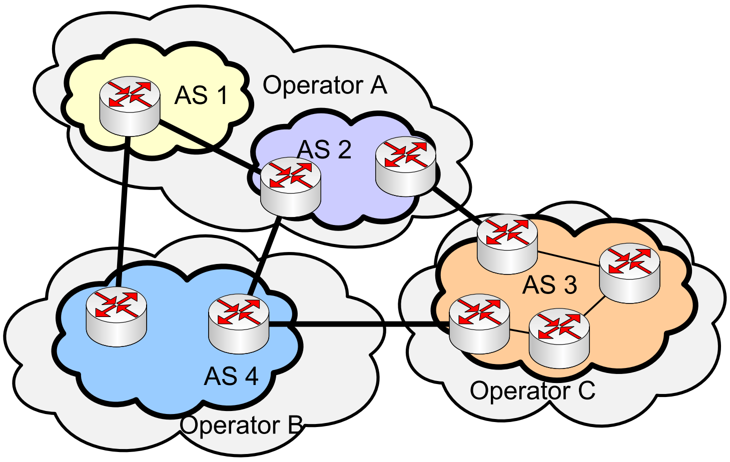


Figure 6.16: View of a node in the AS 3 of the overall network

In the chapter 4, we define the Bayesian risk in terms of utility and cost. The computation of the utility and the cost values yields to a risk function which is minimized to find the optimal decision. The cost function was defined in section 4.1.2 with the following design premises:

1. IP equipment is already deployed by the operators networks, so it is the first option to send the incoming traffic.

2. When IP layer does not provide the desire utility to the flows, a new e2e lightpath is established.

3. The longer the light-path is, the more congestion is reduced in the IP layer.

Such design premises are kept in a multi-domain scenario, where the operators in each domain want to use the currently deployed IP equipment, while providing the required QoS to their customers. Consequently, the cost function must not be changed for the multi-domain scenario.

On the other hand, the utility is based on the delay in each hop. In this multi-domain scenario, where there is not full exchange information, the neighbor domains do not know how many hops there is between the border routers in the network. Therefore, the delay model for the utility function must be changed. Previous algorithms (section 6.2.4) show that it is possible to monitor the mean and the variance of the delay in a given path (see section 6.2.6). Using this information it is possible to use the expression of the utility functions defined in section 4.1.3. Let us show the expression of the utility functions:

$$\begin{aligned}
\mathbb{E}_x[U_{\text{mean}}(x_k^{e2e})] &= \mathbb{E}_x[-x_k^{e2e}] = -\sum_{j=k}^{M}\mathbb{E}_x[x_j] && (6.10)\\
\mathbb{E}_x[U_{\text{step}}(x_k^{e2e})] &= \mathbb{E}_x[U_{\text{step}}(\sum_{j=k}^{M} x_j)] = P(x_k^{e2e} < T_{\max}) && (6.11)\\
P(x_k^{e2e} < T_{\max}) &\sim N(\mu_k^{e2e}, \sigma_k^{e2e}) &&\\
\mathbb{E}_x[U_{\exp}(x_k^{e2e})] &= \lambda e^{\frac{\sigma_k^2{}^{e2e}\lambda^2 - 2\mu_k^{e2e}\lambda}{2}} && (6.12)
\end{aligned}$$

In light of the above equations, it is clear that if the mean delay is computed by monitoring entities in each single domain and this information is exchanged between the domains it is possible to know all the required information to extend the Bayesian decisor to a multi-domain scenario.

6.3.2 Numerical results

Figure 6.17 shows the reference model for the multi-domain scenario. Such scenario has three domains, which have two demands N_1 and N_2 LSPs. The first N_1 LSPs traverse the domains A, B and C, while the N_2 LSPs just traverse the domains B and C. The intra-domain paths, which connects every border router, has five hops and their structure is depicted in Figure 6.7. Let us assume that the amount of cross-traffic in each domain is 90% and the traffic scenario for each domain is M/G/1 (see section 6.2.1). Note that altough every domain knows its own topology, just the domains end-to-end mean delay (μ_i) and variance (σ_i) is used by the Bayesian decisor.

We assume that all domains have the same R_{cost} parameter set to 2. The bandwidth of the LSPs is 5% of the queue occupation (ρ) and $N_{max} = 7$. The value of T_{max} for the U_{exp} is 60 u.t. The K_u and K_c values are set to fill the 50% of the IP layer when there is a demand of 7 LSPs.

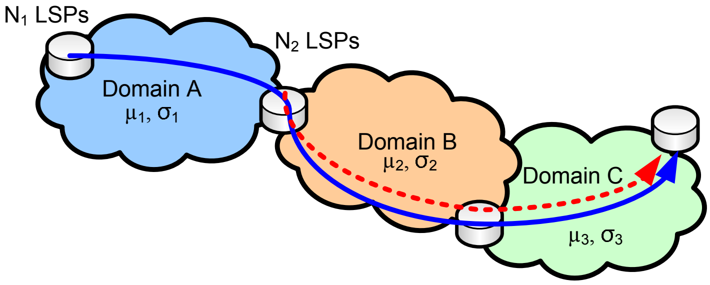


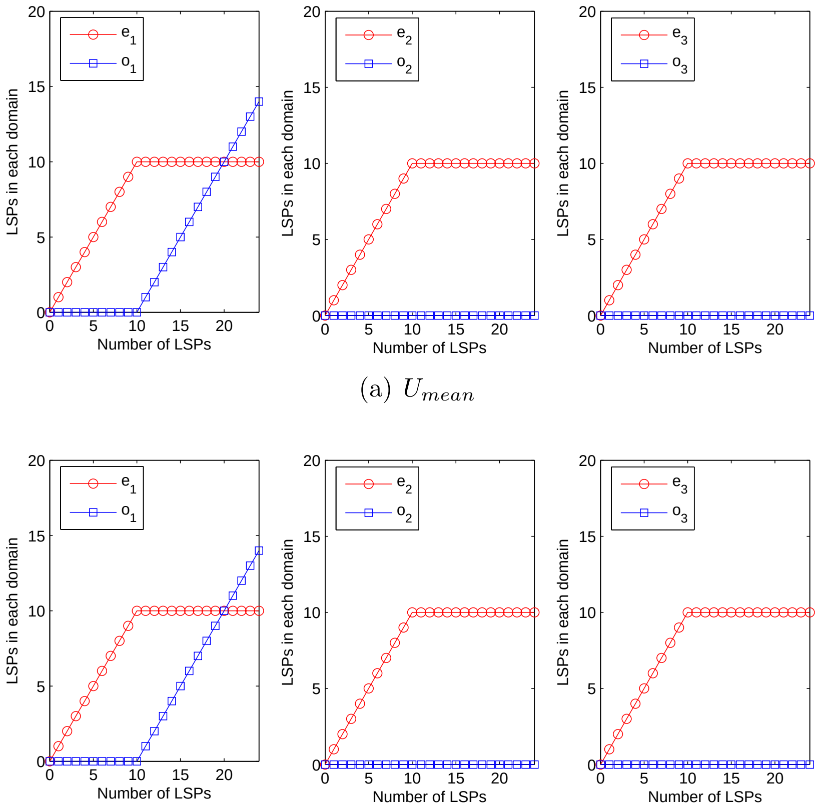
Figure 6.17: Multi-domain reference scenario

Traffic increment in the first domain

This experiment evaluates how the decisor changes the amount of traffic when the amount of incoming traffic in the first domain is increased from 0 to 24 LSPs. Figure 6.18(a) depicts the amount of LSPs that are switched in the IP and WDM layer when the U_{mean} function is used. In light of the results, we can see that while there is enough QoS at the IP layer, the decisor sends the incoming LSPs over the electronic domain, until the QoS is degraded that much that a new lightpath is set up from the first border router in the first domain to the third domain. These results validate the Bayesian decisor, since the behavior is similar to the multi-hop scenario, but the utility functions have been changed to deal with multi-domain environments. Figure 6.18(b) shows the results with U_{step} function. The results are the same for both utility functions and U_{exp} shows the same behavior.

Traffic increment in the second domain

The following experiment shows which decision is made when the second domain increase its amount of offered traffic. In this case, $N_1 = 7$ and N_2 varies from 0 to 24. The amount of LSPs at the IP layer in the first domain (e_1) decreases, since as the amount of traffic in the second domain is larger, the utility perceived at the IP layer is not enough for the services. Figure 6.19(a) shows these results. When the U_{exp} function is used the results



(a) U_{mean}

(b) U_{step}

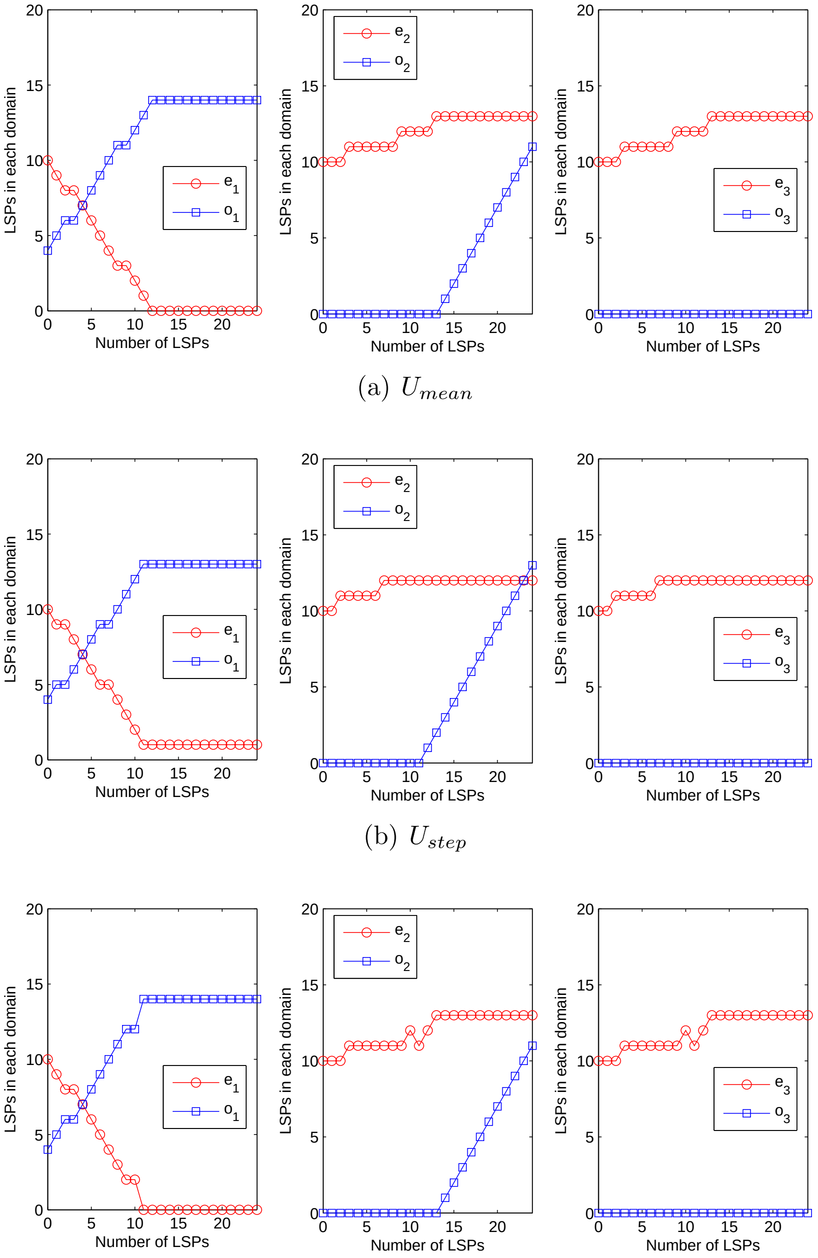
Figure 6.18: Number of LSPs routed in the electronic and optical layer when the first node increases its traffic

are similar (Figure 6.19(c)), but the decision in the second node e_2 allows one LSP more than U_{mean}. U_{step} results are similar. Again the results here show that the extension of the algorithm to other scenarios is possible and that the design requirements are kept in the multi-domain scenario.

6.4 Conclusions

This chapter extends the concepts of the Bayesian decisor to multi-domain scenarios. There are still open issues that the research community is evaluating such as the scalability of the information exchange between multiple domains, but the state of the art says that it is possible to establish optical or electronic connections in multi-domain scenarios with either GMPLS or the PCE architecture. The information provided by each domain comprises intra-domain delay, which can be used by the Bayesian decisor to decide if a new lightpath needs to be created.

In multi-domain scenarios case, the Bayesian decisor operates as in the multi-hop cases with the difference that now each node's delay function comprises the delay introduced by all nodes of that particular domain. This multi-domain scenario could be more complicated and instead of having just a single-path between the domains, it could have multiple paths. This problem would be similar to the problem covered by chapter 5.



(a) U_{mean}

(b) U_{step}

(c) U_{exp}

Figure 6.19: Number of LSPs routed in the electronic and optical layer when the second node increases its traffic

Chapter 7

A Multi-layer solution for Internet Service Providers

This chapter proposes Multi-layer FAN (MFAN) as an evolutionary solution to support Flow-Aware Networking (FAN) in Multi-layer environments. Current network operators are interested in providing their IP layer with QoS solutions, such as Diffserv or FAN. However, such approaches provide QoS at the IP layer only, and it is necessary to combine them with Multi-layer Traffic Engineering on attempts to make QoS mechanisms aware of the underlying WDM switching technology. Our proposal uses the FAN's monitoring parameters to detect which flows are more suitable to be transmitted over the optical lightpaths or over the hop-by-hop connections. Three different policies are introduced to decide on which criteria an IP flow arrival is extracted from the standard FAN architecture to be forwarded onto a transparent lightpath to its destination. The performance of the three proposed policies are discussed in terms of goodput and of queueing delay.

This chapter is organized as follows. Section 7.1 motivates the problem statement and section 7.2 provides an overview of the two main QoS proposals: IntServ and Diffserv. Section 7.3 introduces the concept of Flow-Aware Networking. Section 7.4 describes the MFAN architecture and the three policies to manage the switching of IP flows of an IP flows from the electronic layer to the optical layer. Section 7.5 outlines the benefits and drawbacks of MFAN from simulation. Finally, Section 7.6 concludes this work and proposes a few new perspectives.

7.1 Motivation

The provisioning of Quality of Service (QoS) to applications implemented at the end-nodes is one of the key issues in the engineering of the Next-Generation Internet. Recent studies indicates that multimedia applications are becoming more and more popular, not only Internet Protocol Television (IPTV) [93], but also VoIP [94]. Given their delay and bandwidth restrictions a best-effort packet switched network is not suitable to such applications. Besides network resource overprovisioning, two main approaches for QoS support on IP networks have been proposed in the literature: IntServ and DiffServ. The former is well-known for its lack of scalability due to the soft state of the virtual circuits on which this technique relies on [95]. The latter requires a signaling channel and a control plane based on complex algorithms to address packet marking and metering. Such mechanisms lead to an expensive approach for QoS provisioning. In this context, a new approach, called Flow-Aware Networking (FAN) [96, 97, 98, 99], has appeared as a promising technology to manage congestion control in IP networks and to provide QoS to applications.

Essentially, FAN operates at the packet level and implicitly distinguishes between two different types of flows: streaming (or priority) flows, and elastic (or non-priority) flows. Streaming flows typically refer to voice or video applications (UDP), whereas elastic flows often regard to TCP sessions. The FAN architecture is designed to meet two main objectives: (1) minimize loss and queueing delay suffered by the streaming flows in the routers; and (2) provide a minimum fair rate to the elastic flows. When FAN cannot satisfy these minimum requirements, it rejects new incoming flows. By using such an admission control policy, FAN keeps into service the already admitted flows, thus assuring a minimal QoS under overloaded conditions.

In a first approach, FAN was conceived to operate at the IP level without any information about the underlying layers. With the current technology trends, network operators are gradually migrating to an IP over WDM paradigm, mainly to benefit from the larger transmission capacity offered by optical fibers. Previous work on traffic engineering in multi-layer networks [100, 20, 101, 102] concludes that both the IP and the WDM layers should interact and exchange their information to provide a better QoS. This is the reason why we propose an adaptation of the FAN concept to IP/WDM architectures, and by extension to multi-layer capable routers including optical and electronic switching [20].

Some studies have extended Diffserv concepts to multi-layer scenarios [103, 48, 104, 105]. The authors in [103] propose an architecture for a multi-layer capable router, which includes Diffserv in its IP cards. For instance, the authors in [48] propose a scheme with three CoS that are based on Diffserv. They use SLA to assure a QoS value to each CoS. Essentially, each SLA specifies four QoS parameters: delay, jitter, packet loss and bandwidth. Depending on the flow's CoS the algorithm use different connection models: Optical-LSP, "Virtual circuit connection" and datagram forwarding. The first connection model is a full optical connection. If the QoS requirements are fulfilled, a "Virtual circuit connection" is sent using already created routes, otherwise a new connection is established. Finally, "best-effort" traffic is accommodated on the already available resources. The authors in [105] have provided a new architecture to support Diffserv and MTE in Internet Protocol (IP) over WDM networks.

We define the concept of Multi-layer FAN (MFAN) as a router architecture which is capable of handling optical resources provided by WDM technology within the FAN approach. When a new data flow arrives at a MFAN node, this node first tries to serve this flow's packets at the IP layer. If the IP layer is congested, the flow is transferred to the optical layer whereas this layer benefits from available resources. Figure 7.1 illustrates the architecture of the considered IP/WDM network. The IP packets are routed and forwarded via hop-by-hop lightpaths from source to destination or they can be transmitted through the optical domain use end-to-end lightpaths.

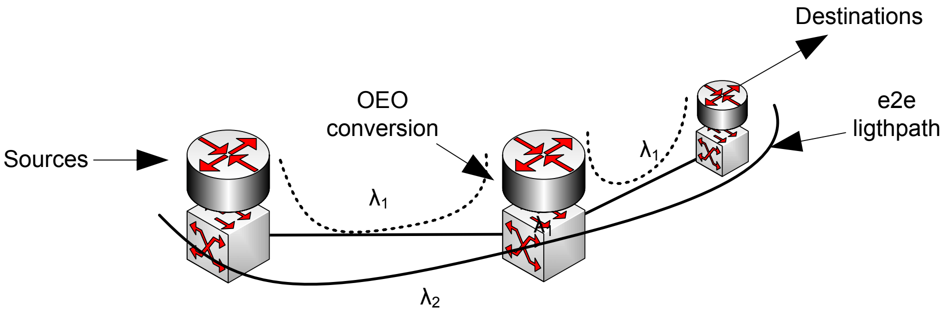


Figure 7.1: End-to-end transparent lightpaths and hop-by-hop opaque lightpaths.

Several policies can be defined to decide which flows are more suitable to be transmitted though the optical layer or via the IP layer. The authors in [48] use Diffserv classification to map the CoS over the kind of connection. If no packet marking is done, it is required to detect features of the incoming flows in order to make such decision. Flow classification is a extensively studied topic [106]. The authors in [104, 107] find flow features to decide which flows should be sent to the optical layer. A first approach is the detection of "elephants" and their transmission through the optical layer [107]. A second approach is to use Diffserv's classification and the flow bandwidth, to decide whether it is transmitted via an Optical

Circuit Switching (OCS) connection or using Optical Burst Switching (OBS). Our approach is to use the implicit classification of FAN and its monitoring parameters to make such decision. This approach does not increase the system complexity and it does not require packet marking.

7.2 Quality of Service at the IP layer

Current packet switched network treat all packets the same way without any preference. They do not provide a better performance to the packets from real-time applications or streaming ones. This behaviour is called “Best Effort”, since the network gives the best possible service to all packets. New applications have appeared in the past years (P2P, grid, VoIP, etc.), which need some service guaranties. In this scenario and with the aim to reduce Operational Expenditure (OPEX), operators are merging all their legacy networks (telephony, ATM,...) to a single IP network, which should be able to provide all different services offered by previous ones, but suc an all-IP-based network needs to support QoS provisioning.

Quality of Service (QoS) is a network skill to offer different network performance to the different applications over the network. QoS is defined in terms of four parameters: bandwidth, delay, jitter and packet loss probability.

QoS is directly related to the network architecture and network congestion. QoS allows to offer new services to the network, because it can support high bandwidth dedicated services, reduce packet loss probability, avoid and manage network congestion and organize traffic.

Two main approaches has been proposed to provide QoS:

- **Integrated Services (Intserv):** this architecture provides an individual treatment to each network flow. This approach clearly does not scale to the global Internet, because the number of flows in the Internet is huge and treating them individually is an unmanageable problem.

- **Differentiated Services (Diffserv):** this architecture treats the traffic based on data aggregation, providing the data in the same group with the same QoS. This architecture solves the scalability problem of IntServ. DiffServ does not assure absolute QoS guaranties, but relative treatment preference between classes. DiffServ requires complicated traffic engineering to be efficient. Pricing is part of the contract and depends on the traffic volume emitted per class and the class itself. The absence of a direct link between cost and price complicates the pricing, so necessary for the ROI.

Although IntServ was proposed in 1994 by the IETF [108] and DiffServ was defined in 1998 ([109],[110]) none of them has yet been deployed by operators, because of the main problems previously cited (scalability and cost). Therefore, new proposals for QoS have appeared. A new QoS architecture was proposed in [96] and it is called Flow-Aware Networking (FAN).

In the following sections, an extended overview of Intserv and Diffserv is provided. Section 7.3 explains in detail FAN architecture.

7.2.1 Integrated Services

The Integrated Services architecture [108] provides QoS based on the flow concept, in a way that mimics QoS enforcement in the telecommunications network. A data flow is a continuous traffic stream created by one user and all its packets require the same QoS. Therefore, the minimum QoS unit in Intserv is the flow.

A key point in Intserv is the flow definition. Flows can be identified using the source and destination address, the source and destination ports and protocol. In IPv6 flow identification could be done using the Flow Label and the source and destination addresses, or just like is done in IPv4.

The IntServ architecture defines three categories of services the traditional best-effort services, guaranteed services and controlled load services [111].

- **Guaranteed Services:** this service is guaranteed a minimum bandwidth and maximum delay. Each router in the e2e path must offer such requested guarantees.

- **Controlled Load Services:** the quality in this service is like a priorized traffic. The network gives priority to this traffic, but it does not assure any delay or bandwidth constraint.

- **Best Effort Services:** this service has no QoS specifications. The network delivers the traffic as soon as possible, following the conventional best-effort approach.

IntServ defines a set of service commitments that the network must fulfil to support IntServ [108]. These commitments can be classified into three categories:

- *QoS commitments:* are the maximum and minimum delays. Guaranteed services, which have a minimum bandwidth and maximum delay guaranteed, specify its requirements using the traffic characteristics (Tspec) of the flow along with the reservation characteristics (Rspec). Controlled load services only specify its traffic characteristics (Tsec) and they are negotiated at session setup, since they do not have any QoS parameter to be assured.

- *Resource-sharing commitments:* are collective policies that affect to flows transported by the network. There are three categories: multi-entity link-sharing, multi-protocol link-sharing and multi-service link-sharing. These categories are defined to share the same link among multiple entities or organizations, protocols or services.

- *Resource reservation commitments:* are defined to improve the network utilization and avoid performance degradation of the services.

Based on all these commitments, a Reference Implementation Framework (RIF) was defined with four components: [108] the packet scheduler, the admission control routine, the classifier and the reservation set-up protocol.

The packet scheduler has a set of queues where the incoming packets from different stream are forwarded. The classifier separates packets into classes so the packet scheduler can provide the desired QoS. The admission control routine is an algorithm that determines the acceptance or refusal of an incoming new flow into the system. Its decision is based on whether or not the already accepted flows are guaranteed their QoS levels. The Resource Reservation Protocol (RSVP) [112] lets the network find an adequate route to support each flow's QoS and further adapt such routes to failures and changes in the network.

The instantiation of the RIF in a router (Figure 7.2) is defined in [108]. The routing agent deals with the computation of the routing tables. The reservation setup agent is in charge of setting aside the resources for the flows, once the admission control has accepted it. The manager agent handles the routing operation. The media transfer layer takes care of packet processing and forwarding.

7.2.2 Differentiated Services

The Differentiated services (DiffServ) approach [110] is based on the idea that traffic is already classified, according to the QoS requirements for each packet. This class information is included into the packet. This is different from of Intserv which provides QoS based on

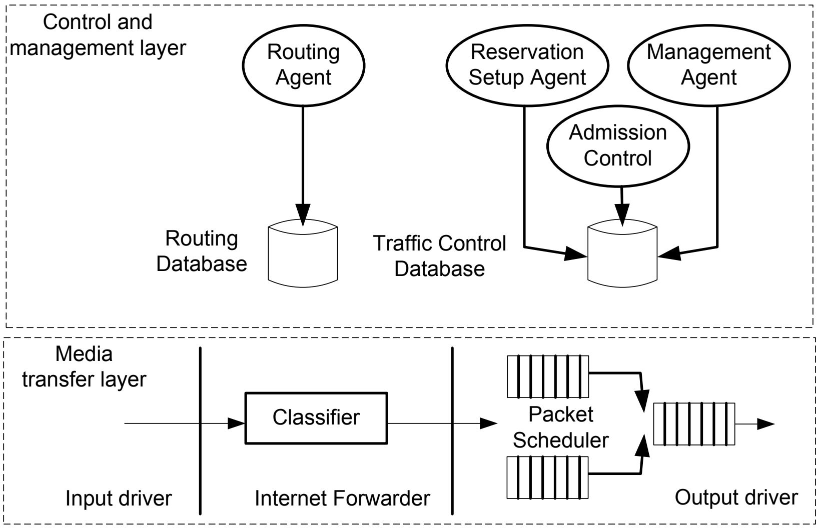


Figure 7.2: Implementation Reference Model for Routers

each flow requirements, thus allowing DiffServ scheme to be scalable. Additionally, DiffServ core routers are capable of forwarding the packets with different Per-Hop Behaviour (PHB).

Essentially, the PHB is the QoS information written on each packet using Differentiated Service (DS) field. The current specification allows thirty two traffic classes and it also includes congestion notification information. The DS field is added to the IP header instead of Type of Service field. This compatibility tries to help in the DiffServ implementation. The DiffServ architecture does not use any extra signaling information, only the routers need to be configured to support different PHBs.

Three main service classes are defined in DiffServ architecture:

- **Expedited Forwarding:** this is the service that is provided greatest quality. This service would be similar to a virtual dedicated line or a CBR ATM circuit. It is assured a minimum bitrate, a maximum packet loss rate, a mean and maximum delay and a maximum delay jitter.

- **Assured Forwarding:** this is a priority service, whereby no QoS parameters are assured, but traffic is processed with certain priority. Four classes are defined within this service class, with three types of guarantees per class depending on the discard packet probability. Therefore, there are twelve possible services using these category. Using the four classes, routers can be programmed to use a certain amount of resources for each class (bitrate, queueing size, etc.).

- **Best Effort:** no guarantees are assured for this service.

If we compare Intserv and DiffServ services we can see that the Expedited Forwarding service is equivalent to the Guaranteed service, while DiffServ's Assured Forwarding service could be quite close to Intserv's Controlled Load Service.

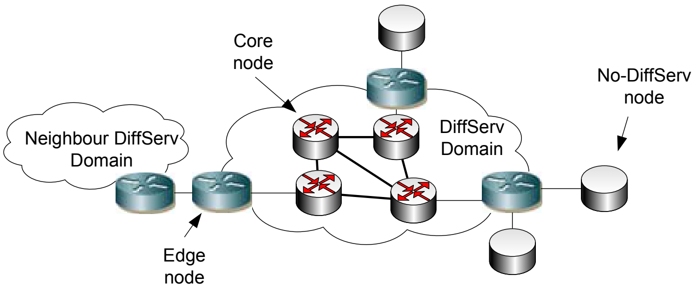


Figure 7.3: DiffServ components

The DiffServ architecture has two types of nodes: edge and core nodes [110]. Edge nodes are in the limit of the DiffServ domains and they can be connected to other DiffServ domains or to Non-DiffServ domains (see Figure 7.3). Depending on the traffic direction, edges nodes can work as Ingress or Egress nodes (Figure 7.4). A DiffServ ingress node classifies incoming traffic into a DiffServ category in order to fulfill the Service Level Specification (SLS) of packets. The SLS is a set of parameters that specifies the traffic profile and the rules to be followed by the classifier The classifier (Figure 7.4) categorises the packets using a set of filters that select the packet according to its DS field value. The traffic can be also metered, re-marked, policed and shaped (or dropped) to assure that the traffic stream is conformed to the rules defined by the SLS.

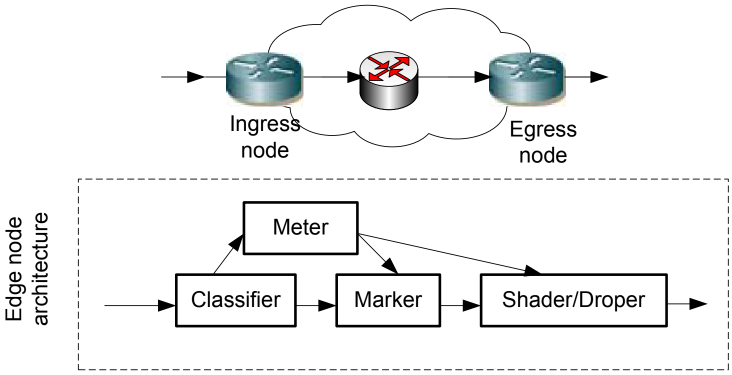


Figure 7.4: Ingress and Egress nodes (DiffServ)

The DiffServ architecture has many variations and alternatives that are focused in solving the problems of the Intserv architecture. Further details concerning DiffServ can be found in [113].

7.3 Review of Flow Aware Networking

Flow-Aware Networking (FAN) was proposed in [96] as a new approach to offer Quality of Service to the Next-Generation Internet, based on implicit classification and admission control of incoming flows, and flow-based scheduling. A flow can be considered a stream of packets with same header attributes, whose packets do no exceed a given interarrival time. FAN simplifies network operations leading to potentially significant cost reduction in the IP backbone because it increases network efficiency. Additionally, FAN requires no change of existing protocols and it can be implemented as an individual device connected to each border router interface.

Essentially, FAN performs implicit classification of flows, choosing from two categories: streaming (high-priority) and elastic (low-priority). Moreover, it defines an admission control mechanism which seeks two objectives: On the one hand, it gives preference to streaming flows on attempts to minimize packet loss and the delay they experience but, at the same time, it aims at ensuring a minimum useful data rate (also known as goodput) to elastic flows.

To this end, FAN defines two parameters: the Priority Load (PL) and the Fair Rate (FR). Fair Rate is an estimation of the bandwidth that an incoming flow would receive if admitted, while Priority Load is an estimation of the service rate of streaming packets in the queue.

Incoming flows are denied access to the system when the FAN architecture can not guarantee a given performance level (delay and fair rate). This admission control mechanism is depicted in Figure 7.5.

The complete process operates as follows: When a packet arrives at the system, the admission control module finds the flow it belongs to, namely f_n, and evaluates whether such f_n is in its inner Protected Flow List (PFL). This list stores the flow identifier of each flow already accepted by FAN and transmitted over the IP layer. If $f_n \in$ PFL, then the packet is served. Otherwise, the packet is part of a new flow which must go through the FAN admission control process. When so, it is tested whether $PL < Th_{PL}$ and $FR > Th_{FR}$, that is, if a given QoS guarantees defined by the Th_{PL} and Th_{FR} thresholds are maintained

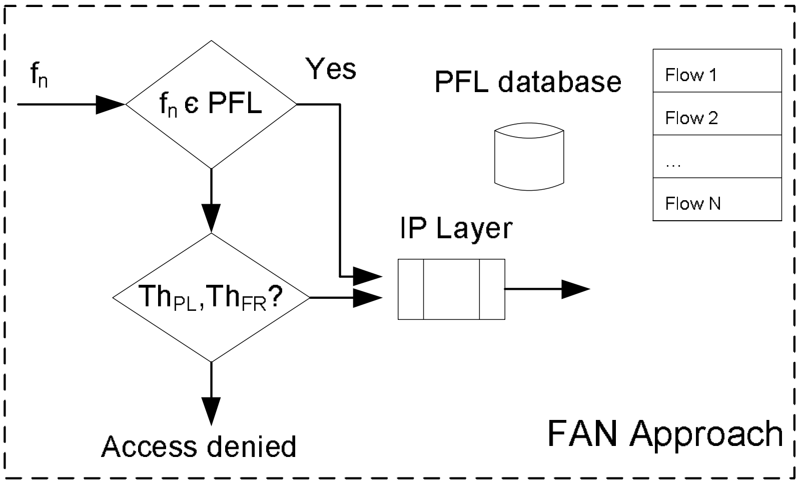


Figure 7.5: FAN Admission Control Flow Diagrams.

or not. If this is the case, the new flow is accepted; otherwise, it is rejected. Therefore, the acceptance-rejection process is defined based on PL and FR. Figure 7.6 defines the Admission Control region depending on the values of PL and FR. On the top-left part of the figure, the acceptance region is found, where $PL < Th_{PL}$ and $FR > Th_{FR}$. The value of PL is roughly estimated every tenths of milliseconds (packet time-scale) and FR is estimated every hundredths of milliseconds (flow time-scale).

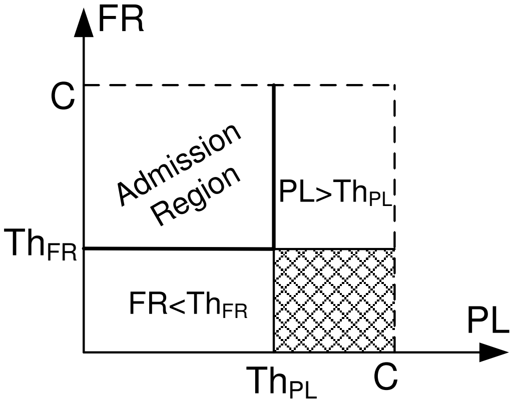


Figure 7.6: Admission Control regions.

Although flows already accepted are somehow protected, only those flows which transmit at a lower rate than Th_{FR} are treated as streaming flows (high-priority). All the others are considered elastic flows thus receiving less preference. This is so in order to avoid that flows abuse from the system resources. This classification is called “Implicit service differentiation". After such classification of flows is performed a Priority Fair Queue (PFQ) algorithm, as defined in [98] (which is based on the Start Fair Queue (SFQ) [114]).

Basically, the PFQ is a Push In First Out (PIFO) queue, which stores packet information (flow identifier, size and memory location) and a timestamp which is determined by the SFQ algorithm. The PFQ queue is split into two areas delimited by a priority pointer (see Figure 7.7), whereby streaming flows are temporally stored at the priority queue area (at the head of the queue), and the elastic flows are stored at the tail of the queue. Preference is given to the priority area since it is served before the non-priority area. Finally, the queue stores non-and high-priority packet count statistics, which are further used to compute the values of PL and FR.

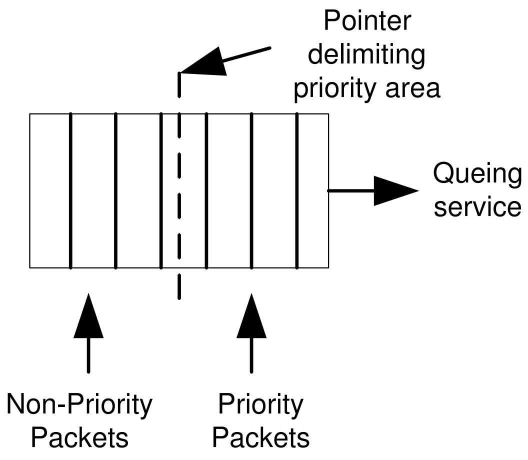


Figure 7.7: Priority Fair Queueing architecture.

In addition, an Active Flow List (AFL) is maintained by the PFQ. This list is similar to the PFL defined above, but it also stores the amount of packets transmitted per flow in the recent past. The flows with the greatest amount of transmitted packets (also known as greatest "backlog") may be discarded under severe congestion conditions. It has been shown in [98, 115, 116] that the AFLs does not suffer from scalability problems. An alternative algorithm to PFQ, called Priority Deficit Round Robin (PDRR), is proposed in [99]. The performance of PDRR is shown similar to that of PFQ [99], but its complexity is $O(1)$ ([99]), whereas PFQ's complexity is $O(log(N))$, like SFQ [114]. In this work PFQ is used, since for simulation purposes the behavior is the same.

It is worth remarking that FAN architectures have been tested [117, 118], patented [119, 120], standardized [121] and commercialized [122]. In [118], the authors compared flow-based and packet-based routers; their results showed that flow-based approach offers enhanced performance in terms of packet processing.

Additionally, FAN architectures have received recently more attention from the Grid community[1]. For instance, the authors in [123, 124, 125] have evaluated FAN architecture under Grid traffic and showed that FAN architecture performs better than DiffServ under a sample Grid environment. In conclusion FAN architectures are a promising approach for QoS provisioning.

7.4 Extension of FAN to an IP over WDM network architecture

Figure 7.8 proposes an architecture for the Multi-layer Flow-Aware Networking (MFAN) node. As it is illustrated, a MFAN node deals with the optical layer and its subsequent queue, which can be used for routing traffic when the IP layer is near congestion.

The optical switching technology used at the WDM Layer is based on Wavelength Selective Switches (WSS) and ROADM. Data flows to be transmitted at the optical layer are buffered in a simple FIFO queue at the source MFAN node. The transmission technique at the optical layer is based on optical bursts with random size. An intermediate node between the source node and the destination node may insert along the lightpath between this pair of nodes its own optical bursts. Similarly, an intermediate node may extract upstream optical bursts if this is their final destination. Such transparent optical circuit which links a source to multiple destinations is known as a light-trail [126, 127]. Traffic inserted at an intermediate node along a light-trail may be viewed as cross-traffic for the optical layer. We can then estimate that MFAN nodes are located at the ingress and egress nodes of a transparent optical cloud.

MFAN assumes that the QoS provided by FAN at the IP layer is sufficient. For this reason, the extended Multi-layer FAN node uses the IP resources whenever it is possible (that is, as long as $PL < Th_{PL}$ and $FR > Th_{FR}$), and requests extra optical resources once either PL or FR fall out their ranges. Therefore, the MFAN solution does not try to improve the QoS provided to incoming flows, but to enable the acceptation of a new flows at the transparent optical layer which would otherwise be rejected at the opaque IP layer.

In the MFAN architecture, the flows that use the optical resources are stored in the

[1] "It's a very promising technology and has significant potential, addressing a number of issues in a way no one else is today." Joe Mambretti, EETimes, 08/06/2007

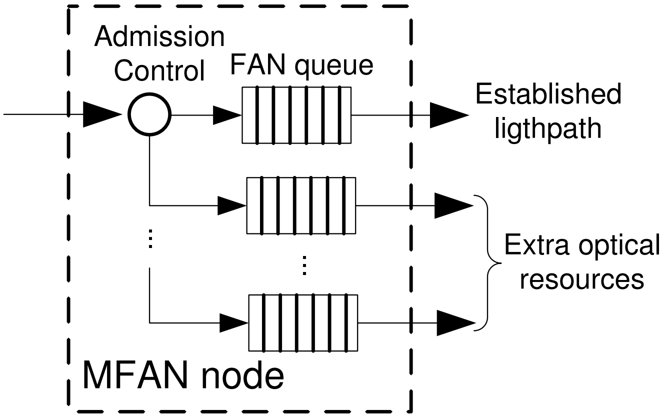


Figure 7.8: Multi-layer Flow-Aware Networking (MFAN) node architecture.

PFLλ list. This list is looked up when incoming packets arrive at the MFAN node, to see whether such packet belongs to an already accepted flow or is part of new flow arrival (see Figure 7.9). If it is a new flow, it is first tried to be routed over the electronic layer (whether $PL < Th_{PL}$ and $FR > Th_{FR}$). If the flow is denied access at the IP layer, then the optical layer is tried, first checking whether there are optical resources available, and secondly evaluating the optical queue threshold (OQ_{th}). If it is successful, then the flow is accepted. In what follows, the optical queue shall be referred as the electrical queue connected to the optical layer.

The following defines three different policies about what to do when the optical layer accepts a flow in a situation of congested electronic layer:

- **Newest-flow policy.** Those incoming flows, which cannot be accepted by the FAN queue, are sent over the optical layer, only if the occupancy of the optical queue is below the threshold OQ_{th}. Admission control is used also in the following two policies.

- **Most-Active-flow policy.** The flow with the greatest "backlog" in the AFL (existing flow) is transferred to the optical layer, thus releasing some space in the electrical layer for the new incoming flow.

- **Oldest-flow policy.** With this policy, the flows that are active and have been around for a longer time in the FAN queue are moved to the optical queue, thus

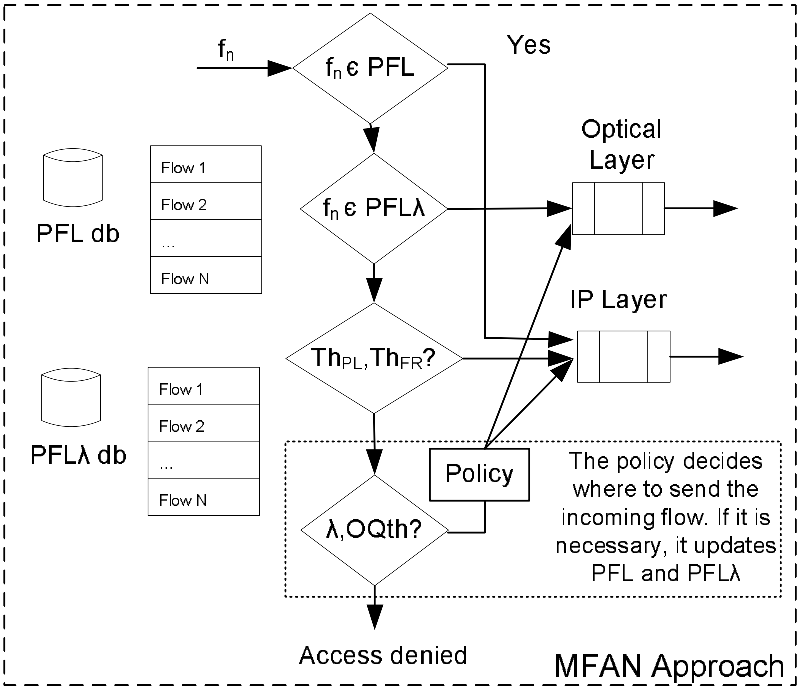


Figure 7.9: MFAN Admission Control Flow Diagram.

releasing space for the new flow arrival. The age of flows is available by FAN in the PFL, thanks to the incoming order to the system.

According to the previous definition of Most-Active- and Oldest-flow Policy, MFAN chooses the flow to send over the optical layer. However, the FAN queue can be congested due to the streaming flows ($PL > Th_{PL}$) or elastic flows ($FR < Th_{FR}$). It is reasonable to switch through the optical layer only the type of flows that are causing such congestion to the FAN queue. This is feasible since flows are classified and stored at the FAN's monitoring system. Figure 7.10 summarizes this idea.

It is important to note that the MFAN system uses only the FAN's monitored parameters and the FIFO queue state (OQ_{th}). Furthermore, the MFAN approach reduces the amount of memory required by FAN. The reason is that, for two links, FAN needs to save two PFL tables as well as two AFL tables. However, the information saved by MFAN is: one PFL, one AFL and one $PFL\lambda$ tables. $PFL\lambda$ does not save any monitoring information, but only the identifiers of the flows routed though the optical queue. Therefore, if FAN scales at Gbps networks, MFAN architecture also does.

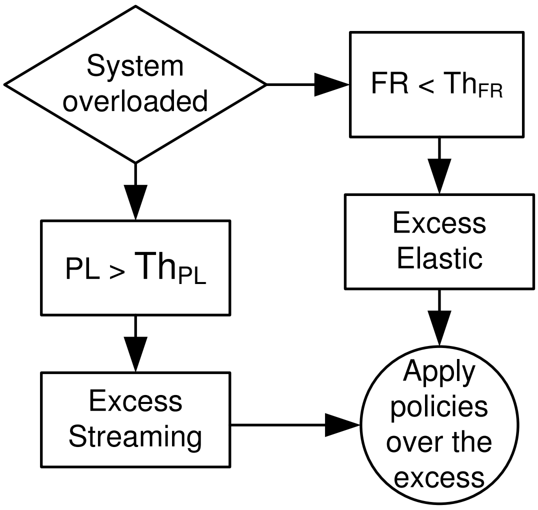


Figure 7.10: Optimal decision of the flows.

7.5 Experiments and results

7.5.1 Simulation scenario description

To study the performance of the three different policies introduced above, we have simulated a scenario with four TCP/Reno and four UDP traffic sources in a two-hop network (see Figure 7.11) using ns2[2]. As illustrated, the light-trail (end-to-end transparent optical link) has been simulated by a direct connection between the first and the third nodes, whereas the IP connection traverses all three nodes. For the sake of simplicity, it is assumed that a single light-trail is available at the optical layer and that bursts are composed by a single IP packet.

This scenario considers the same input traffic parameters used by the authors in [98, 99] to validate FAN. Essentially, flows arrive following a stationary Poisson process, given the fact that the UDP sources (streaming flows) simulate phone calls and TCP (elastic) flow arrivals are well-known to follow this distribution (see [128]). The UDP sources have been simulated with an on/off process, whereby the duration of both periods are exponentially distributed with mean 0.5 seconds. Additionally, the "on"-period rate is 64 Kbps, with constant packet length of 190 bytes. The UDP session length is exponentially distributed with mean 60 seconds. On the other hand, the TCP job size follows a truncated Pareto distribution with tail index $\alpha = 1.5$ and 1 KB mean value, always in the range of 8 KB

[2] http://www.isi.edu/nsnam/ns/

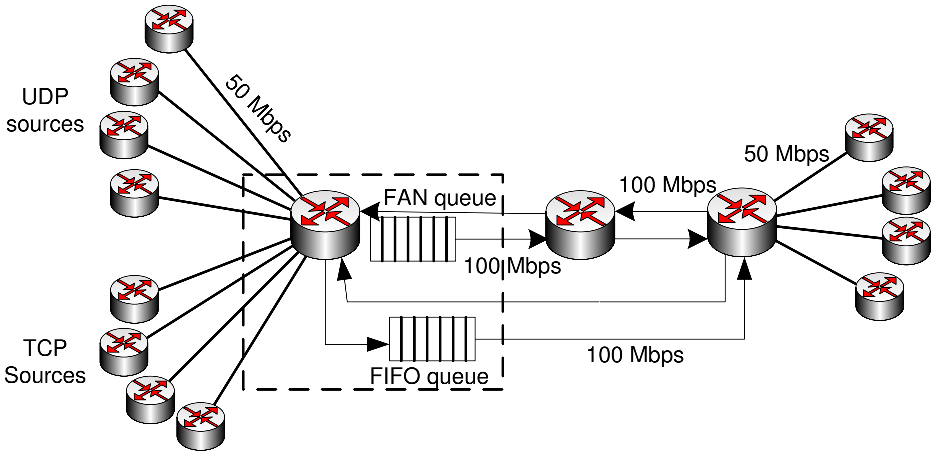


Figure 7.11: MFAN Simulation Scenario.

and 1 MB. Finally, the buffer sizes considered follow the well-known rule of $Q = \overline{RTT} \times C$, as given by [129].

The system's traffic load is considered 110%, in order to study the admission control mechanism. In a first experiment, MFAN is evaluated in a TCP-dominated environment, whereby 80% of the total traffic volume is TCP and the remaining 20% is UDP [99]. In such scenario, it is expected to observe that FR exceeds the threshold value Th_{FR}. However, in the second experiment, we test the MFAN system in a UDP-dominated scenario. In such case, it is expected that $PL > Th_{PL}$. In all the experiments, it is worth remarking that $Th_{PL} = 80\%$ and $Th_{FR} = 5\%$ [99].

With this configuration, we have focused on the following performance metrics: mean delay experienced by the streaming packets and average goodput of the elastic flows in the optical queue. The performance of the FAN queue is not explained here, since it has been validated in previous studies [98, 99]. Besides, one of the MFAN premises is that the QoS given at the IP layer by FAN is adequate.

Finally, the reader should note that the backbone link capacity is 100 Mbps, which is much smaller than typical optical capacities, but significantly reduces the simulation time. The results obtained with this value should remain for higher capacities.

Figure 7.12 summarizes the MTE problem for the MFAN approach. The traffic demands are modeled using the state of the art in TCP and UDP connections, the network resources are our MFAN nodes, which are connected as depicted in Figure 7.11. The objective of our approach is to select the better flows to use the optical resources, when it is mandatory, thus is when FAN queue is saturated. The solution of our problem is this request for extra optical resources and the selection of the most suitable flows.

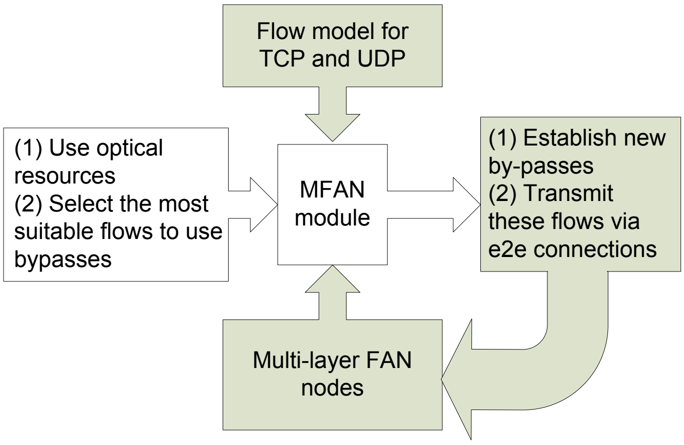

Flow model for TCP and UDP
(1) Use optical resources
(2) Select the most suitable flows to use bypasses
MFAN module
(1) Establish new by-passes
(2) Transmit these flows via e2e connections
Multi-layer FAN nodes

Figure 7.12: Blocks Definition MFAN Problem

7.5.2 Admission control in the optical queue

This first experiment aims to show the benefits of introducing admission control in the optical queue, since this was proposed in FAN to minimize the service degradation that arises under congestion. In light of this, Figure 7.13 shows the results of a simulation example that was carried out both with and without admission control. In both cases, we have considered the Newest-flow policy in the TCP-dominated scenario.

Figure 7.13 illustrates the average goodput for the TCP flows and the mean delay suffered by the UDP packets in the optical queue. As shown, the case without admission control offers less performance (high delay and low goodput) than when admission control is employed. Indeed, this is the case since, the more flows accepted (when no admission control is used) the more load in the queue. Furthermore, it is noteworthy that it is possible to adjust a given desired QoS just by varying the value of the optical queue threshold parameter OQ_{th}.

7.5.3 Implicit classification of FAN

FAN's implicit classification decides which flows are considered streaming (high-priority) and which others are elastic (low-priority). Following the FAN architecture, a situation with Fair Rate under its threshold indicates that the system is congested due to the elastic

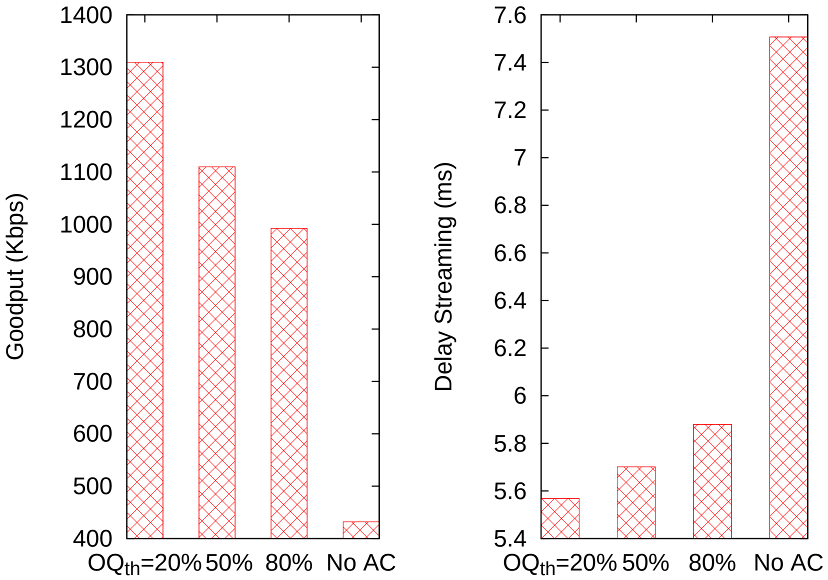


Figure 7.13: Average goodput and delay in the optical queue with and without admission control.

flows, whereas if the Priority Load threshold is exceeded the flows causing congestion are the streaming ones.

For instance, Figure 7.14 shows the evolution of Fair Rate and Priority Load in the TCP-dominated scenario described in Section 7.5.1. As shown, the Fair Rate is out of its nominal range, which means that the system is heavily loaded due to elastic traffic. This is reasonable since 80% of the simulated traffic is TCP. Thus, it makes sense to move the elastic flows to the optical queue, in order to relief the IP layer.

Figure 7.15 illustrates the evolution of FR and PL in the UDP-dominated scenario. In this case, PL is out of range, while FR is most of the simulation time over $Th_{FR} = 5\%$.

The implicit classification information of FAN can be used by the Most-Active-flow and Oldest-flow policies to move to the optical layer the "most appropriate" flows in terms of congestion at the IP layer. When the system is loaded due to elastic flows, some of these should be sent to the optical layer and viceversa. Clearly, the Newest-flow policy makes no use of such information, since it just switches the incoming flows to the optical layer.

7.5.4 Flow routing policies over the optical queue

This experiment aims to study the behaviour of the three policies defined in Section 7.4: Newest-flow, Most-Active-flow and Oldest-flow. As previously stated, the difference among

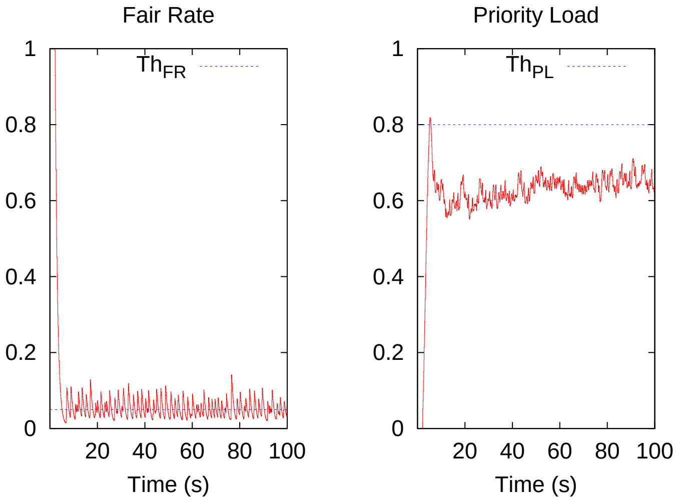


Figure 7.14: Fair Rate and Priority Load evolution in a TCP-dominated scenario (Newest-flow policy and $OQ_{th} = 80\%$).

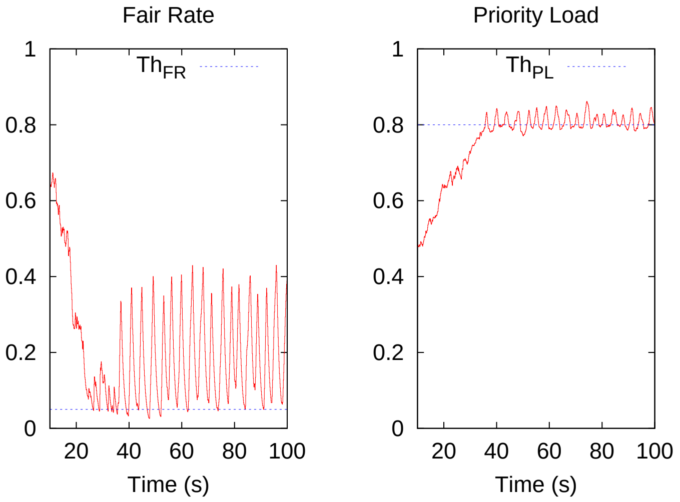


Figure 7.15: Fair Rate and Priority Load evolution in a UDP-dominated scenario (Newest-flow policy and $OQ_{th} = 80\%$).

them is the choice of which flow is to be transmitted over the optical queue upon congestion of the IP layer. Therefore, the following focuses on the performance of the optical queue. Firstly, the results are shown when the system is TCP-dominated and, secondly, the performance of MFAN under a UDP-loaded scenario is studied.

MFAN Performance in an TCP-dominated scenario

Figure 7.16 shows the total number of UDP and TCP flows switched in the optical layer during the 100 seconds that the simulation lasts for the three policies, with OQ_{th} in the range 10% to 90% of the total queue length.

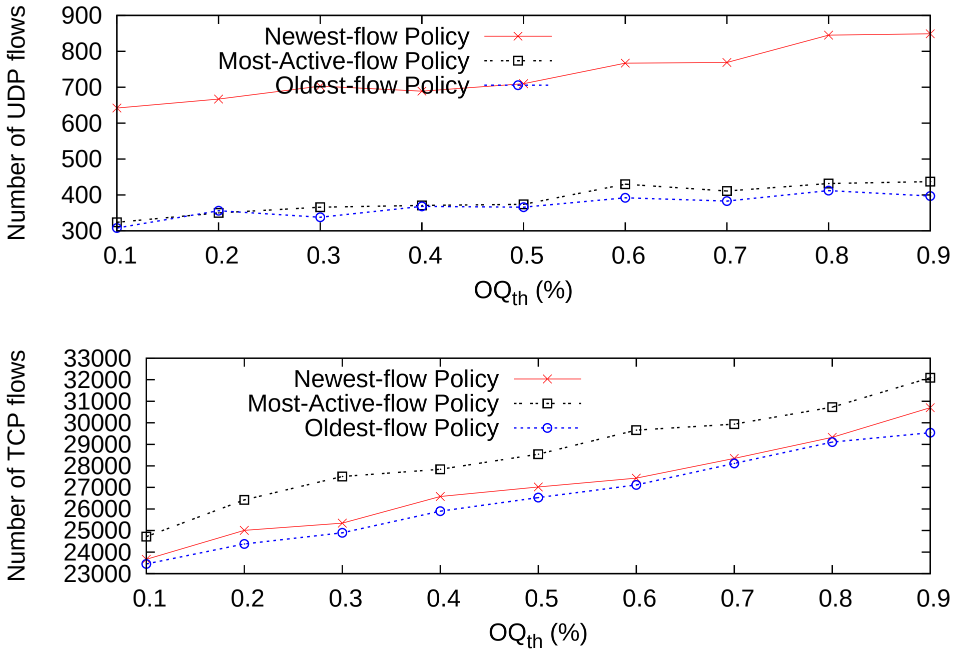


Figure 7.16: Total number of UDP and TCP flows switched in the optical layer (over 100 seconds). TCP-dominated scenario.

First of all, it is important to notice that only a few UDP flows are routed over the optical layer in the cases of Oldest- and Most-Active-flow policies. This is because, with these two policies, only some UDP flows are detected as elastic flows (false positives). On the other hand, it can be seen that the Newest-flow policy sends a greater number of UDP flows through the optical queue. This has a tremendous impact on the performance of the optical queue since the UDP flows, which do not suffer congestion control, increase the overall delay in the optical queue (see Figure 7.17).

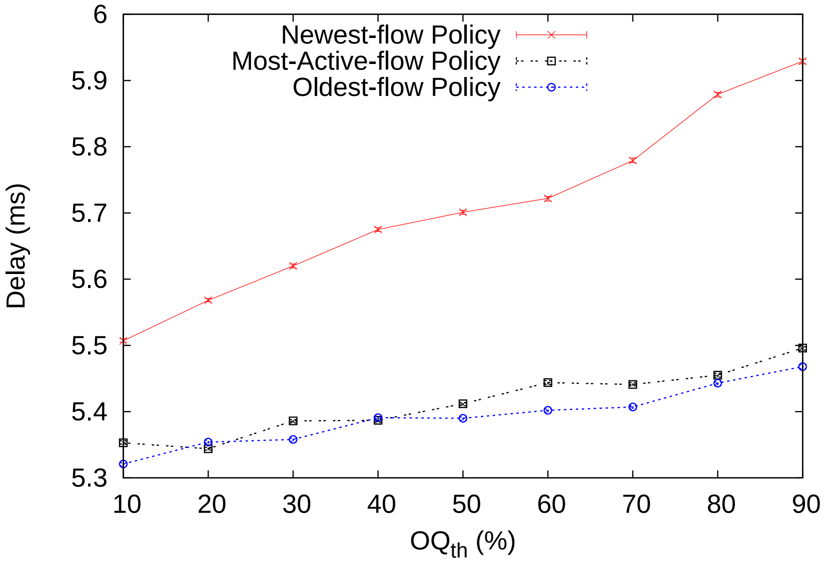


Figure 7.17: Mean delay of the UDP packets in the optical queue (Confidence intervals=95%). TCP-dominated scenario.

Finally, Figure 7.18 shows the average goodput of the TCP flows in the optical queue, for different queueing threshold OQ_{th} values. Again, the Newest-flow policy shows the worst results (that is, low goodput values). Concerning the other two, the Oldest-flow policy presents the best results among the three policies, given that the number of TCP flows accepted is smaller than those accepted by the Most-Active-flow policy (see Figure 7.16 bottom). The reason for this is that the Oldest-flow policy is more accurate at detecting the heaviest flows, since the Most-Active-flow only considers the "backlog", which is a short-term measure of the heaviness of the flows.

MFAN Performance in an UDP-dominated scenario

This section shows the results when the greatest amount of traffic is UDP. Figure 7.19 represents the number of TCP and UDP flows sent to the optical layer in the simulation. The amount of UDP flows is almost the same for all policies, but the Newest-flow policy sends more TCP flows to the optical layer.

According to the amount of UDP and TCP flows in the optical domain (Figure 7.19), the greatest delay is achieved when using the Newest-flow policy (see Figure 7.20). However, the difference among the policies is below 0.1 ms. If we compare this results with the ones in the TCP-dominated scenario (Figure 7.17), we realize that the number of UDP flows is

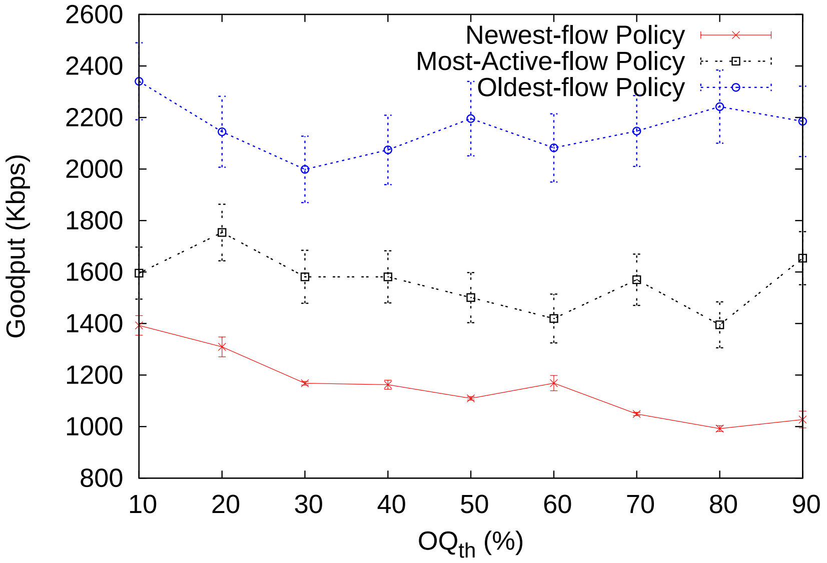


Figure 7.18: TCP flows goodput in the optical queue (Confidence intervals=95%). TCP-dominated scenario.

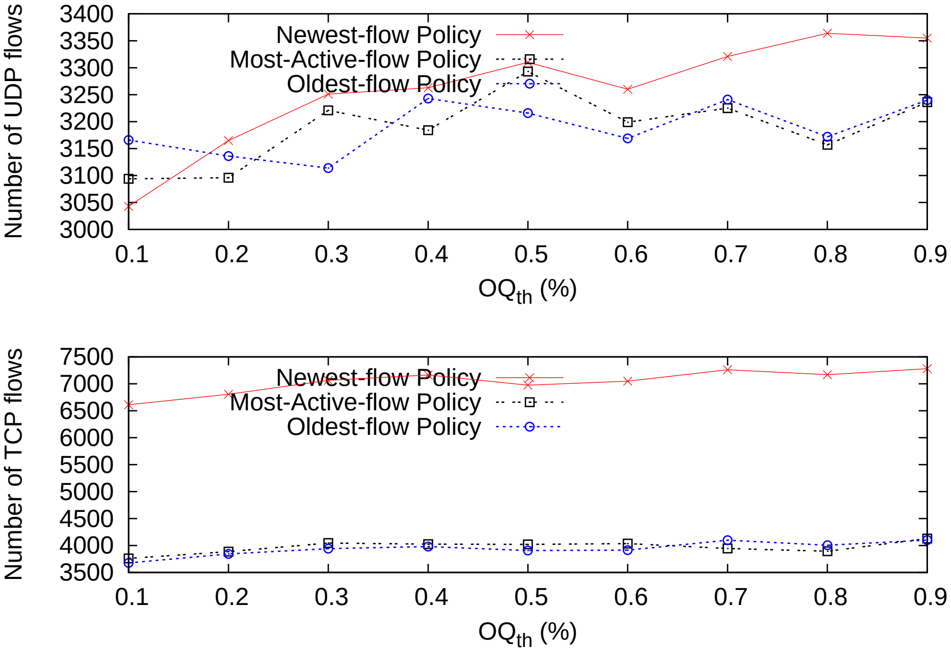


Figure 7.19: Total number of UDP and TCP flows switched in the optical layer (over 100 seconds). UDP-dominated scenario.

decisive for the queuing delay. The number of TCP flows does not greatly influence the UDP queueing delay, because of the congestion avoidance mechanism.

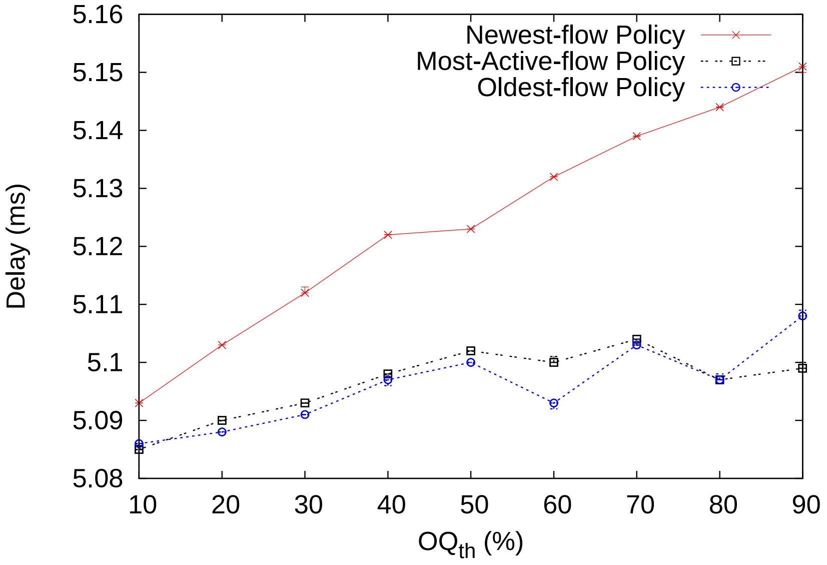


Figure 7.20: Mean delay of the UDP packets in the optical queue (Confidence intervals=95%). UDP-dominated scenario.

The goodput achieved by the TCP flows is represented in Figure 7.21. In this scenario the performance of the Most-Active-flow policy is better than Oldest-flow policy. The Most-Active-flow policy sends to the optical layer the flows which are transmitting a greatest amount of traffic. Moreover, as $PL > Th_{PL}$, this sends only the flows classified as streaming ones. This pattern matches with the TCP flows that are in the slow start phase and do not reach the congestion avoidance phase. When these flows are sent to the optical queue, they can increase their TCP transmission window, thus increasing their goodput. However, “old” TCP flows sent to the optical layer are flows in the congestion avoidance phase, so they do not achieve higher bit rates.

Performance study with different traffic profiles

According to the results of the previous sections, we discover that the performance of the policies depends on the traffic profile observed in the network. The best goodput is achieved when the Oldest-flow policy is used over a TCP-dominated scenario, whereas the Most-Active-flow policy outperforms when the system is mostly loaded by UDP flows. This evolution is shown in Figure 7.22. When the proportion of UDP flows is greater than

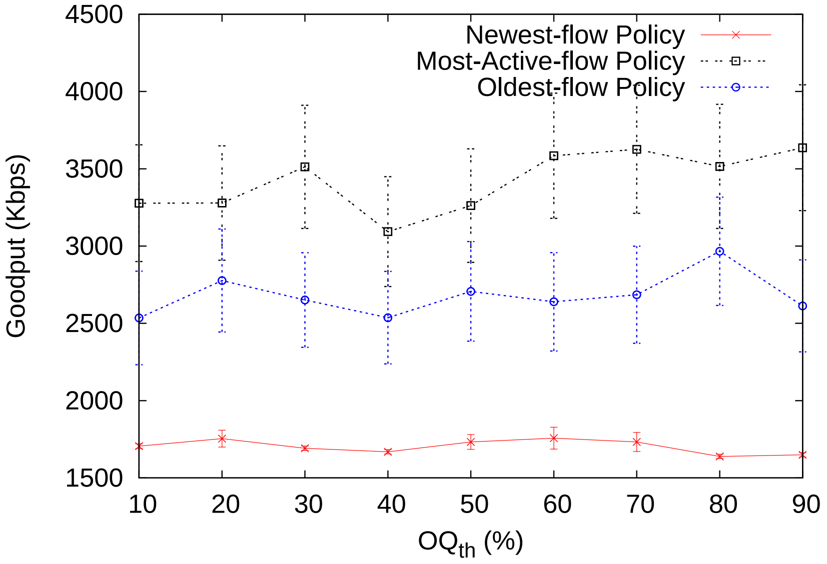


Figure 7.21: TCP flows goodput in the optical queue (Confidence intervals=95%). UDP-dominated scenario.

TCP flows, a better goodput is achieved by the Most-Active-flow policy. The UDP delay is almost the same as with the Oldest- and Most-Active-flow policies. In light of this, a new "Hybrid policy" could be defined on attempts to exploit the benefits of both policies. Such a "Hybrid policy" would use the Most-Active-flow policy when the system is loaded by UDP traffic and it would follow the Oldest-flow one when TCP traffic dominates.

7.6 Conclusions and future work

This work's contributions are two-fold: First, it proposes an extension to the Flow-Aware Networking architecture by including an optical layer upon the congested IP layer. This new architecture is a simple extension of FAN which uses the same monitoring parameters, but includes a new one, the OQ_{th}, to keep FAN's admission control at the optical layer.

And, secondly, this work proposes and analyses three different policies concerning the choice of which flows are moved to the optical layer. The simulations show that the best possible choice, in terms of delay and goodput experienced by the flows, is to switch the heaviest flows found in the IP layer over the optical domain. This is possible using the Most-Active- and Oldest-flow policies which continuously monitor the current flows in the IP layer. Among these two, when the traffic profile is TCP-dominated, the latter is more

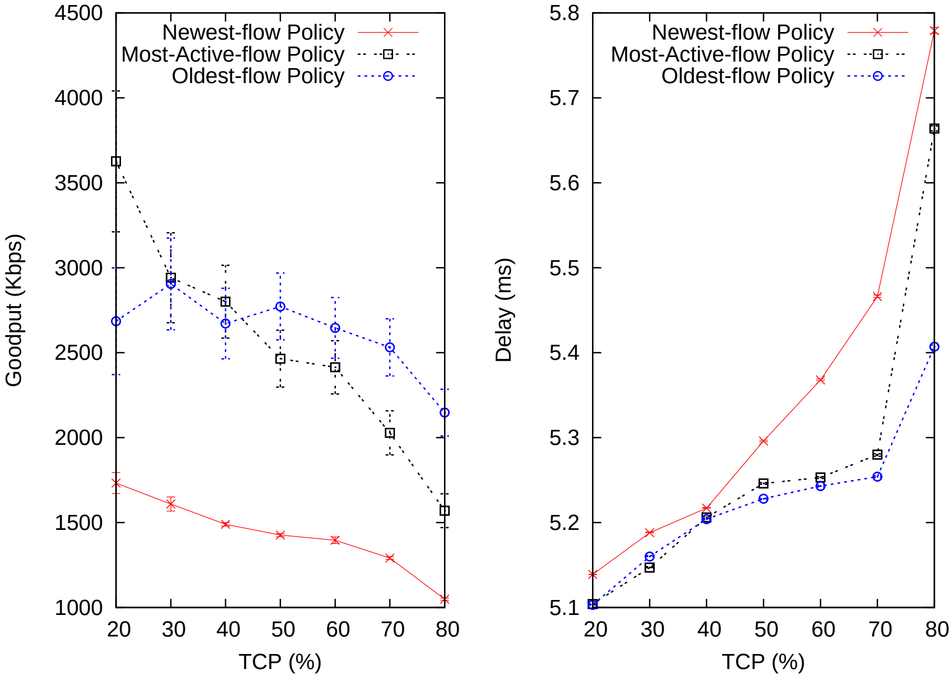


Figure 7.22: Policies performance evaluation when the percentage of TCP varies. ($OQ_{th} = 70\%$)

accurate at detecting the heaviest flows since it monitors flows over a longer period of time. However, if the system had a congestion due to UDP flows, the Most-Active-flow policy shows better results.

As future work a “Hybrid policy”, which combines the benefits of both above, will be studied as a solution for all kind of traffic profiles. Moreover, it is interesting to investigate the performance behaviour of MFAN nodes in a more complex topology and its impact in the optical layer with limited resources. In addition to this, the performance of MFAN shall be evaluated with other traffic applications, such as P2P and Grid.

Chapter 8

Conclusions

This chapter is divided into three section. Section 8.1 summarizes the main contributions of this Ph.D. thesis. Then, the objectives of this work are evaluated in section 8.2. Finally, section 8.3 proposes future work and new research lines.

8.1 Main contributions

This thesis addresses the issue of the end-to-end provisioning of QoS in multi-layer and multi-domain network operators. To achieve such objectives three solutions are proposed: "Threshold-based" algorithms, the Bayesian decisor and Multi-layer FAN (MFAN). A common objective of the three solutions is to improve the QoS in multi-layer networks while managing the optical and electronic resources sensibly. The contributions of this thesis are outlined at the end of each chapter, but the following lists a summary of them:

1. **Multi-layer Traffic Engineering (MTE) strategies are feasible for current networks operators.** Chapter 2 shows that the integration of MTE mechanisms in current network operators is feasible. This chapter not only evaluates the impact at the control plane level, but also at the data plane.
   This contribution has led the following publication:

   - J. E. Gabeiras, V. López, J. Aracil, J. P. Fernández Palacios, C. García Argos, O. González de Dios, F.J. Jiménez Chico and J. A. Hernández: *Is Multi-layer Networking Feasible?*, in Optical Switching and Networking, accepted.

2. **Bayesian decisor for multi-layer and multi-domain architectures.** This thesis proposes a novel methodology to deal with the multi-layer problem of choosing which

amount of traffic should travel through the IP or the optical domain. This Bayesian decisor is defined and validated not only in a single-domain scenario, but also in multi-domain networks.
This contribution has led the following publications:

- V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios O. González de Dios: *Performance evaluation of a Bayesian decisor in a multi-hop IP over WDM network scenario*, in Optical Networking Design and Modeling (ONDM), Feb 2009.
- V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios and O. González de Dios: *A Bayesian decision theory approach for the techno-economic analysis of an all-optical router (extended version)*, in Computer Networks, July 2008, Vol. 52, Issue 10, pp. 1916-1926.
- V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios O. González de Dios: *A Bayesian decision theory approach for the techno-economic analysis of an all-optical router*, in Optical Networking Design and Modeling (ONDM), May 2007. Published in Lecture Notes in Computer Science (LNCS). Within the top five papers award.

3. **Extension of Flow-Aware Networking to multi-layer networks.** The integration of multi-layer networks remains an open issue in QoS solutions for the IP layer. This work proposes the integration of FAN with IP over WDM networks, without the introduction of new monitoring techniques, but just efficiently using FAN's monitoring parameters.
   This contribution has led the following publication:

   - V. López, C. Cárdenas, J. A. Hernández, J. Aracil and M. Gagnaire: Extension of the *Flow-Aware Networking (FAN) architecture to the IP over WDM environment*, in IEEE International Telecommunication NEtworking WorkShop on QoS in Multiservice IP Networks (IT-NEWS/QoS-IP 2008), February 2008. Within the top ten papers award.

8.2 Assessment of the objectives

The specific objectives of this thesis are: “Characterization of the main traffic sources in current IP over WDM networks”, “Definition of techniques to improve the multi-layer integration and offer Quality of Service", "Identify the monitoring parameters and techniques to operate in multi-domain networks providing Quality of Service" and "Integration of the proposed solutions with the current protocols and architectures".

Concerning traffic characterization, chapter 3 validates the Fractional Brownian Motion (FBM) as a model to simulate the behavior of current traffic in networks. Experiments are carried out to test that the queuing performance under a FBM traffic and the state of the art match. This traffic model is assumed for the traffic characterization in the Bayesian decisor model. Chapter 6 evaluates the performance of monitoring delay techniques, based not only on well-known Poisson traffic, but also in a real trace from CAIDA. Such real traffic is not memory-less like the exponential traffic, which enhances the value of this evaluation.

The main contribution of this thesis is the identification of MTE problems and the proposal of different MTE traffic mechanisms. Chapter 2 proposes a "Threshold-based" algorithm, which allows the network operator to avoid congestion at the IP layer, offloading Longest or Largest flows through the lightpaths. The performance of these "Threshold-based" algorithms show that when lower congestion levels are reached at the IP layer, a greater amount of lightpaths are requested to the WDM layer. Usually both objectives performance and resource utilization are opposite. Chapter 3 proposes a Bayesian decisor to trade-off between the performance and the resource utilization. This contribution is very valuable, since it offers a new methodology which can be adapted to the operator requirements. This adaptation is realistic, because QoS constraints as well as resource utilization parameters are included in the model. This novel methodology is extended not only for a single-node, but also to multi-node paths (chapter 4) and to general network topologies (chapter 5). These chapters show the coherent behavior of the Bayesian decisor in all the scenarios. The third contribution to the MTE field is Multi-layer FAN (MFAN). Chapter 7 reviews FAN technology and proposes a node model for a multi-layer scenario. The MFAN proposal does not increase the complexity of FAN and takes advantage of the functionalities provided by it. Three policies to deal with the flow routing problem were defined and evaluated, discussing their performance in scenarios with different traffic profiles.

Chapter 6 evaluates some techniques to monitor the end-to-end delay in single-domain networks. Such single-domain techniques are used to provide a delay model to the Bayesian decisor. This new delay model helps us to define the Bayesian decisor for multi-domain networks. This model is also evaluated in chapter 6, validating the behavior of the Bayesian decisor in multi-domain scenarios.

All proposed solutions are in line with the current standards for backbone networks. Chapter 2 provides a wide discussion about the feasibility of MTE strategies in current network operators. Chapter 6 carries out a detailed description of the current protocols for multi-domain backbone scenarios to show that the Bayesian decisor proposal can be easily integrated in these architectures. Finally, the main QoS technologies (Intserv and Diffserv) are introduced in the chapter 7. They are compared with Flow-Aware Networking (FAN), thus providing a reasonable motivation for Flow-Aware Networking architectures. Multi-layer FAN (MFAN) is proposed as an evolutionary solution for FAN networks, which have to operate in multi-layer topologies. Consequently, the proposals of this dissertation are not just interesting for academia, but also for real operators.

In light of the results, we can say that objectives of this work have been fulfilled. Multi-layer architectures are a novel area with a lot of open problems (routing, dynamic grooming, multi-domain interaction, etc.). This dissertation contributes with the definition of appropriated MTE strategies for the different problems that current network operators have to deal with. The extension to multi-domain networks and the definition of evolutionary solutions enhance the contributions of this work.

8.3 Future work

The results presented throughout this thesis open new research lines for future work that extends the research in multi-layer and multi-domain networks. Let us outline some of these future research lines:

- **Centralized vs. distributed architecture.** The interaction among the entities in our proposal, assumes that a point in the network contains all the state information. This centralized approach is feasible for small networks, but it does not scale to large ones. Therefore, a future research line is how to adapt the “Threshold” and Bayesian decisor algorithms to work distributedly and compare its performance with the centralized solution.

- **Multi-service networks.** The “Threshold” algorithm just takes into account a single threshold for service provisioning. The Bayesian decisor evaluates its performance with three different types of services, but they do not share network resources in our study. However, current networks are migrating to a single infrastructure in order to reduce their OPEX and CAPEX. Consequently, the Bayesian decisor should include mechanisms to share resources among different services and the "Threshold-based" algorithm should include support for different QoS applications.

- **Evaluation of timing parameters in MTE strategies.** Every algorithm included in a network can operate at three levels: Network Planning, Network Engineering and Real-Time Management. Network planning decisions are valid for months or years. Network engineering timing is done in weeks or days, and real-time management in hours, minutes or seconds. The strategies proposed in this work do not consider when the algorithms may be run. The MFAN solution is clearly shown as an on-line mechanism for TE provisioning. The "Threshold-based" and the Bayesian decisor algorithms may be used as short or long term mechanisms. However, to solve such question, it is required to evaluate how such algorithms could be implemented in real network equipment and how the delay of the monitoring information can affect their performance.

- **Economic impact of Multi-layer Traffic Engineering mechanisms.** The deployment of MTE mechanisms in real networks, not only depends on the technical capabilities of the mechanisms, but also on their impact in the CAPEX and the OPEX. This work proposes MTE strategies and a preliminary case study is included in the dissertation. However, a realistic cost model not only for the IP layer, but also for the optical components is mandatory to evaluate the impact of MTE algorithms in economic terms. If we want to evaluate the impact on the OPEX, the evaluation of the timing parameters is also required. So this future line research is joint with the previous future research line.

- **MTE strategies with resilience support.** The scope of this dissertation is not focused on resilience issues, but this is a key topic for the development of MTE strategies in operators. The integration of these MTE strategies should support the interaction with resilience techniques that are under research for backbone technologies.

Conclusiones

Este capítulo tiene dos secciones. La primera resume las contribuciones principales de la tesis y se evalúan los objetivos de esta tesis en la segunda sección.

Contribuciones principales

Esta tesis estudia la provisión de calidad de servicio extremo a extremo en redes multicapa y multidominio. Para conseguir este objetivo se han propuesto tres soluciones: algoritmos basados en umbrales, el decisor bayesiano y Multi-layer FAN (MFAN). Un objetivo común de las tres soluciones es mejorar la calidad de servicio en redes multicapa utilizando los recursos de la capa IP y óptica cuando son necesarios. Las contribuciones de esta tesis se muestran en cada capítulo, pero la siguiente lista resume las principales:

1. **Las estrategias de Ingeniería de Tráfico multicapa son viables para los operadores de red actuales.** El capítulo 2 muestra que la integración de mecanismos de Ingeniería de Tráfico multicapa es posible. Este capítulo no sólo evalúa el impacto del plano de control, sino también el plano de datos.
   Esta contribución ha dado lugar a la siguiente publicación:

   - J. E. Gabeiras, V. López, J. Aracil, J. P. Fernández Palacios, C. García Argos, O. González de Dios, F.J. Jiménez Chico and J. A. Hernández: *Is Multi-layer Networking Feasible?*, in Optical Switching and Networking, aceptado.

2. **Decisor bayesiano para redes multicapa y multidominio.** Esta tesis propone una metodología novedosa para tratar el problema multicapa sobre la cantidad de tráfico que debe ser enviado usando los recursos IP y ópticos. Este decisor bayesiano está definido y validado no sólo para un único dominio, sino para redes multidominio.
   Esta contribución ha dado lugar a la siguiente publicaciones:

- V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios O. González de Dios: *Performance evaluation of a Bayesian decisor in a multi-hop IP over WDM network scenario*, in Optical Networking Design and Modeling (ONDM), Feb 2009.
- V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios and O. González de Dios: *A Bayesian decision theory approach for the techno-economic analysis of an all-optical router (extended version)*, in Computer Networks, July 2008, Vol. 52, Issue 10, pp. 1916-1926.
- V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios O. González de Dios: *A Bayesian decision theory approach for the techno-economic analysis of an all-optical router*, in Optical Networking Design and Modeling (ONDM), May 2007. Published in Lecture Notes in Computer Science (LNCS). Entre los cinco mejores artículos.

3. **Extensión de Flow-Aware Networking a redes multicapa.** La integración de redes multicapa con soluciones de calidad de servicio en la capa IP es aún una línea de investigación abierta. Este trabajo propone la integración de FAN con redes IP sobre WDM, sin añadir complejidad, sino usando los parámetros de monitorización de FAN.
   Esta contribución ha dado lugar a la siguiente publicación:

   - V. López, C. Cárdenas, J. A. Hernández, J. Aracil and M. Gagnaire: Extension of the *Flow-Aware Networking (FAN) architecture to the IP over WDM environment*, in IEEE International Telecommunication NEtworking WorkShop on QoS in Multiservice IP Networks (IT-NEWS/QoS-IP 2008), February 2008. Entre los diez mejores artículos.

Evaluación de los objetivos

Los objetivos específicos de esta tesis son: “Estudiar las fuentes de tráfico en las redes IP y ópticas y encontrar modelos que permitan analizar el impacto en las redes.”, “Encontrar técnicas para mejorar la integración multicapa y ofrecer calidad de servicio”, “Identificar los parámetros principales a monitorizar en un escenario que pretende trabajar con múltiples dominios y ofrecer calidad de servicio” e “Integrar de las soluciones propuestas con los protocolos y arquitecturas comúnmente utilizadas”.

Respecto a la caracterización de tráfico, el capítulo 3 valida el proceso “Fractional Brownian Motion (FBM)” como un modelo adecuado para simular el comportamiento del tráfico en redes actuales. Se han realizado experimentos para comprobar que el rendimiento de una cola con tráfico FBM y el estado del arte encajan. Este modelo de tráfico se utiliza como modelo para el decisor bayesiano. El capítulo 6 evalúa el rendimiento de técnicas de monitorización de retardo con tráfico poisoniano y con una traza real. Este tráfico real no es un proceso sin memoria lo cual aumenta el valor del estudio.

La principal contribución de esta tesis es la indentificación de problemas de Ingeniería de Tráfico multicapa y la propuesta de algoritmos para resolver dichos problemas. El capítulo 2 propone un algortimo basado en umbral, que permite al operador reducir la congestión en la capa IP, descargando el tráficoyer en la malla fotónica. El rendimiento de estos algoritmos muestra que cuando se obtiene una menor congestión en la red se utiliza una mayor cantidad de recursos en la capa óptica. Normalmente, el rendimiento y la utilización de recursos son dos objetivos contrarios. El capítulo 3 propone un decisor bayesiano para buscar un compromiso entre el rendimiento y la utilización de recursos. Esta contribución es importante ya que ofrece una nueva metodología que se adapta a los requisitos del operador. Esta adaptación es realista, ya que se incluye calidad de servicio y la utilización de los recursos en el modelo. Este metodología se ha desarrollado no sólo para el caso de un único nodo, sin para múltiples nodos (capítulo 4) y para topologías generales (capítulos 5). Estos capítulos muestran el comportamiento del decisor bayesiano en todos los escenarios. La tercera contribución en este objetivos es Multi-layer FAN (MFAN). El capítulo 7 revisa la tecnología FAN y propone un modelo de nodo para un escenario multicapa. Esta propuesta no aumenta la complejidad de FAN, sino que utiliza los parámetros de monitorización de FAN. Tres políticas para decidir que tipo de tráfico se debe enviar a la capa óptica se han definido, evaluado y se ha mostrado su rendimiento con distintos perfiles de tráfico.

El capítulo 6 evalúa algunas técnicas para monitorizar el retardo extremo a extremo intra-dominio. Estas técnicas son utilizadas para dar un modelo de retardo al decisor bayesiano. Este nuevo modelo de retardo permite definir el decisor bayesiano para redes multi-dominio. Este modelo es evaluado en el capítulo 6, validando el comportamiento del decisor bayesiano en escenarios multidominio.

Todas las soluciones propuestas están alineadas con los estándares para redes troncales. El capítulo 2 ofrece una amplia discusión sobre la viabilidad de estrategias de Ingeniería de Tráfico multicapa en operadores de red actuales. El capítulo 6 lleva a cabo una detallada descripción de los protocolos actuales en redes multidominio para troncales de red. Muestra que el decisor bayesiano puede ser integrado fácilmente en esta arquitectura. Las principales tecnologías de calidad de servicio (Intserv y Diffserv) son introducidas en el capítulo 7. Estas tecnologías son comparadas con Flow-Aware Networking (FAN), ofreciendo una motivación razonada para las arquitecturas FAN. Multi-layer FAN (MFAN) extiende FAN en redes multicapa, por lo que no añade complejidad y podría integrarse con facilidad. Por lo tanto, las propuestas de este trabajo no sólo quedan justificadas desde el punto de vista académico, sino para operadores reales.

Tras esta evaluación se puede ver que se han cumplido los objetivos. Las arquitecturas multicapa son un area con problemas abiertos. Esta tesis contribuye con la definición de estrategias multicapa para diferentes problemas que los operadores de red tiene que solucionar. La extensión para redes multidominios aumenta la contribución de este trabajo.

References

[1] A. Martínez, J. Aracil, and J. de Vergara, “Optimizing offset times in Optical Burst Switching networks with variable Burst Control Packets sojourn times,” *Optical Switching and Networking*, vol. 4, no. 3-4, pp. 189–199, 2007.

[2] Cisco Systems Inc., “Converge IP and DWDM Layers in the Core Network,” 2007, White Paper.

[3] M. Maier, *Optical Switching Networks*, 1st ed. Cambridge University Press, February 2008.

[4] H. Ishio, J. Minowa, and K. Nosu, “Review and status of wavelength-division-multiplexing technology and its application,” *Lightwave Technology, Journal of*, vol. 2, no. 4, pp. 448–463, 1984.

[5] A. H. Gnauck, G. Charlet, P. Tran, P. Winzer, C. Doerr, J. Centanni, E. Burrows, T. Kawanishi, T. Sakamoto, and K. Higuma, “25.6-Tb/s C+L-Band Transmission of Polarization-Multiplexed RZ-DQPSK Signals,” in *Conference on Optical Fiber Communication/National Fiber Optic Engineers Conference (OFC/NFOEC)*, 2007.

[6] T. S. El-Bawab, *Optical Switching*, 1st ed. Springer, April 2006.

[7] IETF, “Generalized Multi-Protocol Label Switching (GMPLS) Architecture (RFC 3945),” October 2004.

[8] ITU-T, “Architecture for the Automatically Switched Optical Network (ASON) - Rec. 8080/Y.1304,” 2001.

[9] IETF, “Fast Reroute Extensions to RSVP-TE for LSP Tunnels (RFC 4090),” May 2005.

[10] Cisco Systems Inc., "Cost Comparison Between IP-over-DWDM and Other Core Transport Architectures," 2007, White Paper.

[11] M. N. Ellanti, S. S. Gorshe, L. G. Raman, and W. D. Grover, *Next Generation Transport Networks: Data, Management, and Control Planes*, 1st ed. Springer, April 2005.

[12] N. Larkin, D. Limited, and U. Enfield, "ASON and GMPLS The battle of the optical control plane," 2002. [Online]. Available: http://www.dataconnection.com/network/download/whitepapers/asongmpls.pdf

[13] P. Tomsu and C. Schmutzer, *Next Generation Optical Networks: The Convergence of IP Intelligence and Optical Technologies (Radia Perlman Series in Computer Networking and Security)*. Prentice Hall PTR, September 2001.

[14] IETF, "Requirements for Traffic Engineering Over MPLS (RFC 2702)," 1999.

[15] ——, "Generalized MPLS Signaling - RSVP-TE Extensions (RFC 3473)," January 2003.

[16] ——, "Generalized MPLS Signaling - CR-LDP Extensions (RFC 3472)," January 2003.

[17] ——, "Traffic Engineering (TE) Extensions to OSPF Version 2 (RFC 3630)," September 2003.

[18] ——, "Intermediate System to Intermediate System (IS-IS) Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS) (RFC 4205)," October 2005.

[19] ——, "Link Management Protocol (LMP) (RFC 4204)," October 2005.

[20] K. Sato, N. Yamanaka, Y. Takigawa, M. Koga, S. Okamoto, K. Shiomoto, E. Oki, and W. Imajuku, "GMPLS-based photonic multilayer router (Hikari router) architecture: an overview of traffic engineering and signaling technology," *IEEE Communications Magazine*, vol. 40, no. 3, pp. 96–101, March 2002.

[21] IETF, "IP over Optical Networks: A Framework (RFC 3717)," March 2004.

[22] Optical Internetworking Forum (OIF), "User to Network Interface Signaling Specification," April 2001.

[23] Cisco Systems Inc., "Cisco Segmented Generalized Multiprotocol Label Switching for the IP Next-Generation Network," April 2006, White Paper.

[24] IETF, "Overview and Principles of Internet Traffic Engineering (RFC 3272)," May 2002.

[25] R. Dutta, A. E. Kamal, and G. N. Rouskas, Eds., *Traffic Grooming for Optical Networks: Foundations, Techniques and Frontiers*, 1st ed. Springer, August 2008.

[26] S. Roh, W. So, and Y. Kim, "Design and Performance Evaluation of Traffic Grooming Algorithms in WDM Multi-Ring Networks," *Photonic Network Communications*, vol. 3, no. 4, pp. 335–348, 2001.

[27] P. Arijs, D. Colle, and P. Demeester, "Optimization models for cost savings in stacked ring network design," in *Informi Telecommunications Conference*, 2000.

[28] O. Gerstel, R. Ramaswami, G. Sasaki, and S. Xros, "Cost-effective traffic grooming in WDM rings," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 618–630, 2000.

[29] H. Zhu and B. Mukherjee, "A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks," *IEEE/ACM Transactions on Networking*, vol. 11, no. 2, pp. 285–299, 2003.

[30] F. Farahmand, X. Huang, and J. Jue, "Efficient Online Traffic Grooming Algorithms in WDM Mesh Networks with Drop-and-Continue Node Architecture," in *IEEE Broadnets*, October 2004.

[31] J. Kuri, N. Puech, M. Gagnaire, E. Dotaro, and R. Douville, "Routing and wavelength assignment of scheduled lightpath demands," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 8, pp. 1231–1240, October 2003.

[32] B. Puype, Q. Yan, D. Colle, S. De Maesschalck, I. Lievens, M. Pickavet, and P. Demeester, "Multi-layer traffic engineering in data-centric optical networks," in *Optical Networking Design and Modeling (ONDM)*, 2003.

[33] J. Comellas, R. Martinez, J. Prat, V. Sales, and G. Junyent, "Integrated IP/WDM routing in GMPLS-based optical networks," *IEEE Network*, vol. 17, no. 2, pp. 22–27, March 2003.

[34] M. Gagnaire, M. Koubaa, and N. Puech, "Network Dimensioning under Scheduled and Random Lightpath Demands in All-Optical WDM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 9, pp. 58–67, 2007.

[35] J. Jimenez, O. Gonzalez, B. Puype, T. Cinkler, P. Hegyi, R. Muñoz, R. Martínez, F. Gálan, and R. Morro, "Multilayer traffic engineering experimental demonstrator in the Nobel-II project," in *BroadBand Europe*, 2007.

[36] J. E. Gabeiras, V. López, J. Aracil, J. Fernández, C. García, O. González, F. Jiménez, and J. Hernández, "Is Multi-layer Networking Feasible?" to appear.

[37] Y. Nakahira, "Traffic Driven IP Optical Path Rearrange Network System Using PSC/LSC Multi-Layer GMPLS," in *European Conference on Optical Communications (ECOC)*, 2004.

[38] M. Ruffini, D. O'Mahony, and L. Doyle, "A Testbed Demonstrating Optical IP Switching (OIS) in Disaggregated Network Architectures," in *IEEE Tridentcom*, 2006, pp. 1–3.

[39] A. Taniguchi, Y. Sameshima, S. Okamoto, T. Otani, Y. Okano, Y. Tsukishima, and W. Imajuku, "Operational Evaluation of ASON/GMPLS Interdomain Capability over a JGN II Network Testbed," *IEEE Communications Magazine*, vol. 46, no. 5, pp. 60–66, May 2008.

[40] H. Kojima, T. Takeda, and I. Inoue, "A study on multilayer service network architecture in IP optical networks," in *Asia-Pacific Conference on Communications (APCC)*, vol. 2, 2003.

[41] L. R. Ford and D. R. Fulkerson, *Flows in Networks (Rand Corporation Research Studies Series)*. Princeton Univ Pr, June 1962.

[42] IST Project FP6-506760, "Next generation Optical network for Broadband European Leadership (NOBEL)." [Online]. Available: http://www.ist-nobel.org/

[43] M. Ruffini, D. O'Mahony, and L. Doyle, "A cost analysis of Optical IP Switching in new generation optical networks," in *International Conference on Photonics in Switching*, October 2006, pp. 1–3.

[44] M. Ruffini, D. Kilper, D. O'Mahony, and L. Doyle, "Cost Study of Dynamically Transparent Networks," in *Conference on Optical Fiber Communication/National Fiber Optic Engineers Conference (OFC/NFOEC)*, 2008, pp. 1–3.

[45] W. Wei, Q. Zeng, Y. Ouyang, and D. Lomone, "High-Performance Hybrid-Switching Optical Router for IP over WDM Integration," *Photonic Network Communications*, vol. 9, no. 139-155, p. 17, March 2005.

[46] K. Matsui, T. Sakurai, M. Kaneda, J. Murayama, and H. Ishi, "A multi-layered traffic engineering architecture for the electronic/optical hybrid network," in *IEEE Pacific Rim Conference on Communications, Computers and signal Processing (PACRIM)*, vol. 1, August 2003, pp. 293–296.

[47] T. Kimura, K. Kabashima, M. Aoki, and S. Urushidani, "Proposal and Comparison of QoS Schemes for IP-over-Optical Multilayer Networks," *IEICE Transactions on Communications*, vol. 88, pp. 3895–3903, 2005.

[48] W. Colitti, K. Steenhaut, A. Nowe, E. E. Monreal, and L. Tran, "Multilayer QoS in integrated IP over Optical Networks," in *International Conference on Communications and Electronics (ICCE)*, October 2006, pp. 1–6.

[49] A. Elwalid, D. Mitra, and Q. Wang, "Distributed Nonlinear Integer Optimization for Data-Optical Internetworking," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1502–1513, August 2006.

[50] S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1176–1188, September 1995.

[51] Y. Choi and S. Bahk, "QoS scheduling for multimedia traffic in packet data cellular networks," in *IEEE International Conference on Communications (ICC)*, vol. 1, May 2003, pp. 358–362.

[52] ITU-T, "ITU-T Recommendation Y.1541 - Network Performance Objectives for IP-Based Services," February 2003.

[53] 3GPP, "3GPP Recommendation s.r0035-0 v1.0. - Quality of Service," September 2002.

[54] ITU-T, "ITU-T Recommendation G.107 : The E-model, a computational model for use in transmission plannings," March 2005.

[55] A. Erramilli, P. Pruthi, and W. Willinger, "Self-Similarity in High Speed Network Traffic Measurements: Fact or Artefact," in *VTT symposium*, 1995.

[56] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 953–962, August 1995.

[57] R. G. Clegg, "A practical guide to measuring the Hurst parameter," *International Journal of Simulation: Systems, Science & Technology*, vol. 7, pp. 3–14, 2006.

[58] S. French and D. Ríos Insúa, *Statistical decision theory.* Oxford University Press Inc., 2000.

[59] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, and C. Diot, "Measurement and analysis of single-hop delay on an IP backbone network," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 908–921, August 2003.

[60] J. A. Hernández and I. W. Phillips, "Weibull mixture model to characterise end-to-end Internet delay at coarse time-scales," *IEE Proc. Communications*, vol. 153, no. 2, pp. 295–304, April 2005.

[61] R. G. Clegg, "Markov-modulated on/off processes for long-range dependent internet traffic," *ArXiv Computer Science e-prints*, October 2006.

[62] M. Pioro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks (The Morgan Kaufmann Series in Networking)*, 1st ed. Morgan Kaufmann, July 2004.

[63] "DRAGON (Dynamic Resource Allocation via GMPLS Optical Networks)." [Online]. Available: http://dragon.east.isi.edu

[64] IETF, "A Border Gateway Protocol 4 (BGP-4) (RFC 1771)," March 1995.

[65] IEFT, "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering (RFC 4726)," November 2006.

[66] IETF, "Inter-Domain MPLS and GMPLS Traffic Engineering - Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions (RFC 5151)," November 2006.

[67] S. Okamoto, H. Otsuki, and T. Otani, “Multi-ASON and GMPLS network domain interworking challenges,” *IEEE Communications Magazine*, vol. 46, no. 6, pp. 88–93, 2008.

[68] R. Munoz, R. Martınez, F. Galan, R. Morro, H. Foisel, S. Szuppa, J. Jimenez, O. Gonzalez, H. Dentler, E. Escalona *et al.*, “Experimental interconnection and interworking of the multi-domain (ASON-GMPLS) and multi-layer (TDM-LSC) NOBEL2 Testbeds,” in *European Conference on Optical Communications (ECOC)*, 2007.

[69] IETF, “A Path Computation Element (PCE)-Based Architecture (RFC 4655),” August 2006.

[70] R. Douville, J. Le Roux, J. Rougier, and S. Secci, “A service plane over the PCE architecture for automatic multidomain connection-oriented services,” *IEEE Communications Magazine*, vol. 46, no. 6, pp. 94–102, 2008.

[71] F. Aslam, Z. Uzmi, and A. Farrel, “Interdomain path computation: Challenges and Solutions for Label Switched Networks,” *IEEE Communications Magazine*, vol. 45, no. 10, pp. 94–101, 2007.

[72] N. Ghani, Q. Liu, A. Gumaste, D. Benhaddou, N. Rao, and T. Lehman, “Control Plane Design in Multidomain/Multilayer Optical Networks,” *IEEE Communications Magazine*, vol. 46, no. 6, pp. 78–87, 2008.

[73] S. Sanchez-Lopez, X. Masip-Bruin, E. Marin-Tordera, J. Sole-Pareta, and J. Domingo-Pascual, “A hierarchical routing approach for GMPLS based control plane for ASON,” in *IEEE International Conference on Communications (ICC)*, vol. 3, 2005.

[74] Q. Liu, M. Kök, N. Ghani, and A. Gumaste, “Hierarchical routing in multi-domain optical networks,” *Computer Communications*, vol. 30, no. 1, pp. 122–131, 2006.

[75] X. Yang and B. Ramamurthy, “Inter-domain dynamic routing in multi-layer optical transport networks,” in *IEEE Global Telecommunications Conference (Globecom)*, vol. 5, 2003.

[76] Y. Zhu, A. Jukan, M. Ammar, and W. Alanqar, “End-to-End service provisioning in multigranularity multidomain optical networks,” in *IEEE International Conference on Communications (ICC)*, vol. 3, 2004.

[77] A. Hadjiantonis, M. Ali, H. Chamas, W. Bjorkman, S. Elby, A. Khalil, and G. Ellinas, “Evolution to a Converged Layer 1, 2 in a Global-Scale, Native Ethernet Over WDM-Based Optical Networking Architecture,” *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 5, pp. 1048–1058, 2007.

[78] F. Verdi, M. Magalhaes, and E. Madeira, “On the Performance of Interdomain Provisioning of Connections in Optical Networks using Web Services,” in *IEEE International Symposium on Computers and Communications (ISCC)*, 2006.

[79] T. Lehman, X. Yang, C. Guok, N. Rao, A. Lake, J. Vollbrecht, and N. Ghani, “Control Plane Architecture and Design Considerations for Multi-Service, Multi-Layer, Multi-Domain Hybrid Networks,” in *High-Speed Networks Workshop*, 2007, pp. 67–71.

[80] M. Yannuzzi, X. Masip-Bruin, and O. Bonaventure, “Open issues in interdomain routing: a survey,” *IEEE Network*, vol. 19, no. 6, pp. 49–56, 2005.

[81] R. Dutta, A. Kamal, and G. Rouskas, *Traffic Grooming for Optical Networks: Foundations and Techniques.* Springer, 2008.

[82] V. Paxson, “Strategies for sound internet measurement,” in *ACM SIGCOMM conference on Internet measurement.* ACM New York, NY, USA, 2004, pp. 263–271.

[83] IETF, “A One-way Delay Metric for IPPM (RFC 2679),” September 1999.

[84] K. Claffy, “Measuring the Internet,” *IEEE Internet Computing*, vol. 4, no. 1, pp. 73–75, 2000.

[85] CAIDA OC48 Trace Project, “CAIDA OC48 Traces 2003-04-24 (collection).” [Online]. Available: http://www.caida.org/

[86] A. Dempster, N. Laird, D. Rubin *et al.*, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[87] M. Yannuzzi, X. Masip-Bruin, S. Sánchez-López, E. Tordera, J. Solé-Pareta, and J. Domingo-Pascual, “Interdomain RWA based on stochastic estimation methods and adaptive filtering for optical networks,” in *IEEE Global Telecommunications Conference (Globecom)*, 2006.

[88] W. Lee, "Spanning Tree Method for Link State Aggregation in Large Communication Networks," in *IEEE Infocom*, vol. 95, 1995, pp. 297–302.

[89] S. Nelakuditi, Z. Zhang, R. Tsang, and D. Du, "Adaptive Proportional Routing: A Localized QoS Routing Approach," *IEEE/ACM Transactions on Networking*, vol. 10, no. 6, pp. 790–804, 2002.

[90] K. Lui, K. Nahrstedt, and S. Chen, "Routing with topology aggregation in delay-bandwidth sensitive networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 1, pp. 17–29, 2004.

[91] J. Szigeti, I. Ballok, and T. Cinkler, "Efficiency of information update strategies for automatically switched multi-domain optical networks," in *International Conference on Transparent Optical Networks (ICTON)*, vol. 1, 2005.

[92] G. Liu, C. Ji, and V. Chan, "On the scalability of network management information for inter-domain light-path assessment," *IEEE/ACM Transactions on Networking*, vol. 13, no. 1, pp. 160–172, 2005.

[93] Y. Xiao, X. Du, J. Zhang, F. Hu, and S. Guizani, "Internet Protocol Television (IPTV): The Killer Application for the Next-Generation Internet," in *IEEE Communications Magazine*, vol. 45, no. 11, Toronto, Ont., Canada,, November 2007, pp. 126–134.

[94] TeleGeography, "International telecommunications traffic," 2007. [Online]. Available: http://www.telegeography.com

[95] P. Pan and H. Schulzrinne, "YESSIR: a simple reservation mechanism for the Internet," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 2, pp. 89–101, 1999.

[96] J. Roberts, "Quality of Service by flow-aware networking," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 358, no. 1773, pp. 2197–2207, 2000.

[97] N. Benameur, A. Kortebi, S. Oueslati, and J. W. Roberts, "Selective service protection in overload: differentiated services or per-flow admission control?" in *International Telecommunications Network Strategy and Planning Symposium (NETWORKS)*, June 2004, pp. 217–222.

[98] A. Kortebi, S. Oueslati, and J. W. Roberts, “Cross-protect: implicit service differentiation and admission control,” in *Workshop on High Performance Switching and Routing HPSR.*, 2004, pp. 56–60.

[99] ——, “Implicit Service Differentiation using Deficit Round Robin,” in *International Teletraffic Congress (ITC)*, August 2005.

[100] S. De Maesschalck, M. Pickavet, D. Colle, and P. Demeester, “Multi-layer traffic grooming in networks with an IP/MPLS layer on top of a meshed optical layer,” in *IEEE Global Telecommunications Conference (Globecom)*, vol. 5, December 2003, pp. 2750–2754.

[101] B. Puype, Q. Yan, S. De Maesschalck, D. Colle, M. Pickavet, and P. Demeester, “Optical cost metrics in multi-layer traffic engineering for IP-over-optical networks,” in *International Conference on Transparent Optical Networks (ICTON)*, vol. 1, July 2004, pp. 75–80.

[102] M. Vigoureux, B. Berde, L. Andersson, T. Cinkler, L. Levrau, M. Ondata, D. Colle, J. Fernandez-Palacios, and M. Jager, “Multilayer traffic engineering for GMPLS-enabled networks,” *IEEE Communications Magazine*, vol. 43, no. 7, pp. 44–50, July 2005.

[103] M. Hirano, M. Aoki, N. Matsuura, T. Kurimoto, T. Miyamura, M. Goshima, and S. Urushidani, “A scalable switch architecture for ultra-large IP and lambda switch routers,” in *International Conference on Telecommunications (ICT)*, vol. 2, 2003.

[104] G. Lee and J. Choi, “Flow Classification for IP Differentiated Service in Optical Hybrid Switching Network,” *Lecture Notes in Computer Science*, pp. 635–642, 2005.

[105] W. Colitti, K. Steenhaut, and A. Nowe, “Multilayer traffic engineering and DiffServ in the next generation internet,” in *International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE)*, 2008, pp. 591–598.

[106] K. Lan and J. Heidemann, “A measurement study of correlations of Internet flow characteristics,” *Computer Networks*, vol. 50, no. 1, pp. 46–62, 2006.

[107] T. Fioreze, M. Wolbers, R. van de Meent, and A. Pras, “Offloading IP Flows onto Lambda-Connections,” *Lecture Notes in Computer Science*, vol. 4785, p. 183, 2007.

[108] IETF, “Integrated Services in the Internet Architecture: an Overview (RFC 1633),” June 1994.

[109] ——, “Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers (RFC 2474),” December 1998.

[110] ——, “An Architecture for Differentiated Services - RFC 2475,” December 1998.

[111] P. P. White, “RSVP and integrated services in the internet: a tutorial,” *IEEE Communications Magazine*, vol. 35, no. 5, pp. 100–106, May 1997.

[112] IETF, “Resource Reservation Protocol (RFC 2205),” September 1997.

[113] C. Gbaguidi, H. J. Einsiedler, P. Hurley, W. Almesberger, and J. Hubaux, “A Survey of Differentiated Services Proposals for the Internet,” April 1998.

[114] P. Goyal, H. M. Vin, and H. Cheng, “Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 5, pp. 690–704, 1997.

[115] A. Kortebi, L. Muscariello, S. Oueslati, and J. W. Roberts, “Evaluating the number of active flows in a scheduler realizing fair statistical bandwidth sharing,” *International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, vol. 33, no. 1, pp. 217–228, 2005.

[116] A. Kortebi, L. Muscariello, S. Oueslati, and R. J. W., “Minimizing the overhead in implementing flow-aware networking,” in *Architectures for Networking and Communications Systems*, October 2005.

[117] N. Benameur, S. Oueslati, and J. Roberts, “Experimental Implementation of Implicit Admission Control.”

[118] J. Park, M. Jung, S. Chang, S. Choi, M. Young Chung, and B. Jun Ahn, “Performance Evaluation of the Flow-Based Router Using Intel IXP2800 Network Processors,” in *International Conference on Computational Science and Its Applications (ICCSA)*, 2006.

[119] S. Oueslati and J. Roberts, “Method and a device for implicit differentiation of quality of service in a network,” April 2004, uS Patent App. 10/825,756.

[120] J. Roberts, S. Oueslati, and A. Kortebi, “Procede et dispositif d’ordonnancement de paquets pour leur routage dans un rseau avec dtermination implicite des paquets traiter en priorit,” 2006.

[121] ITU-T E.417, “Framework for the network management of IP-Based networks,” 2005.

[122] Anagran, “Intelligent Flow Routing for Economical Delivery of Next-Generation Network Services,” 2007.

[123] C. Cardenas, M. Gagnaire, V. Lopez, and J. Aracil, “Admission control for Grid services in IP networks,” in *IEEE First Symposium on Advanced Networks and Telecommunications Systems (ANTS)*, 2007.

[124] ——, “Performance Evaluation of the Flow-Aware Networking (FAN) architecture under Grid environment,” in *IEEE/IFIP Network Operations and Management Symposium (NOMS)*, 2008.

[125] C. Cardenas and M. Gagnaire, “Performance comparison of the Flow-Aware Networking (FAN) architectures under GridFTP traffic,” in *ACM/SIGAPP Symposium on Applied Computing (SAC)*, 2008.

[126] A. Gumaste, G. Kuper, and I. Chlamtac, “Optimizing light-trail assignment to WDM networks for dynamic IP centric traffic,” in *IEEE Workshop on Local and Metropolitan Area Networks (LANMAN)*, April 2004, pp. 113–118.

[127] Y. Ye, H. Woesner, R. Grasso, T. Chen, and I. Chlamtac, “Traffic grooming in light trail networks,” in *IEEE Global Telecommunications Conference (Globecom)*, 2005.

[128] F. P. Kelly, S. Zachary, and I. Ziedins, *Stochastic Networks: Theory and Applications.* Oxford University Press, USA, September 1996.

[129] V. Jacobson, “Congestion Avoidance and Control,” in *ACM SIGCOMM*, Stanford, CA, August 1988, pp. 314–329.

List of publications

All publications are ranked chronologically per topic.

Publications related to thesis

Books

1. Book chapter contribution: V. López, J. Aracil, *Traffic models for IP over WDM networks, "Enabling Optical Internet with Advanced Network Technologies"*, Ed. Springer, Series "Computer Communications and Networks". Editors: J. Aracil and F. Callegatti.

Journals

2. J. E. Gabeiras, V. López, J. Aracil, J. P. Fernández Palacios, C. García Argos, O. González de Dios, F.J. Jiménez Chico and J. A. Hernández: *Is Multi-layer Networking Feasible?*, in Optical Switching and Networking, accepted.

3. V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios and O. González de Dios: *A Bayesian decision theory approach for the techno-economic analysis of an all-optical router (extended version)*, in Computer Networks, July 2008, Vol. 52, Issue 10, pp. 1916-1926.

International conferences

4. V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios O. González de Dios: *Performance evaluation of a Bayesian decisor in a multi-hop IP over WDM network scenario*, in Optical Networking Design and Modeling (ONDM), Feb 2009.

5. V. López, C. Cárdenas, J. A. Hernández, J. Aracil and M. Gagnaire: Extension of the *Flow-Aware Networking (FAN) architecture to the IP over WDM environment*, in IEEE International Telecommunication NEtworking WorkShop on QoS in Multiservice IP Networks (IT-NEWS/QoS-IP 2008), February 2008. Within the top ten papers award.

6. V. López, J. A. Hernández, J. Aracil, J. P. Fernández Palacios O. González de Dios: *A Bayesian decision theory approach for the techno-economic analysis of an all-optical router*, in Optical Networking Design and Modeling (ONDM), May 2007. Published in Lecture Notes in Computer Science (LNCS). Within the top five papers award.

Other publications

Books

7. Chapter Editor: J. A. Hernández and V. López, *Optical Burst Switching, "Enabling Optical Internet with Advanced Network Technologies"*, Ed. Springer, Series "Computer Communications and Networks". Editors: J. Aracil and F. Callegatti.

Journals

8. J. A. Hernández, P. Reviriego, J. L. García-Dorado, V. López, D. Larrabeiti, J. Aracil: *Performance evaluation and design of Polymorphous OBS networks with guaranteed TDM services*, in IEEE/OSA Journal of Lightwave Technology, accepted.

9. J. A. Hernández, J. Aracil, V. López, J. L. García-Dorado and L. De Pedro: *Performance analysis of asynchronous best-effort traffic coexisting with TDM reservations in polymorphous OBS networks*, in Phot. Network Communications, August 2008.

10. C. Cárdenas, M. Gagnaire, V. López and J. Aracil: *Admission control in Flow-Aware Networking (FAN) architectures under GridFTP traffic, in Optical Switching and Networking*, January 2009, Vol. 6, Issue 1, pp. 20-28.

11. J. L. García-Dorado, J. E. López de Vergara, J. Aracil, V. López, J. A. Hernández, S. López-Buedo, L. de Pedro: *Utilidad de los flujos netFlow de RedIRIS para anlisis*

de una red acadmica (On the use of RedIRIS's netFlow information for academic networks), in RedIRIS Journal, April 2008, No. 82-83, ISSN 1139-207X.

12. J. A. Hernández, J. Aracil, V. López and J. E. López de Vergara: *On the analysis of burst-assembly delay in OBS networks and applications in delay-based service differetiation*, in Phot. Network Communications, August 2007, Vol. 14, No. 1, Pages. 49-62.

13. P. Chanclou, J. Fernández Palacios, S. Gosselin, V. López, E. Zouganeli: *Overview of the Optical Broadband Access Evolution - A joint paper of operators in the IST network of excellence e-Photon/ONe*, in IEEE Communications Magazine August 2006, Vol. 44, Issue. 8, Pages. 29-35.

International conferences

14. C. Cárdenas, M. Gagnaire, V. López and J. Aracil: *Performance Evaluation of the Flow-Aware Networking (FAN) architecture*, in IEEE-IFIP Network Operations and Management Symposium (NOMS 2008), April 2008.

15. V. López, J. L. García-Dorado, J. A. Hernández, J. Aracil: *Performance comparison of scheduling algorithms for IPTV traffic over Polymorphous OBS routers*, in International Conference on Transparent Optical Networks (ICTON) - Mediterranean Winter, December 2007.

16. C. Cárdenas, M. Gagnaire, V. López, and J. Aracil: *Admission control for Grid services in IP networks*, in Advanced Networks and Telecommunication Systems (ANTS), December 2007. Within the top seven papers award.

17. P. Aguilar-Jiménez, V. López, S. Sánchez, M. Prieto, D. Meziat: *Design and Implementation of Synthesizable SpaceWire Cores*, in International SpaceWire Conference, September 2007.

18. P. Chanclou, J. Fernández Palacios, S. Gosselin, V. López, E. Zouganeli: *Overview of the Optical Broadband Access Evolution - A joint paper of operators in the IST network of excellence e-Photon/ONe.* in The 2nd Institution of Engineering and Technology International Conference on Access Technologies, June 2006.

19. J. A. Hernández, J. Aracil, V. López, J. P. Fernández Palacios, O. González de Dios: *A resilience-based comparative study between Optical Burst Switching and Optical Circuit Switching technologies*, in IEEE ICTON, June 2006.

20. O. González de Dios, I. de Miguel, V. López: *Performance evaluation of TCP over OBS considering background traffic*, in ONDM, May 2006.

National conferences

21. O. González de Dios, I. de Miguel, V. López, R. Durán, N. Merayo, J. F. Lobo: *Estudio y simulación de TCP en redes de conmutación óptica de ráfagas (OBS) (Study and simulation of TCP in Optical Burst Switched (OBS) networks)*, in Telecom I+D, November 2005.