Universitat Politècnica de Catalunya

QoS Provisioning in Optical Packet Networks for Metropolitan and Wide Area Environments

Davide Careglio

Advisor: Josep Solé i Pareta

computer architecture department

A Thesis Presented to the Universitat Politècnica de Catalunya in Fulfilment of the Requirements for the Degree of

Doctor en Ingeniería de Telecomunicación

December 2004

Contents

List of Figures **v**

List of Tables **ix**

Abstract **xi**

Resumen **xv**

Structure of the thesis **xix**

1 Introduction **1**
1.1 Towards all-optical networks 1
1.2 Motivations and environments 5
1.2.1 Metro area networks 7
1.2.2 Wide area networks 7
1.3 Methodology and thesis outline 8

2 Optical packet switching: an overview **13**
2.1 Main concepts 13
2.2 Packet formats and network operations 14
2.3 Metro area network context 17
2.3.1 Examples of OPS-based metro network prototypes 21
2.4 Wide area network context 24
2.4.1 Node architectures 24
2.4.2 Techniques for contention resolution 26
2.4.3 Example of OPS-based backbone network prototypes 28

3 Introduction to the OPS-based metro area networks **33**

4 Multi-PON architecture **37**
4.1 State-of-the-art 37
4.1.1 MAC protocol 37
4.1.2 Scheduling algorithms 38
4.1.3 QoS provisioning 40
4.2 Contributions 41
4.2.1 Simulation scenario 41

4.3 Performance evaluation . . . 42
4.4 Optimization . . . 47
4.5 QoS provisioning . . . 47
4.5.1 Problem formulation . . . 47
4.5.2 QoS strategy: the Limited Attempts (LA) technique . . . 49
4.5.3 Performance evaluation . . . 49
4.5.4 Comparison with other QoS techniques . . . 53
4.6 Summary . . . 54

5 Multi-ring architecture **57**
5.1 State-of-the-art . . . 57
5.1.1 MAC protocol . . . 58
5.1.2 Scheduling algorithm . . . 59
5.1.3 Traffic measurements . . . 60
5.1.4 Fairness control . . . 61
5.1.5 QoS provisioning . . . 61
5.2 Contributions . . . 61
5.2.1 Simulation scenario . . . 61
5.3 Performance evaluation . . . 63
5.4 Optimization . . . 65
5.5 QoS provisioning . . . 66
5.5.1 Problem formulation . . . 66
5.5.2 Heuristic solution . . . 70
5.5.3 Performance evaluation . . . 71
5.5.4 Optimization of the QoS mechanism . . . 73
5.6 Summary . . . 82

6 Benchmarking **83**
6.1 Methodology and network scenario description . . . 83
6.2 Multi-PON versus multi-ring . . . 85
6.3 Multi-ring versus passive multi-ring, SDH, Ethernet, and RPR . . . 87
6.3.1 Benchmarked solutions . . . 87
6.3.2 Resource dimensioning . . . 87
6.4 Example of CAPEX analysis . . . 91
6.5 Example of OPEX analysis . . . 91
6.6 Conclusions and perspectives . . . 93

7 Introduction to the OPS-based wide area network **97**
7.1 State-of-the-art . . . 97
7.2 The connection-oriented OPS network . . . 99
7.3 Problems addressed in this thesis . . . 100
7.4 Simulation scenario . . . 103

8 The OVC setup in connection-oriented OPS networks 105
8.1 Problem description . . . 105
8.2 OWSA algorithms . . . 107
8.3 Performance evaluation . . . 108
8.4 Quality differentiation at the OVC setup . . . 111
8.5 Summary . . . 112

9 QoS management in connection-oriented OPS networks 115
9.1 State-of-the-art . . . 115
9.2 The Service Category-to-Algorithm Wavelength Selection technique . . . 117
9.2.1 Example of defining three different OPS service categories . . . 117
9.2.2 Performance evaluation . . . 121
9.2.3 Optical buffer architecture to integrate different SCAWS . . . 122

10 Conclusions and future works 129

A Acronyms 133

B Related publications 135
B.1 Papers . . . 135
B.2 Papers in revision . . . 137
B.3 Project deliverables . . . 137
B.4 Other publications . . . 138

Bibliography 139

List of Figures

1.1 Trend of the optical networking technology 2
1.2 Optical circuit switching solution 3
1.3 Optical burst switching solution 4
1.4 Optical packet switching solution 5
1.5 Network segments 6

2.1 An example of optical packet formats: (a) out-of-band control channel, (b) in-band control channel 14
2.2 (a) Data incoming from client layers can be placed in different optical packet formats: (b) synchronous, fixed-length packets; (c) asynchronous, fixed-length packets; (d) synchronous, variable-length packets; (e) asynchronous, variable-length packets 16
2.3 Schematic example of an OPS node for metro networks 18
2.4 Example of physical topologies for metro networks 19
2.5 Example of composite physical topologies for metro networks 19
2.6 Structure of the AWG-STAR metro network 22
2.7 Structure of the HORNET metro network 22
2.8 Structure of the DBORN metro network 23
2.9 A generic OPS node architecture 25
2.10 Schematic example of the (a) single-stage and (b) multi-stage node architecture 25
2.11 Schematic example of the (a) feed-forward and (b) feedback node architecture 25
2.12 Buffer configurations with 4 FDLs: (a) Degenerate $k_j = j - 1$, $\mathbf{Q}_4 = \{0, D, 2D, 3D\}$, (b) Non-degenerate $k_j = (j-1)^2$, $\mathbf{Q}_4 = \{0, D, 4D, 9D\}$ 27
2.13 Structure of the WASPNET switch node 29
2.14 Structure of the DAVID switch node 29

3.1 Metro network architectures considered in this thesis 34

4.1 Multi-PON architecture 38
4.2 Timing structure of the wavelength channels 39
4.3 Throughput as a function of the offered load under uniform traffic matrix 43
4.4 Throughput as a function of the offered load under diagonal traffic matrix 44
4.5 Throughput as a function of the offered load under dynamic diagonal traffic matrix 45

4.6 Throughput as a function of the offered load comparing the Greedy and the Frame-based algorithm 46
4.7 Average end-to-end delay as a function of the offered load comparing the Greedy and the Frame-based algorithm 46
4.8 Throughput as a function of the offered load comparing the original and the optimized solution 48
4.9 Maximum end-to-end delay as a function of the offered load comparing the original and the optimized solution 48
4.10 Throughput as a function of the offered load comparing TM and HM techniques and considering ($h = 3$, $k = 7$) 50
4.11 Throughput as a function of the offered load comparing different values of (h, k) and using the TM technique 51
4.12 Packet loss rate as a function of the offered load comparing different values of (h, k) and using the TM technique 52
4.13 Maximum end-to-end delay as a function of the offered load comparing different values of (h, k) and using the TM technique 52
4.14 Throughput as a function of HP traffic relative load percentage at 100% total load using the TM technique 54
4.15 Throughput as a function of the offered load comparing the AP, RED and LA 55
4.16 Maximum end-to-end delay as a function of the offered load comparing the AP, RED and LA 55

5.1 Architectures of (a) PMR node with transmission and reception decoupling and (b) MR node with erasure capability 58
5.2 Example of multi-slot forwarding in the multi-ring. Colors in slot represent packet destinations 59
5.3 Scheduling at the Hub 60
5.4 Throughput as a function of the offered load under uniform traffic matrix 63
5.5 Relative throughput per node for total load on the ring of 0.7, without fairness control (solid line) and with SAT for two values of Q 64
5.6 Throughput as a function of the offered load under diagonal-3 traffic matrix 65
5.7 Throughput as a function of the offered load under diagonal traffic matrix with spatial reuse (dashed line) and without spatial reuse (solid line). (a) Network with 16 rings and 4 wavelengths per ring, (b) Network with 4 rings and 16 wavelengths per ring 67
5.8 Throughput as a function of the offered load under diagonal-7 traffic matrix with traffic fluctuation comparing (a) the original and (b) the optimized solution 68
5.9 Throughput as a function of the offered load under uniform traffic matrix. GS load is fixed to 30% 71
5.10 Throughput as a function of GS traffic relative load assuming 100% total load under diagonal traffic matrix 72
5.11 Scheduling wavelength-to-wavelength permutations at the Hub 73

5.12 Example of multi-slot forwarding in the multi-ring network with wavelength-to-wavelength permutations 74
5.13 Throughput as a function of GS traffic relative load under the uniform traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration 78
5.14 Throughput as a function of GS traffic relative load under the diagonal traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration 79
5.15 Throughput as a function of GS traffic relative load under the power-of-ten traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration 80
5.16 Throughput as a function of GS traffic relative load under the very unbalanced traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration 81

6.1 The network dimensioning methodology 84
6.2 The considered switch architecture 86
6.3 Node structures. (DMUX: wavelength demultiplexer; MUX: wavelength multiplexer; NPR: network processing receiver; NPT: network processing transmitter; SW: STM-1/STM-4 switch; Eth SW: Ethernet switch; XC: cross-connect; DAB: data aggregation board. (a) Point-to-point Ethernet Hub + Node, (b) SDH node, and (c) RPR node. . 88
6.4 Node capacity (in Gbit/s) required in the different network architectures for the three traffic volumes 89
6.5 Example of OPEX analysis: relative annual OPEX cost comparison in the different network architecture for the three traffic volumes 92
6.6 Possible introduction scenario of the different metro technologies . . . 93

7.1 The considered switch architecture 98
7.2 Connection-oriented OPS network 100
7.3 Contention resolution techniques in connection-oriented OPS networks. 101

8.1 Example of OVC forwarding table configurations able to avoid and produce contentions. 106
8.2 Packet loss rate as a function of the relative load at different overall load. 107
8.3 OWSA algorithms. 109
8.4 Packet loss rate as a function of the offered load comparing the RND, RR, BLC, and GRP algorithms under uniform traffic matrix. 110
8.5 Packet loss rate as a function of the offered load comparing the RND, RR, BLC, and GRP algorithms under power-of-two traffic matrix. . . 110

8.6 Packet loss rate as a function of the offered load comparing the RND, RR, BLC, and GRP algorithms under unbalanced traffic matrix. . . . 111
8.7 Procedure for HQ and LQ OVC. 112
8.8 Packet loss rate as a function of the offered load. HQ load increases from 5% to 50% with respect to the overall load under uniform traffic matrix. 113
8.9 Packet loss rate as a function of the offered load with HQ load is 25% under (a) power-of-two traffic matrix and (b) unbalanced traffic matrix. 113

9.1 TSWS algorithms. 118
9.2 LBWS algorithms. 119
9.3 SKWS algorithms. 120
9.4 (a) Packet loss rate, (b) Forwarding opacity, and (c) Out-of-sequence packets as a function of D normalized to the average packet duration, comparing TSWS, LBWS and SKWS 121
9.5 Non-degenerate buffer configuration with 6 FDLs. BE packets can use delays $\{0, D, 2D, 3D\}$, while the RT and LS packets can use delays $\{0, 3D, 6D, 9D\}$ 123
9.6 Packet loss rate as a function of D normalized to the average packet duration. 124
9.7 Packet loss rate as function of the buffer length B. 125
9.8 Packet loss rate as function of traffic relative load percentage. 125

List of Tables

2.1 Example of node architectures 26

5.1 Average running times for the four solutions 72

6.1 Node types and traffic assumptions 84
6.2 Major components quantities for mean traffic (80G scenario) 85
6.3 Transport resources required in the different architectures 90
6.4 Example of CAPEX analysis: cost relative to the passive multi-ring . 91

9.1 PLR, FO and OS comparing SCAWS technique with EQWS and MIN-GAP both adopting a buffer threshold technique 124

Abstract

Over the past few years, Internet has grown at an exponential rate and bandwidth demands have been continuously increasing. At the beginning, the introduction of Wavelength Division Multiplexing (WDM) allowed to accommodate such a demand, without combine a replacement of the current redundant and inefficient multi-layer structure. The challenges have recently migrated from transmitting high-capacity optical signals over long distances to effectively switching and managing that data in optical domain. These functions, currently performed in the electronic domain, are the main causes of the large bottleneck to the scalability and growth of Internet.

On the other hand, it is expected that future networks should transport heterogeneous traffic services including multimedia and interactive applications, which require proper treatments (e.g., delay or bandwidth guarantees). QoS provisioning seems therefore a mandatory task.

Definitely, optical networks with QoS capabilities are the main goal for the next generation telecommunication infrastructures. An intense debate has been ongoing about which is the optical network model to adopt, aiming at identifying the degree of the optical transparency to be achieved, and the proper flexibility of the optical interconnections. In the perspective of network optimization, the implementation of packet switching techniques directly in the transport network will bring more statistical sharing of the physical resources to reduce the connection costs and cope with the gap between transmission speed and switching capacity. In this direction, two different approaches have been currently developing in the research community, namely Optical Burst Switching (OBS) and Optical Packet Switching (OPS), which are attractive solutions to the electronic bottleneck with promise of transparency, high capacity, flexibility, and little electromagnetic interference.

In this thesis we focus on the OPS solutions, which pretends to be a long-term solution requiring very fast all-optical switches (switching time in the order of nanoseconds), and advanced optical components such as tunable wavelength converters and 3R optical regenerators. Since OPS is based on statistical multiplexing, packet contentions may arise at the nodes. Therefore, a contention resolution policy must be applied to reduce the packets losses and make the statistical multiplexing more efficient. In the electrical packet switching, contention resolution techniques typically exploit the space domain, by means of deflection routing, and the time domain, by means of queuing. In the optical packet switching, the lack of optical RAMs imposes the use of a pool of fiber delay lines (FDLs), which are bulky and not scalable and offer limited buffering capabilities (few tens of delays at maximum). In contrast, the use of WDM links and wavelength converters allows to solving contention also in frequency domain, by means of wavelength multiplexing. This contention resolution technique is the most effective one in packet switching since it does not incur in additional delay and maintains both the correct packet order and the same hop distance to the destination.

In this context, we deal with the problem of providing QoS capabilities for both metropolitan and wide area environments.

For the metro environment, the networks are generally *buffer-less*, in the sense

that once information enters the network, it remains in the optical domain and does not face any buffers until it is delivered to its destination. In order to avoid collision on the individual WDM channels, Medium Access Control (MAC) protocols are needed, which may integrate the support of QoS provisioning.

Different network architectures can be envisaged from this simple concept. We focus on two particular architectures, namely multi-PON and multi-ring respectively. Both architectures with their respective functional mechanisms (including the MAC protocols) have been proposed by the DAVID consortium, which is a project funded by the Fifth Information Society Technologies program of the European Union. Our contributions on this topic address different problems. We firstly identify the service requirements for the generic OPS-based metro networks according to what developed in electrical, standardized MANs. Afterwards we perform an extensive and exhaustive performance evaluation to evaluate the mechanisms proposed in previous works and identify their weaknesses. We hence design optimized mechanisms to improve the performance and provide QoS support to all types of services previously identified. Finally, in order to assess the cost/effectiveness of the OPS-based solutions considered in this thesis, we compare them in a benchmarking study. This study also compares the results with non-OPS technologies based on electrical switching such as SDH, RPR and Ethernet.

For the wide area environment, we focus on the connection-oriented OPS network scenario where nodes have limited buffer capabilities and are subject to asynchronous variable length packets. In such a scenario, we address two problems, namely the problem of setting up of the Optical Virtual Circuits (OVCs), properly configuring the forwarding table at the nodes, and the problem of providing QoS.

Concerning the former problem, at the OVC setup, each node must assign both the output port and the output wavelength to the OVC in such a way that the packets belonging to that OVC are always switched to the same output. This double setup problem is different with respect to the *classical* RWA problem in circuit-switched network because here the wavelengths are shared among several OVCs (in a packet-switched basis). In this study we do not deal with the problem of selecting the output port, which depends on the routing protocol but we are interested in the election of the wavelength, which may be set locally by each node using a *OVC-to-wavelength setup assignment* (OWSA) algorithm. In particular we show that intelligent OWSA procedures can considerably improve the performance of the switches. The intelligence relies on grouping the flows coming from the same input wavelength allowing to obtain the conflict-free situations and hence reducing the contention probability.

Concerning the latter problem, existing solutions to provide QoS in OPS networks are based on the following strategy: 1) design a contention resolution algorithm, which minimizes the Packet Loss Rate (PLR), 2) apply a QoS mechanism (some form of resources reservation on top of the contention resolution algorithm) able to differentiate the PLR among two or more classes. Given that we are dealing with a connection-oriented model, here we suggest a new method based on the well known ATM scheme of defining different service categories, which consists of defining different OPS service categories, each one based on a different contention resolution algorithm specifically designed to cope with the requirements of that category. With this technique, besides

the PLR, also the packet delay and the computational complexity are considered as QoS metrics.

The work presented here is part of the research activities performed by the Broadband Communications Systems Group at the Advanced Broadband Communications Centre (CCABA) of the Universitat Politècnica de Catalunya. In particular, the work was carried out within three relevant research projects. Two projects funded by the European Union: the IST-1999-11742 DAVID (Data And Voice Integration over DWDM) project, and the COST Action 266 (Advanced Infrastructures for Photonic Infrastructures). The last one funded by the Spanish Ministry of Science and Technology (MCyT): TRIPODE under contract FEDER-TIC2002-04344-C02-02.

Finally, it has to be underlined that this work will continue to be developed and will be part of the new Integrated Project NOBEL (Next generation Optical network for Broadband European Leadership) founded by the European Union within the Sixth Framework Programme.

Resumen

El gran crecimiento y expansión de Internet en los últimos años, con el consecuente incremento de usuarios y tráfico, ha hecho que aumente la necesidad de ancho de banda en las redes de telecomunicación actuales. Esta demanda se debe en gran medida a la popularización de Internet y a la explosión de nuevos servicios que de ello se deriva y que exigen a la red mayor eficiencia, en términos de optimización de recursos y prestaciones. Es de esperar que en un futuro inmediato esta tendencia continúe, sobre todo a la vista del fuerte incremento del tráfico de datos frente al tráfico de voz.

Inicialmente, estas necesidades se han atendido integrando la tecnologa WDM (*Wavelenght Division Multiplexing*) a las actuales arquitecturas multi-capas sin la necesidad de reemplazar las funcionalidades redundantes e ineficaces de esta arquitectura. Por esta razón, el desafío de la futura generación de redes de telecomunicación punta a pasar de la simple transmisión de señales ópticas de gran capacidad a efectivamente conmutar y gestionar esta cantidad de datos en el domino óptico. Estas funcionalidades, actualmente realizadas por componentes eléctricos, son las que actualmente causan un cuello de botella en la escalabilidad y crecimiento de Internet.

Por otro lado, se espera que las futuras redes transporten servicios heterogéneos que incluyen tanto transferencia de datos como transmisión de aplicaciones multimedia e interactivas. Cada servicio por lo tanto necesita un requerimiento y tratamiento particular (por ejemplo garantir un limite en el retraso extremo-extremo o en el ancho de banda). En este entorno, proporcionar calidad de servicio (*Quality of Service*, QoS) resulta ser un factor obligatorio.

Definitivamente, las redes ópticas con soporte a la QoS son el principal objetivos de la próxima generación de redes de telecomunicación. La primera etapa de esta migración prevee pasar de los actuales sistemas punto a punto hacia interconexiones basadas en redes de conmutación de circuitos ópticos ASON. Esta solución es capaz de proporcionar conexiones ópticas bajo demanda de manera rápida y flexible a través de un plano de control basado en el paradigma GMPLS.

No obstante la introducción de la gestión automática de los recursos, ASON hace un uso relativamente estático de las longitudes de onda, no aprovechando todo el potencial que la tecnologa óptica puede ofrecer con una explotación más dinámica de los recursos. Con el objetivo de optimizar los recursos disponibles de manera automática, la implementación de la técnica de conmutación de paquetes directamente en las redes de transporte pretende ser una solución válida que aproveche la multiplexación estadística directamente en el dominio óptico para proporcionar mayor fiabilidad y alta velocidad y bajas interferencias electromagnéticas. En esta dirección, se están desarrollando dos nuevos enfoques llamados conmutación de ráfagas ópticas (*Optical Burst Switching*, OBS) y conmutacin de paquetes opticos (*Optical Packet Switching*, OPS). Ambas soluciones son igualmente atractivas y permiten eliminar el cuello de botella entre la transmisin óptica y la conmutacin.

Esta tesis se centra en redes OPS que pretende ser una solución a largo plazo en cuanto requiere conmutadores de alta velocidad y componentes ópticos avanzados como los conversores sintonizables de longitud de onda y regenadores completamente

ópticos. Siendo OPS una tecnología basada en la multiplexación estadística, dos o más paquetes pueden contender el mismo recurso lo que obliga la implementación de algoritmos de resolución de contenciones. En la conmutación de paquetes eléctricos, se usa tanto el dominio espacial, enviando unos paquetes hacia otros puertos de salidas, como el dominio temporal enviando los paquetes a las colas eléctricas. En OPS el uso del dominio temporal es muy limitado debido a la falta de memorias RAM ópticas que impone el uso de fibras de retrasos (*Fiber Delay Lines*, FDLs). Las FDLs son muy costosas, su uso muy engorroso, solo pueden proporcionar retrasos discretos y por lo tanto su uso debe limitarse a unas pocas decenas. Por otro lado, el uso de enlaces WDM y conversores de longitud de onda permite explotar el dominio de la frecuencia para resolver las contenciones.

En este ámbito el objetivo de esta tesis es el desarrollo de nuevos mecanismos para proporcionar QoS en redes OPS tanto en entorno metropolitano como de rea extendida.

Por lo que concierne el entorno metropolitano, las redes son generalmente *bufferless*, en el sentido que una vez transmitida la informacin a la red, esa se queda en el dominio óptico sin encontrar ninguna cola en el camino hasta alcanzar su destino. Para evitar contenciones, protocolos de acceso al medio compartido (*Medium Access Control*, MAC) son necesarios y pueden integrase con mecanismos para proporcionar QoS.

De este concepto se pueden diseñar varias arquitecturas distintas. En esta tesis nos concentramos en dos arquitecturas basadas en topologas compuestas, llamadas respectivamente redes multi-PON y redes multi-anillos. Ambas han sido desarrolladas en el proyecto de investigacin DAVID financiado por la Unión Europea dentro del quinto programa marco. Nuestras contribuciones abarcan varios aspectos. Antes de todos se han identificado los requerimientos de las futuras redes metropolitanas basadas en OPS, con particular atención en determinar los servicios necesarios de acuerdo con lo que se ha diseñado en la anterior generación eléctrica de redes metropolitanas. Luego para ambas redes se ha seguido el mismo procedimiento. La primera etapa ha sido la evaluación de prestaciones a través de simulaciones con el objetivo de identificar los puntos débiles. Se ha luego pasado a la fase de optimización tanto de la arquitectura de las redes como de los mecanismos que gobiernan su funcionamiento y se han verificado las mejoras. Finalmente se han propuesto mecanismos para proporcionar QoS según los requerimientos definidos anteriormente y se ha hecho un estudio de coste/prestaciones comparando las dos arquitecturas con otras actualmente en comercio como SDH, RPR y Ethernet.

Por lo que concierne el entorno de área extendida, se ha considerado una red de conmutación de paquetes ópticos orientado a la conexión donde los nodos tienen limitadas capacidad de encolamiento. En este contexto, se han tratado dos problemáticas: el establecimiento de las conexiones virtuales ópticas (*Optical Virtual Circuit*, OVC) configurando propiamente las tablas de expedición (*forwarding table*) en los nodos y la provisión de QoS.

Para el primer punto, a la llegada de una petición de establecimiento de una nueva OVC, cada nodo debe asignar un puerto y una longitud de onda de salida a esta OVC. En el entorno OPS, esta tarea es diferente respecto al *clásico* problema

RWA (*Routing and Wavelength Assignment*) presente en las redes de conmutación de circuitos en cuanto en este caso las longitudes de onda están compartidas entre varios OVCs. Mientras la elección del puerto de salida depende de los algoritmos de routing, la elección de la longitud de onda se puede decidir localmente en cada nodo según diferente políticas llamadas *OVC-to-wavelength setup assignment* (OWSA). En esta parte de la tesis se ha estudiado en detalle este problema y se han propuesto diferentes estrategias. En particular se ha demostrado que una buena política de asignación de longitud de onda incrementa notablemente la prestación de un conmutador OPS. Se ha usado la idea de agrupar, siempre que se pueda, los flujos de tráfico que entran en el conmutador por el mismo puerto y misma longitud de onda de manera de disminuir lo máximo posible la probabilidad de contención entre paquetes.

Por lo que concierne proporcionar QoS, el estado del arte indica que hasta el momento se ha siempre seguido la misma técnica basada en: 1) diseñar un algoritmo de resolución de contenciones que minimice la probabilidad de pérdidas de paquetes (*Packet Loss Rate*, PLR) y luego 2) aplicar un mecanismo de reserva de recursos capaz de diferenciar la PLR entre dos o mas clases de tráfico. Considerando que el entorno de estudio es orientado a la conexión, se ha propuesto un enfoque diferente basado en el esquema aplicado en redes ATM donde se definen diferentes categorías de servicio, cada una con su propio tratamiento dentro de la red. En particular se han definido 3 categorías de servicio para entorno OPS y se han desarrollado 3 algoritmos de resolución de contenciones, cada uno pensado para proporcionar el servicio requerido. Con esta técnica, además de controlar la PLR, también se pueden considerar el retraso y la complejidad computacional como métricas QoS.

El trabajo presentado en esta tesis es parte de las actividades de investigación del grupo de Sistemes de Comunicació de Banda Ampla del Centre de Comunicacions Avançades de Banda Ampla (CCABA) de la Universitat Politècnica de Catalunya. En particular, el trabajo ha hecho parte de tres proyectos de investigación. Dos de ellos financiado por la Unión Europea: el proyecto IST-1999-11742 DAVID (Data And Voice Integration over DWDM) y la acción COST 266 (Advanced Broadband for Photonic Infrastructure). El último financiado por el Ministerio de Ciencia y Tecnología (MCyT): el proyecto TRIPODE (FEDER-TIC2002-04344-C02-02).

Finalmente, cabe subrayar que este trabajo se seguirá desarrollado dentro de un nuevo proyecto de investigación financiado por la Unión Europea: el proyecto integrado FP6-506760 NOBEL (Next generation Optical network for Broadband European Leadership).

Structure of the thesis

Environment

Optical Packet Switching: cope with the electronic bottleneck exploiting the statistical multiplexing directly in optical domain; the optical packets travel along the network transparently while the headers are converted in electrical domain at each hop to take forwarding decisions.

In this dissertation

Quality of Service is defined by IETF (RFC 2386) as *"the complete set of service requirements to be met by the network while transporting a flow"*. This dissertation mainly focuses on providing QoS in both wide area and metropolitan environment using OPS environment assigning/managing/controling resources such that not all customers are treated the same.

OPS-based metropolitan area networks

Scenario: two composite network topologies

- Multi-PON topology
- Multi-ring topology

Tasks

- Identification of the service requirements
- Optimization of the architecture and the proposed MAC protocol
- QoS supporting all types of services
- Benchmarking study

Identification of the service requirements

- Guaranteed service: data with real-time constraints.
- Priority service: data near-real-time, less delay and bandwidth-sensitive.
- Best-effort service: data that can be sent at the leisure of the node.

Multi-PON	*Multi-ring*
Related work • MAC protocol • QoS supporting Best-effort and Guaranteed services	Related work • MAC protocol • QoS supporting Best-effort and Priority services
Contributions • Optimization phase • Support of the Priority service	Contributions • Optimization phase • Support of the Guaranteed service

Benchmarking study

A cost-effectiveness study is performed to assess the benefits of the multi-PON and multi-ring solutions comparing them to non-OPS technologies such as SONET, Ethernet and RPR.

OPS-based wide area networks

Scenario: single connection-oriented OPS switch

Tasks

- Analysis of the related work
- Identification of the critical metrics
- Novel techniques for QoS support in connection-oriented OPS networks

Analysis of the related work

- Several contention resolution algorithms ranging from simple random to complex void filling queueing selection
- QoS provisioning based on the following method:
  - design a contention resolution algorithm which minimizes the packet loss rate
  - then apply a QoS mechanism to differentiate the packet loss rate using resource reservation, offset time or hybrid electrical/optical buffers

Identification of the critical metrics

- Packet loss rate
- Out-of-sequence delivery of packets
- Computational complexity

Novel techniques

- Contributions to the development of the connection-oriented OPS network
- Design of original policies to setup optical virtual connections
- QoS management based on defining different categories of service

Chapter 1

Introduction

1.1 Towards all-optical networks

The existing transport infrastructures are based on technologies foreseen during the '90s when the focus was on providing a guaranteed level of performance and reliability for voice calls and leased lines. But the trend in networking of the last decade reveals that the Internet and IP[1] are becoming the dominant solution, reaching world-wide diffusion and acceptance. Born as a research and university network, providing basic services like e-mail and file transfer, Internet has been growing at an exponential rate. The last network generation consequently deployed solutions capable of fulfilling the capacity requirements of present and future data traffic in both wide and metro environments. The optical technology and in particular the Wavelength Division Multiplexing (WDM) systems are been deeply introduced to increase the bandwidth-carrying capacity of a single optical fiber by effectively creating multiple virtual fibers, each carrying multigigabits of traffic per second, on a single fiber. This network evolution leads to today's network infrastructure which typically comprise four layers: IP for carrying applications and services, ATM for an efficient path management of the network, SONET/SDH for a robust transportation, and WDM for a wide capacity.

The economical crisis during 2001 has shown the drawbacks of current four-layer Internet infrastructure which limit its scalability in terms of dimension, services and business. Indeed multilayer architectures suffer from the lowest common denominator effect where any one layer can limit the entire network, as well as add to the cost of the entire network. As a result the existing transport infrastructure has been currently overcome by a new network model based on IP over WDM (i.e., IP traffic transported directly over optical networks) [52] [78]. This means that the intermediate layer functionalities (essential for a proper network behavior) must move to the other layers. In the end, this results in a simpler, more cost-efficient network that will transport wide range of data streams and very large volume of traffic.

On the other hand, the purely connectionless and *best effort* IP paradigm seems however insufficient to fulfil the needs of new multimedia and interactive applications. Works are ongoing to evolve Internet by optimizing the network efficiency with the

[1] see Appendix A for the list of acronyms

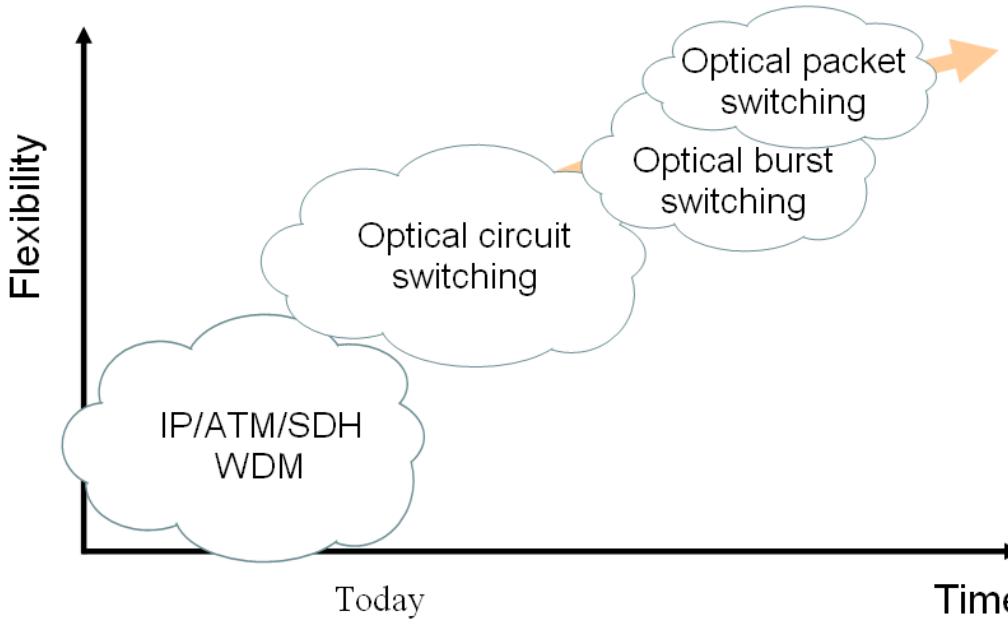

Figure 1.1: Trend of the optical networking technology

integration of Traffic Engineering (TE) [92] functionalities and providing effective Quality of Service (QoS) [89] capabilities. In this direction, proposals such as the Multi-Protocol Label Switching (MPLS) standard [91] and the Differentiated Services [90] architecture can provide the tools for effective QoS management and TE in the Internet.

Definitely, developing optical networks with QoS and TE capabilities is the main goal for next generation telecommunication networks. In this perspective an intense debate has been ongoing about which is the optical network model to adopt, aiming at identifying the degree of optical transparency to be achieved, and the proper flexibility of optical interconnections. Figure 1.1 foresees such trend showing the possible steps from today point-to-point transmission towards more flexible network implementations. We can recognize the current network architecture in the first bubble on the bottom-left side of the Figure 1.1 (i.e., IP/ATM/SDH/WDM).

The first step in the future is the introduction of some flexible mechanisms in WDM networks. The GMPLS technology provides an intelligent control plane (signaling and routing) functionalities for devices that switch in any of these domains: packet, time, wavelength, and fiber with minimum human interaction. This common control plane promises to simplify network operation and increase the network utilization. The basic challenge for an all-encompassing control protocol is the establishment, maintenance, and management of traffic-engineered paths to allow the data plane to efficiently transport user data from the source to the destination [3] [4] [44]. As shown in Figure 1.2, the traffic-engineered paths are physical circuits called *lightpaths* established between two end-points. The network nodes, called *optical cross-connects* (OXCs), are basically space switches which connect input and output ports on a per-wavelength basis, with or without wavelength conversion. When a

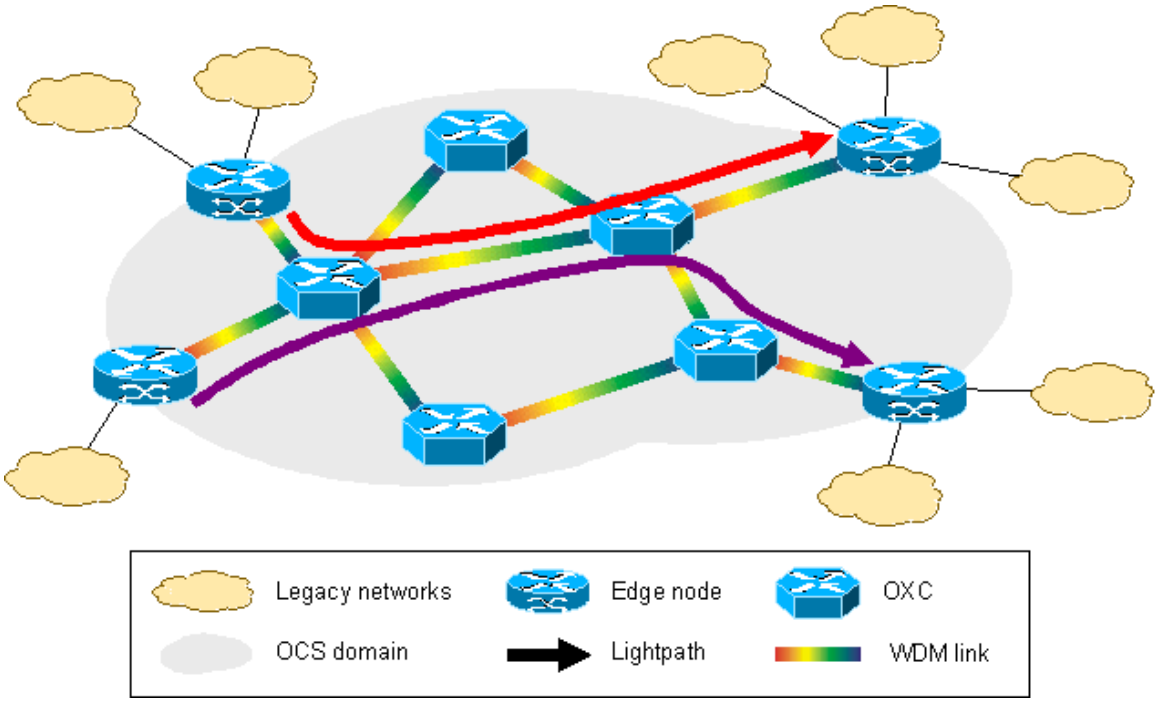

Figure 1.2: Optical circuit switching solution

node requires to establish a new lightpath for a given traffic flow, it sends a signalling request along a route. If the request can be completely accommodated, each OXC reserves the resources configuring its switch matrix. Resulting optical circuit switched (OCS) networks can offer explicit transfer guarantees and some degree of flexibility.

Toward full network utilization, the OCS networks present some drawbacks. They need considerable delay to confirm circuit establishment due to the both propagation delay and optical cross-connects reconfiguration. The network consists of a complex, heterogeneous structure with no easy manageability for protection/restoration, TE, and scalability/upgradability issues. These problems redirect the research community toward solutions able to offer high utilization, probably cope with high traffic churn and a significant portion of bursty traffic [38], deliver connection-oriented services and need to be cost effective too. Handling finer granularity connections appears therefore the aim of next generation networks where packet-based networks, such as Internet, will play a predominant role.

At this point, it is important to consider the emerging all-optical devices. In spite of the extraordinary advances in transmission capacity, optics has not penetrated much into the switching and management part of the network and all-optical networks are still in their infancy. New optical components such as Tunable Wavelength Converters (TWCs), Semiconductor Optical Amplifiers (SOAs), 3R optical regenerators are currently under development aiming at providing very high integration degree and very low power consumption [87] [33].

These new devices open up new possibilities where a finer granularity with respect to the OCS solution can be realized with the Optical Burst Switching (OBS) and

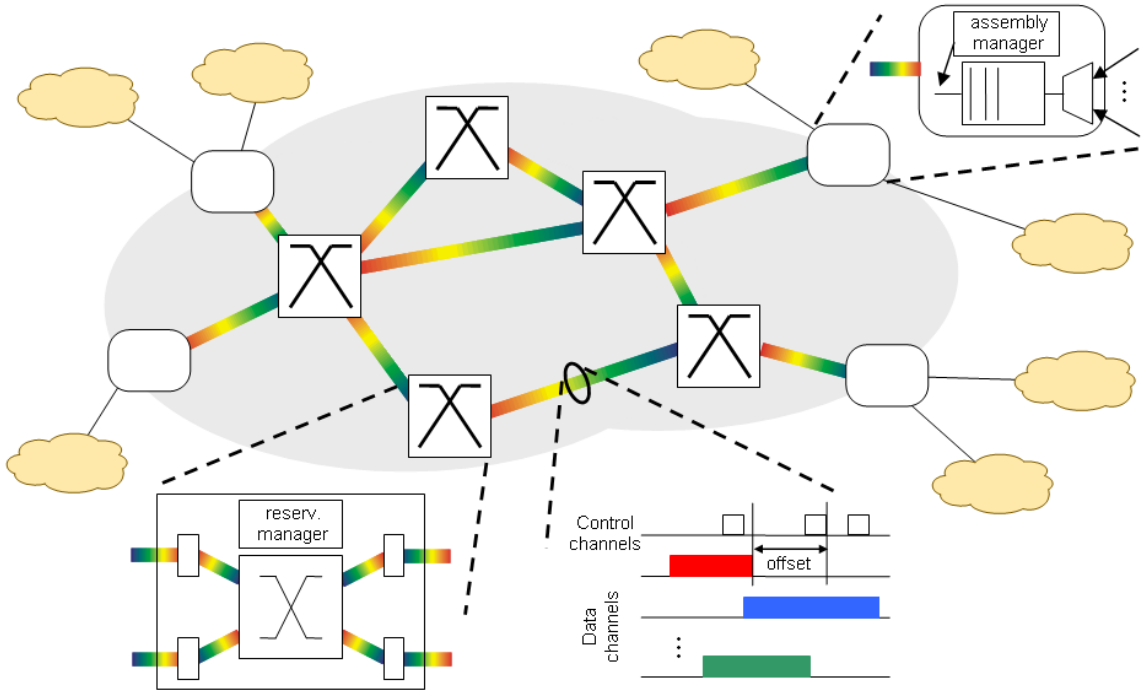

Figure 1.3: Optical burst switching solution

Optical Packet Switching techniques. Both switching technologies do not require a two-way end-to-end signalling, as data is optically transferred across the network without waiting to set-up the entire path. Instead of over-provisioning of circuits, these solutions apply the packet switching techniques directly in the optical transport network bringing more statistical sharing of the physical resources to reduce the connection costs.

OBS networks [86] (see Figure. 1.3) are characterized by a separation of data and control channels. At first, a control packet is sent to support intermediate nodes configuration and resource reservation; meanwhile the source node builds the corresponding burst aggregating incoming packets with the same characteristics (destination, QoS level, etc.); when ready, the burst is sent and optically switched across the network. This means that only control packets are electro-optical converted at each hop to take reservation decisions, while the bursts always remain in the optical domain.

OPS networks [62] (see Figure 1.4) use finest switching granularity and require very fast all-optical switches (switching time in the order of nanoseconds). Control and data information travels together in the same channel; each intermediate node converts the control headers to take switching decisions while the packets always remain in the optical domain.

Both solutions improve wavelength utilization by introducing statistical multiplexing directly in the optical domain. This cause that data contentions may arise at the nodes and therefore contention resolution policies must be applied to reduce the losses probability and make the statistical sharing more efficient. In electrical so-

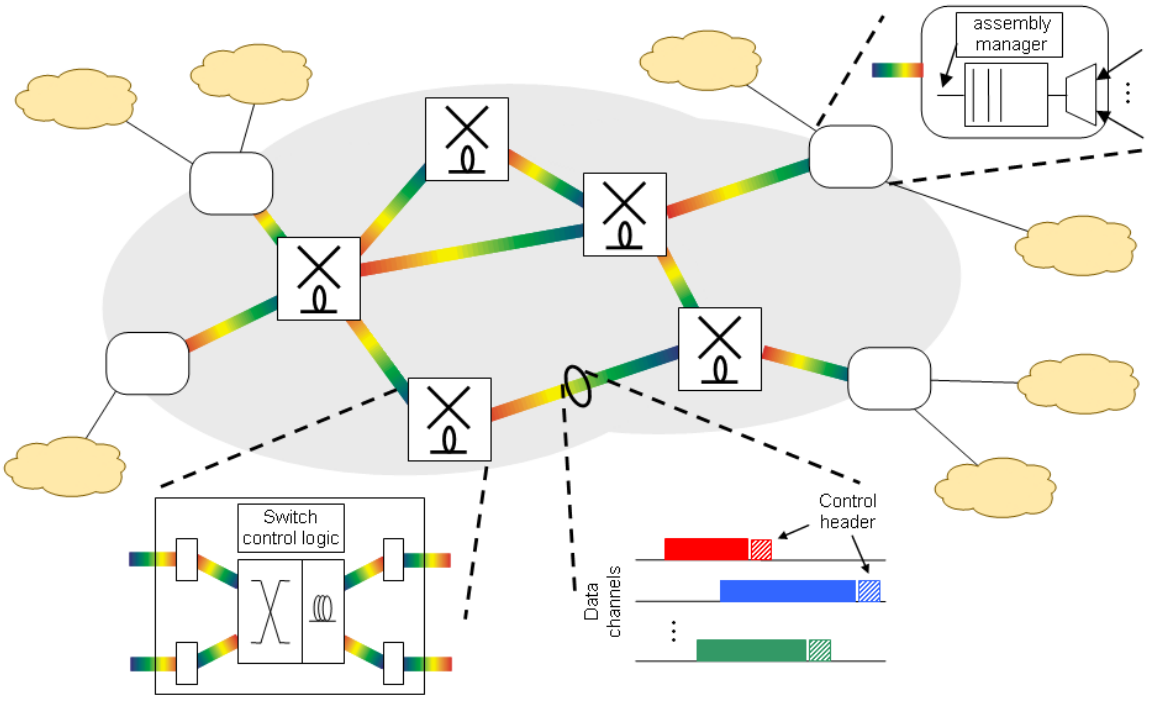

Figure 1.4: Optical packet switching solution

lutions, contention resolution may usually performed in the space domain, by means of deflection routing, and in the time domain, by means of queuing. In optics, the lack of optical RAMs imposes the use of a pool of Fiber Delay Lines (FDLs) [48] which are bulky and not scalable and offer limited buffering capabilities (few tens of delays at maximum [45]). Nonetheless, optics can exploit the WDM links and wavelength converters and therefore solve contention also in frequency domain, by means of wavelength multiplexing [105].

In this thesis, we focus on the OPS networks (i.e., the final step of the current networking evolution), although the same concepts developed here can be easily modified to be effectively applied to OBS networks. This adaptation is an open issue and will be part of future works.

1.2 Motivations and environments

The technology limitation of the optical queuing motivates significant research efforts in recent years dealing with the design of simple scheduling policies able to provide QoS differentiation. The impossibility of pre-emptying packets already buffered makes unfeasible the implementation of conventional fair queuing scheduling commonly used in present day routers. At the same time, QoS schemes must be kept very simple to be effective in OPS where each node must be able to schedule tens of Terabits per second.

Even though numerous works have been done in the past on optical networks,

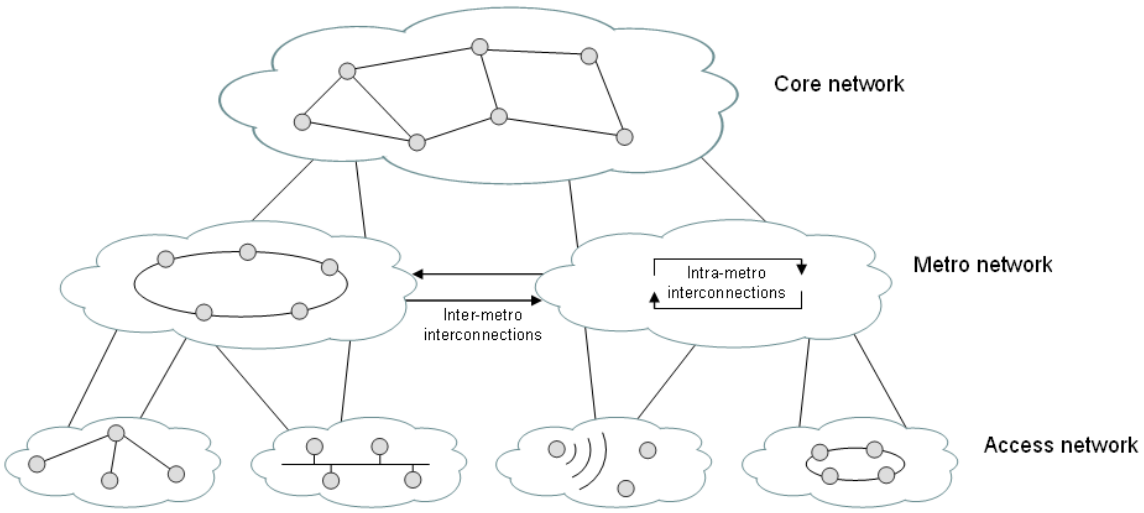

Figure 1.5: Network segments

few examples try to address the problem of a whole network performance analysis and design. At the same time, new emerging optical technologies are continuously opening up new opportunities for the development of novel network architectures and operative mechanisms. In this dissertation we hence deal with OPS networks placing particular attention to the QoS provisioning problem from an end-to-end perspective.

The hierarchy of communication networks can be viewed as consisting of three major segments (Figure 1.5): the access network, metro network, and wide area network[2]. The access network is responsible for collecting end-user traffic and is usually of less than tens of kilometers on its extension. Example of access networks include local/regional Internet Service Providers (ISPs), intra-building Ethernet, etc. The metro network is responsible for aggregating traffic from several access networks and routes it onto another metro (inter-metro traffic), onto the wide or directly delivers the traffic to its destination within the same metro (intra-metro traffic). The metro usually covers distances from tens of kilometers to few hundreds of kilometers. Finally, the wide area network (also known as core or backbone network[3]) interconnects metro networks and are typically more than hundreds of kilometers away.

In this thesis, we focus on both the metro and wide environment. From the networking point of view, the requirements and therefore the design guidelines of the former are generally different than the latter. We therefore carry out separated studies on metropolitan and wide area networks which are briefly described in the following subsections.

[2]Metro and wide are usually referred using the acronym MAN and WAN, respectively

[3]From now on, wide, backbone and core will be indiferently used to referring to this area network

1.2.1 Metro area networks

In the metro environment, the networks are generally *buffer-less*, in the sense that once information enters the network, it remains in the optical domain and does not face any buffers until it is delivered to its destination. Different network architectures can be envisaged from this simple concept (see Chapter 2 for a brief survey). In order to avoid collision on the individual WDM channels, Medium Access Control (MAC) protocols are needed which may integrate the support of QoS provisioning.

In this part of the thesis, we focus on two network architectures proposed in the DAVID project [40], namely multi-PON and multi-ring respectively. DAVID (Data And Voice Integration over DWDM) is a project funded by the Information Society Technologies program of the European Commission with the aim of providing advanced optical DWDM packet-switched network by developing innovative concepts and technologies [43]. The multi-PON and multi-ring architectures with their respective functional mechanisms (including the MAC protocols) have been proposed by other partners of the DAVID consortium.

Our task on this environment include: (i) the performance evaluation of these architectures, (ii) the identification of the drawbacks and of the open issues, (iii) the optimization of the proposed architectures and MAC protocols, and finally (iv) the proposal of QoS mechanisms to support multi-class traffic. The suggested solutions are also compared with new electrical solution targeting the same application environments (such as Resilient Packet Ring [96]) in a cost/performance perspective.

1.2.2 Wide area networks

In wide environment, queueing is achieved by Fibre Delay Lines (FDLs). As previously discussed, these buffers have many drawbacks: they are cumbersome, costly, and do not allow a new incoming packet to overcome other packets already queued, etc. Since the exploitation of the time is limited, frequency domain is deeply exploited to reduce the queueing requirements and, at the same time, to increase the network performance.

Recent works (for instance [16] [21]) suggest the integration of a connection-oriented path management protocol on top of the contention resolution algorithm which both improves the network performance and reduces the control complexity. In this context, protocols such as MPLS [91] can be extended to the OPS environment. These connection-oriented protocols are based on a distributed management scheme that provides and maintains optical paths, which we call Optical Virtual Circuits (OVCs).

In this part of the thesis we focus on the connection-oriented OPS network scenario and we address two problems, namely the problem of setting up of the OVCs, properly configuring the forwarding table at the nodes, and the problem of providing QoS.

Concerning the former problem, at the OVC setup, each node must assign both the output port and the output wavelength to the OVC in such a way that the packets belonging to that OVC are always switched to the same output. This double setup problem is different with respect to the *classical* RWA problem in circuit-switched

network because here the wavelengths are shared among several OVCs (in a packet-switched basis). In this study we do not deal with the problem of selecting the output port which depends on the routing protocol but we are interested in the election of the wavelength which may be set locally by each node using a *OVC-to-wavelength setup assignment* (OWSA) algorithm. In particular we show that intelligent OWSA procedures can considerably improve the performance of the switches. The intelligence relies on grouping the flows coming from the same input wavelength which allows to obtain the conflict-free situations and hence reduce the contention probability.

Concerning the latter problem, existing solutions to provide QoS in OPS networks are based on the following strategy: 1) design a contention resolution algorithm which minimizes the Packet Loss Rate (PLR), 2) apply a QoS mechanism (some form of resources reservation on top of the contention resolution algorithm) able to differentiate the PLR among two or more classes. Given that we are dealing with a connection-oriented model, here we suggest a new method based on the well known ATM scheme of defining different service categories which consists of defining different OPS service categories, each one based on a different contention resolution algorithm specifically designed to cope with the requirements of that category. With this technique, besides the PLR, also the packet delay and the computational complexity can be considered as QoS metrics.

1.3 Methodology and thesis outline

In order to attain the objective of the thesis, we use the following methodology for both metropolitan and wide area network environments:

1. Collect, analyse and identify the drawbacks of the previous works found in the literature on the same subject;

2. Propose novel or optimised solutions to overcome the drawbacks found in the previous step;

3. Set up ad hoc network simulators and simulation scenarios, thus carry out proper simulation evaluations for evaluating the merits of the proposals.

4. If the results are not as good as expected, repeat step 2.

The thesis is organized in four parts.

The first part includes Chapter 2 where we presents an overview of the current state-of-the-art of the optical packet switching paradigm. The application of the OPS concepts to both metropolitan and wide area network environments is also addressed.

The second part is dedicated to the OPS-based metro area networks. Chapter 3 introduces the two OPS-based metro area architectures studied in this thesis, namely the multi-PON and multi-ring networks. Here, we also identify the required services in such environments. In Chapter 4 we discuss the state-of-the-art of the multi-PON architecture and describe our contributions on this subject. The same structure is

used in Chapter 5 where we introduce the related work of the multi-ring architecture followed by the description of our contributions on this network. In Chapter 6 we compare the multi-ring and multi-PON solutions in a benchmarking study. We do not restrict ourselves to detailing the multi-PON and multi-ring performance, but also compare them to non-OPS technologies such as SDH, Ethernet and Resilient Packet Ring (RPR).

The third part is dedicated to the OPS-based wide area network environment. In Chapter 7 we briefly introduce the OPS backbone environment introducing the connection-oriented environment and the problems addressed in this thesis on this topic. In Chapter 8 we deal with the problem of setting up the optical virtual connections, properly configuring the forwarding table at the nodes. In this context, we study setup procedures to improve the switch performance compared to simple random schemes. In Chapter 9 we address the QoS provisioning problem proposing a method based on the well known ATM scheme of defining different service categories. In particular, we define three different OPS service categories based on three different contention resolution algorithms

The fourth and last part includes Chapter 10 where we draw the conclusions and indicate the future works.

PART I

Optical packet switching

Chapter 2

Optical packet switching: an overview

In the following chapter we describe the common concepts that underlying any OPS proposals both in wide and metropolitan scenario (section 2.1). Afterwards, we separate the environments, describing first the state-of-the-art of OPS solutions for metro networks (section 2.3) and then for core networks (section 2.4).

2.1 Main concepts

The implementation of packet switching techniques in the optical domain is a research topic that has been investigated all over the last decade [62] [105]. Several research projects demonstrate the feasibility of the OPS technology and significant progress in terms of component availability and integration has been recently achieved [33].

The target is building transparent optical network equipments capable to carry data-centric traffic at huge bit-rates. Since electronic-based devices may result too slow to perform the required ultra-fast switching operation, the basic idea is to exploit the bandwidth made available by optical components while reducing the electro-optical conversions as much as possible and achieving better interfacing with WDM transmission systems. An example of OPS network with mesh topology is shown in Figure 1.4. Legacy packet-oriented networks (e.g. based on IP, ATM or Gigabit Ethernet) are "clients" of the OPS domain and supply heterogeneous datagrams/cells/frames to it through the ingress *edge systems*, which are responsible for building optical packets. In particular, each edge node has to collect incoming data and arrange them into optical packets, according to the specific format adopted by the network. While performing this operation, it may be required to aggregate small incoming data units or segment long ones in order to fit them properly into the optical container. The edge node is also in charge for creating a packet header and adding control information to it, needed to accomplish a correct path inside the OPS network. Once a packet has entered the OPS domain, it is transparently switched by the nodes according to a statistical multiplexing techniques. As soon as the packet has reached the proper egress edge node, its data content is translated back to the original format

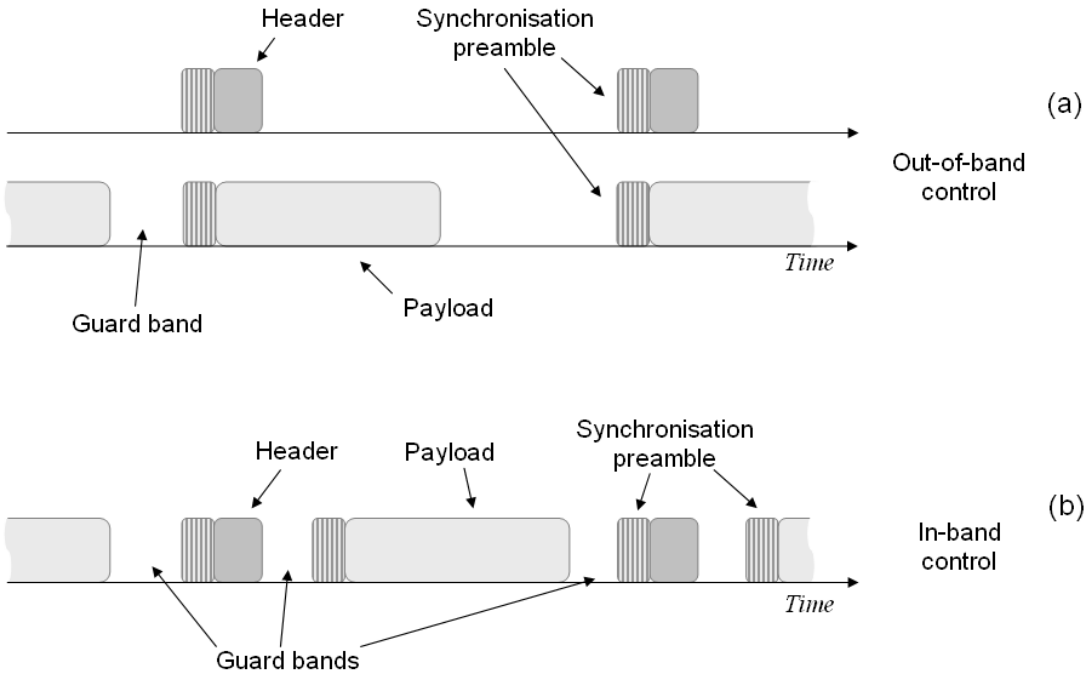

Figure 2.1: An example of optical packet formats: (a) out-of-band control channel, (b) in-band control channel

and delivered to the destination legacy network. Here some reassembling operations may be needed.

2.2 Packet formats and network operations

One of the key issues in OPS is the format of optical packets, that should be carefully chosen taking into account the limits of the optical technology on one side and traffic characteristics as well as transparency requirements on the other. The optical packets are typically composed by header and payload. The header contents are used to control the routing/switching decisions of the optical packet. Different techniques to attach the header to a packet can be used:

- out-of-band control, where the headers are transmitted on a separate wavelength, i.e., the control channel (e.g., [9]);
- in-band control, i.e., serial transmission of header and payload on the same wavelength (e.g., [50]); in this case, header and packet can be transmitted using orthogonal modulation formats to increase the channel utilisation (e.g., [69]).

Figure 2.1 show an example of the different optical packet formats: (a) refers to the case of out-of-band control while (b) to the in-band transmission case.

The figure shows that a suitable idle time intervals, called *guard bands* have to be introduced between header and payload and between contiguous packets accounting

for switching times of the constituent opto-electronic devices as well as payload position jitter [50]. In addition, *synchronization preambles* are placed in front of each header and payload, allowing the receivers to correctly lock on the optical signal.

The out-of-band control transmission is generally used when packets and headers travel together on the same route and have a locked timing relationship. This is especially a viable approach in metro networks where simple ring or star topologies are adopted, as the synchronization between the control and data channels is reasonably easy to maintain. On the other hand, it results complex when applied in a core network which generally consists of a meshed topology.

The next step in the definition of the optical packet format concerns the choice of the size of the entities to be switched as well as the network operation techniques to be adopted. The following alternatives have been proposed in literature, we use Figure 2.2 to compare them supposing that two client packets (Figure 2.2(a) entering in the OPS domain and the in-band control transmission is used for any alternatives:

- **Fixed length packets and synchronous node operation**, FLP-SO [50] (see Figure 2.2(b). The time scale is partitioned into time-slots of fixed duration and switching and transmission functions are performed only at given instances, i.e. at the beginning of each slot. Each payload has a fixed size and is inserted in a time-slot; a header is added into each slot and resulting fixed-length packets are switched independently one-another. Here, some padding to fill-up the last slot may be necessary.

- **Variable length packets and synchronous node operation**, VLP-SO [17] (see Figure 2.2(c). In this case the datagram transmission time is larger than the slot size, and a single header is inserted in the first slot while the payload spans over several time-slots, which are switched altogether as a single "train of slots". Some padding operation to fill-up the last slot may be required.

- **Fixed length packets and asynchronous operation**, FLP-AO [55](see Figure 2.2(d). Incoming datagrams are put into one or more optical packets of a given size, which may be received, switched and transmitted at any time. Insertion of some padding to fill-up the optical payload may be necessary. This case has little interest and will not be considered in the following discussions.

- **Variable length packets and asynchronous node operation**,VLP-AO [99] (see Figure 2.2(e). Incoming datagrams are put into the optical payloads "as they are" and each packet may be received, switched and transmitted at any time.

From the performance point of view, using fixed length packets with synchronous operation is the better choice. Unfortunately, this case is not very well tailored by Internet traffic. Indeed, the IP packets are variable length datagrams and must be fitted into fixed length containers. This may cause severe inefficiencies while at the same time the synchronous operation implies that the input interface of the switches needs to delineate and align the packets arriving from the different input ports. This

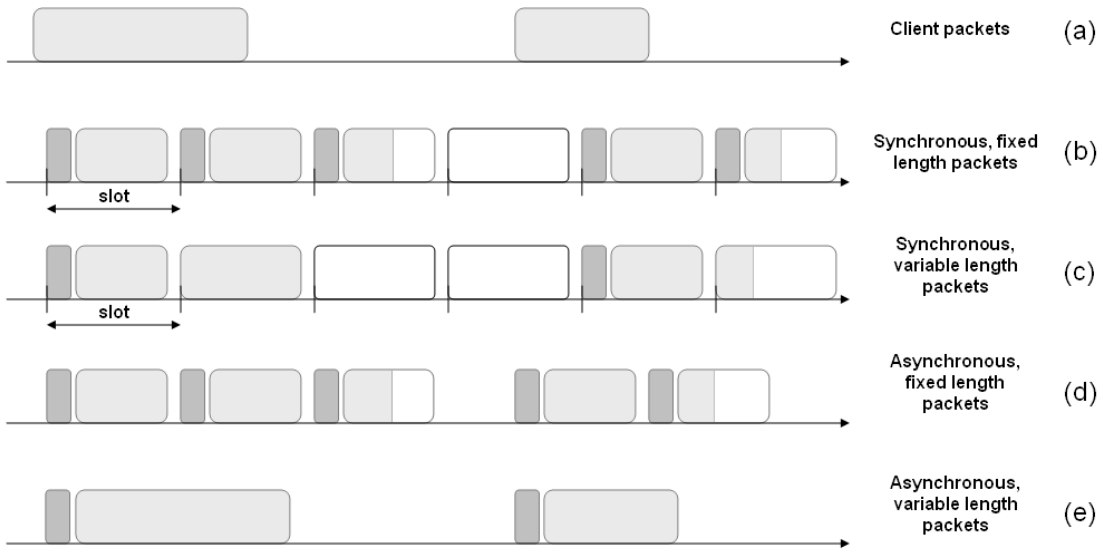

Figure 2.2: (a) Data incoming from client layers can be placed in different optical packet formats: (b) synchronous, fixed-length packets; (c) asynchronous, fixed-length packets; (d) synchronous, variable-length packets; (e) asynchronous, variable-length packets

operation is not trivial and is one of the most arduous, complex and cost tasks [99] for the OPS networks. Thereby, the asynchronous networks generally have lower cost, higher flexibility, robustness and offer better interworking skill with network protocols; on the contrary they have lower overall throughput than synchronous networks, because of the increased contention probability.

For these reasons, current investigations mainly consider the asynchronous, variable length case for the wide networks where generally the available bandwidth is not the principal design constraint. At the same time, asynchronous, variable length case well suits with the in-band header transmission; hence we can assume a scenario with full flexibility where packets and corresponding headers travel together across a meshed network using any available route with whatever wavelength from source to destination node.

On the contrary, the synchronous, fixed-length case is the preferred solution for the metro networks where limiting the cost per transmitted bit is generally the main goal to achieve [30]. At the same time, the utilization of the wavelengths can be further improved using the out-of-band header transmission, i.e., putting the control information on an additional wavelength. In this case we can assume a scenario where packets travel on the data wavelengths while the corresponding headers all together on a single control wavelength; a locked time between packets and headers permits to recognize the correct relationships; while a regular topology fixes the route from the source and destination nodes and does not affect this correspondences [43].

2.3 Metro area network context

Whereas core and access networks are currently experiencing huge innovations, the metro networks are mostly SONET/SDH over WDM rings that carry the increasing amount of data traffic very inefficiently. This results in the so-called *metro-gap*. The gap creates a clear network bottleneck preventing the client benefits. Therefore researchers world-wide make big efforts to investigate new packet-based technologies (e.g., RPR [93] [95]). They natural fit with the now ubiquitous IP traffic, and appear to be the best choices for overtaking the metro-gap in a cost-effective manner.

In the current networking context, a MAN must meet the following requirements [64] [74]:

- *Flexibility.* Capability of handling different granularities of bandwidth and to support a wide range of protocols. Techniques for dynamic allocation of capacity, needed in order to exploit more efficiently the limited network's resources.
- *Cost-effectiveness.* The appropriate network topology and protocols must be identified. Compared to the current technology, CAPEX (Capital Expenditure) and OPEX (Operational Expenditure) must be reduced by a considerable amount.
- *Upgradeability.* Ability to incorporate new technologies in an easy and non-disruptive manner must be present.
- *Scalability.* Ability to add network devices in an easy and non-disruptive way.
- *Efficiency.* High throughputs and short delays should be provided.
- *Fairness.* Starvation of nodes through a regulation of the bandwidth usage must be avoided.
- *Multicasting.* Efficient support to applications such as videoconferences or distributed games implies efficient multicast support.
- *Quality of Service.* Rapid provisioning capabilities and service guarantees to mission-critical data and delay-sensitive applications must be supported. The interaction between metro MAC protocols, and core and access network protocols must be taken into proper account in order to ensure an end-to-end QoS to customers.
- *Bandwidth management.* In order to control the amount of high priority traffic injected in the MAN and avoid congestion situations.
- *Reliability.* The network elements must offer a high degree of reliability. This mandates that critical sub-systems are fully protected and capable of in-service upgrades. Network recovery strategies must cover and work around network failures and ensure continuing availability of crucial services.

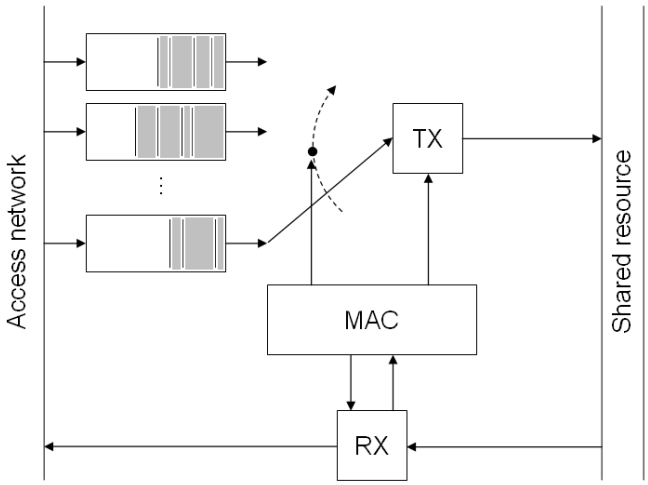

Figure 2.3: Schematic example of an OPS node for metro networks

OPS solutions appear to be a good candidate for future metro architectures to cope with these requirements. Besides the fact that it is a packet-based technology, it also avoids any electronic bottleneck conversion, and cuts the cost through a simplification of the structure. At the same time, the use of WDM channels allows to reaching high network capacity and the flexibility of the statistical multiplexing provides very high resource utilisation.

Since in WDM networks multiple channels are created by transmitting and receiving data on parallel wavelengths, two types of data collisions can occur: (i) transmission collision takes place when two or more nodes send data on the same wavelength channel simultaneously; and (ii) receiver contention takes place when a number of packets (in different wavelengths) must be received at the same time by a given node which has not enough receivers. Therefore, the nodes of the OPS-based metro networks generally use electronic buffer to store the packets coming from access networks accessing the shared wavelength channels by deploying a Medium Access Control (MAC) protocol which aims at either avoiding collisions and contentions or mitigating their impact on the network performance. Once a packet is transmitted, it runs the entire source-destination path without experiencing any additional delay (i.e., buffer-less path). Figure 2.3 shows a schematic example of the structure of an OPS node. It consists of a MAC chip which is in charge of selects from the set of electrical buffers the packets to be sent to the shared resource. To decide when and at which wavelength send the packet, the MAC can also listen the shared resource to identify a free space. At the same time, the MAC listen the shared resource to recognize the packets to be received by its node. It is therefore clear that all these processes strongly depend on several aspects of the metro network which include: network topology, node architectures, MAC protocols, and QoS strategies.

For what regards the physical topology, it can usually be a star [79], a linear bus [73], or a ring [94] as shown in Figure 2.4 for the case of WDM metro network. From a performance perspective, all topologies are currently equally attractive. To increase

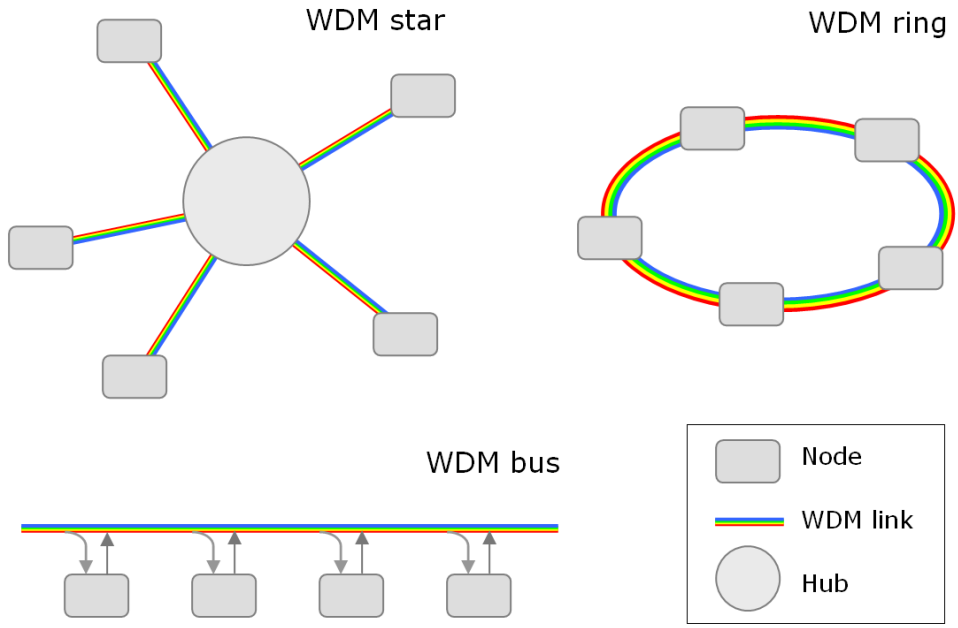

Figure 2.4: Example of physical topologies for metro networks

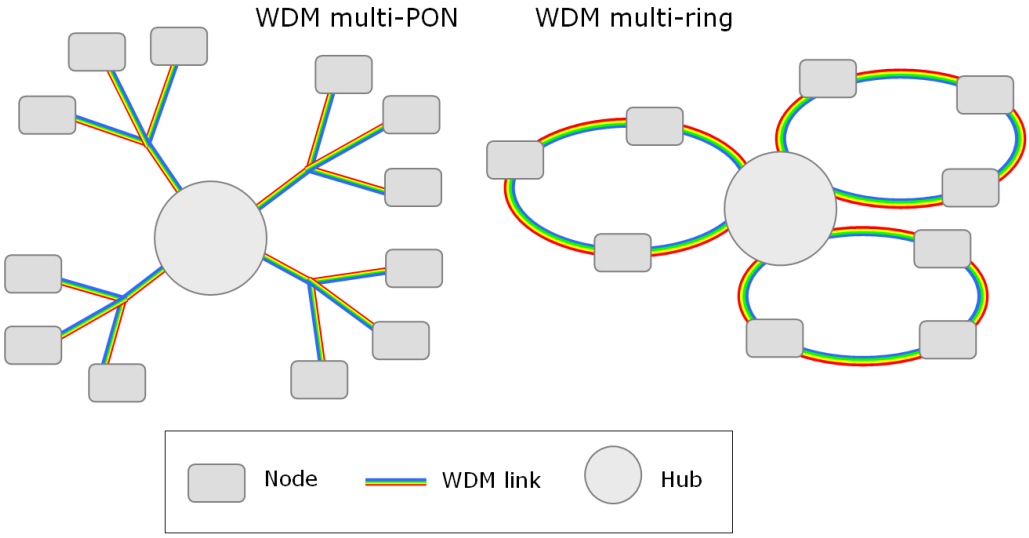

Figure 2.5: Example of composite physical topologies for metro networks

the number of nodes and/or the overall network throughput, it is possible to create composite topologies [9][25][46] as in Figure 2.5. In this case several small networks such as Passive Optical Networks (PONs) or rings can be connected through a central node forming multi-PON or multi-ring topology. In such topologies, the central node is usually referred as *Hub*.

For what regards the node architecture, the use of WDM systems opens up a wealth of possibility mainly related with the transmitters and receivers devices:

- Fixed Transmitter(s) and Fixed Receiver(s) (FT-FR);
- Tunable Transmitter(s) and Fixed Receiver(s) (TT-FR);
- Fixed Transmitter(s) and Tunable Receiver(s) (FT-TR);
- Tunable Transmitter(s) and Tunable Receiver(s) (TT-TR).

Fixed transceivers, which can only access some predetermined channels, are ready available in the market, but considerations often restrict the installation of a large number of such transceivers at each node. For TT-FR and FT-TR structures, no channel collisions will occur and simple MAC protocols can be employed, but the maximum number of nodes will be limited by the number of available channels. Systems based on the TT-TR structure are probably the most flexible in accommodating a scalable user population, but they also have to deal with the channel-switching time overhead of the transceiver [79].

In addition to the source and destination nodes, the star and composite topologies have a central node (i.e., the Hub). In this case, the Hub can be based on different architectures:

- Passive Star Coupler (PSC) [34] [66]. It is an N-input, N-output device with the property that the power from each input is divided equally among all the outputs. This implies that each transmitted packet is broadcasted at the Hub to all receivers which may simplify the design of the MAC protocol but presents poor performance due to the high redundancy and high transmission collision probabilities at the output ports of the Hub. Hence, it is a suitable solution only for low to medium network loads.
- Arrayed Waveguide Grating (AWG) [6] [82]. It is an N-input, N-output device which provides static space routing dependent on the input (port, wavelength)-tuple and offers frequency periodicity. Unlike the PSC the AWG allows for spatial wavelength reuse, i.e., each wavelength can be applied at all AWG input ports simultaneously without resulting in transmission collisions at the AWG output ports [74]. The design of the MAC protocol may result more complex and some feedback information about channels usage must be sent to the transmitter nodes to avoid both transmission collision and receiver contention.
- Semiconductor Optical Amplifier (SOA)-based switch [36] [43]. It is an N-input, N-output device which provides full connectivity between any input and

output ports. In this case, the configuration of the switch must be regulated by a Switch Control Logic (SCL) which is in charge of assigning output resources according to the incoming packets. The design of the MAC protocol may be easier than the previous case, but the switch needs advanced optical devices.

A large number of MAC protocols for metro WDM networks have been proposed and investigated in the literature. Many of those MAC protocols make use of well known *random access schemes* such as ALOHA and CSMA (Carrier Sense Multiple Access) or *controlled access schemes* such as token ring and FDDI and adopt them to the high-speed multichannel WDM environment. The interested reader can refer to [102] [79] [80] for an extensive survey on media access techniques.

At the same time, very few works have been developed to provide QoS in metro WDM networks [75] [49]. This is in contrast with the well defined schemes adopted in electrical metro network such as FDDI and DQDB [97] and in new electrical solution RPR [96]. Without lose of generality, we can classify the different type of QoS provided by the standardized metro network within one of the following classes:

- Isochronous. It refers to precesses that ensure that data flow continuously and at a steady rate in closing timing.
- Synchronous. It refers to data that have real-time constraints, implying that they need to be delivered in a certain time with a maximum delay. Synchronous data usually have reserved bandwidth.
- Asynchronous. It refers to data that can be sent at the leisure of the node. Asynchronous data can only use the leftover bandwidth. The asynchronous service may include more than one sub-service with different priority among them [97].

2.3.1 Examples of OPS-based metro network prototypes

In this section, we briefly describe three OPS-based metro networks currently under development in different research/university centers. These examples demonstrate the world-wide interest and the viability of such solutions.

AWG-STAR (Figure 2.6) is an NTT solution based on an AWG based star metro WDM network which interconnect n nodes [82]. Each node has 2 fixed-tuned transmitters and n fixed-tuned receivers. Each node transmits data packets on the same 1.55 μm wavelength and the corresponding header on the same 1.3 μm wavelength. To enable communication between any arbitrary pair of nodes, wavelength have to be converter at the central AWG. A WDM coupler routes the header to the optical header analyzer which determines the target wavelength on which the packet has to be transmitted to reach the destination node. The data packet is amplified and forwarded to the wavelength converter which consists of multiple light source each operating at different wavelength.

HORNET (Figure 2.7), which stands for Hybrid Opto-Electronic Ring NETwork, is a WDM time-slotted ring developed at the Stanford University [94]. HORNET has

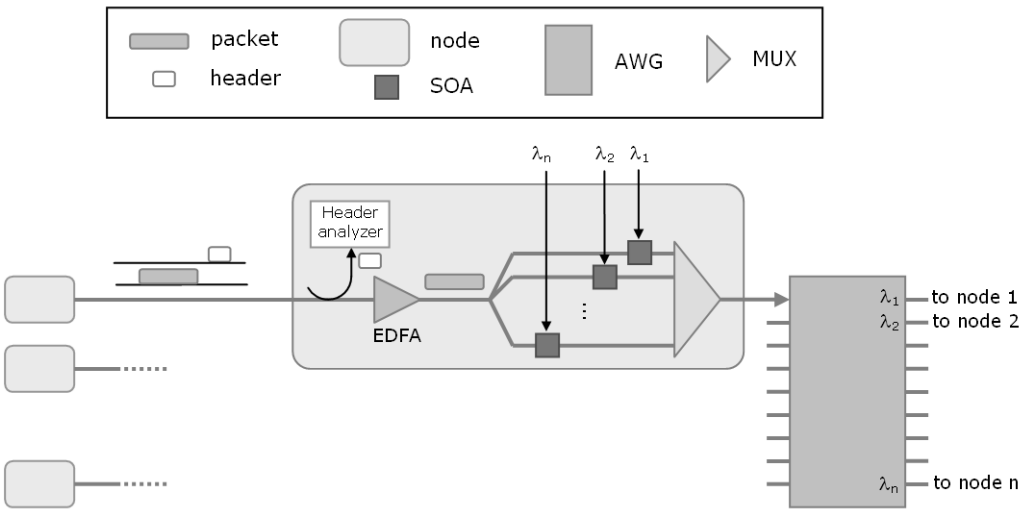

Figure 2.6: Structure of the AWG-STAR metro network

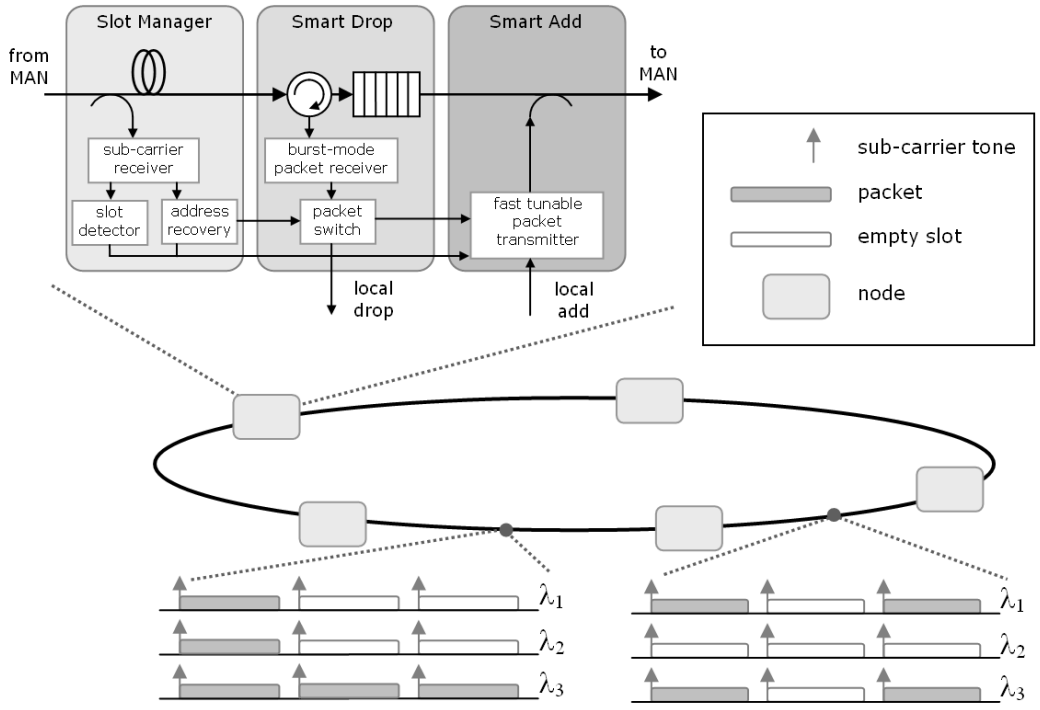

Figure 2.7: Structure of the HORNET metro network

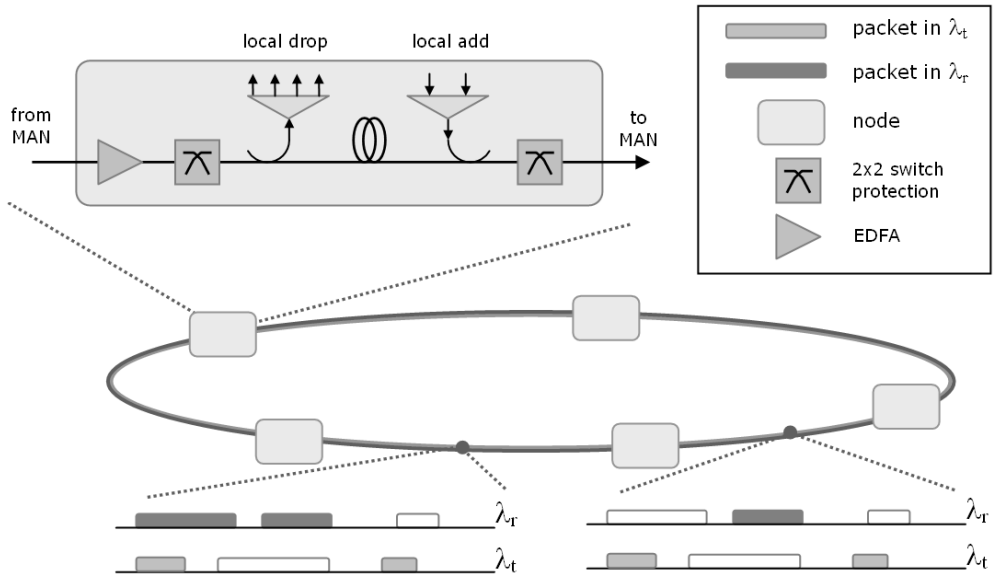

Figure 2.8: Structure of the DBORN metro network

a tunable transmitter, fixed receiver design (TT-FR), and the nodes use a CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) MAC protocol to govern access to the wavelengths (see [94] for more details on the access protocol). HORNET can use either slotted or variable-length packets- a characteristic which can provide flexibility. It can be based on two counter-rotating rings to offer cut-fibre protection, nevertheless, a failure will result in halving the available bandwidth. Multicast can be provided via node-by-node re-transmission, but no protocol is included and evaluated to handle multicast traffic and incoming traffic at the nodes. Neither QoS strategies nor fairness mechanisms are implemented. No performance evaluations to assess the merits of such architecture are available.

DBORN (Figure 2.8), which stands for Dual Bus Optical Ring Network, is based on a unidirectional fibre ring organized around a Hub developed at Alcatel Research and Innovation [73]. This architecture uses a spectral separation of upstream and downstream flows from/toward the Hub, forming a dual logical bus structure. Nodes dynamically read data on the downstream bus and write on the upstream bus, while the Hub interconnects the buses through a wavelength conversion. The spectral separation avoids the use of erasing functionality at the nodes increasing the nodal cascadability. A simple collision avoidance MAC is implemented through power detection utilizing a photodiode and a fixed-length delay line. The network can support any client packets and makes it easy to add/ remove nodes to/from the network. Some cost studies have shown the benefits of this architecture [73]. A performance evaluation is not available; neither QoS strategies nor fairness mechanisms have been implemented. The current protection mechanism is based on duplicating the network components.

2.4 Wide area network context

As stated in previous sections, the key topics in wide networks are the contention resolution algorithms and the node architectures designs. In this section we give a short overview of the current proposals.

2.4.1 Node architectures

The general architecture of a OPS node is presented in Figure 2.9, showing the main functional blocks:

- *Input interface* used to demultiplex the wavelengths on incoming fibers, to synchronize packets, to tap some optical power, packet header information extraction, and packet header removal.
- *Optical switching* used to perform the packet switching in optical domain solving the possible contentions. This block can be formed by different sub-blocks:
  - *Optical space switch* is a non-blocking space switching matrix used to physically interconnect the input and output ports;
  - *Optical buffer* used to solve contentions in the time domain;
  - *Wavelength converters* used to solve contentions in the wavelength domain.
- *Output interface* used to regenerate the optical signals, insert the packet header, and to multiplex wavelengths back to the fiber.
- *Switch control logic* used to perform header processing and rewriting, routing table lookup, as well as to set-up the optical devices to proper arrange the optical switching.

Several proposals for an OPS node architecture can be found in literature [104]. We can categorise them in two fundamental ways:

- There can be a single stage or multiple stages;
- The contentions can be solved either using a feed-forward or feedback configuration.

Single-stage architectures are generally easier to implement but may require complex scheduling algorithm to solve packet contentions. To increase the buffering capacity, the multi-stage architecture cascades many small switches, forming a larger switch with a greater buffer depth. Figure 2.10 depict a schematic example of the single stage and multi-stage node architecture.

In a feed-forward configuration (see Figure 2.11(a), when packets enter in a stage, the possible contentions are solved once and then the packets are sent to the next stage or to the output; this implies that the packets travel along a constant number

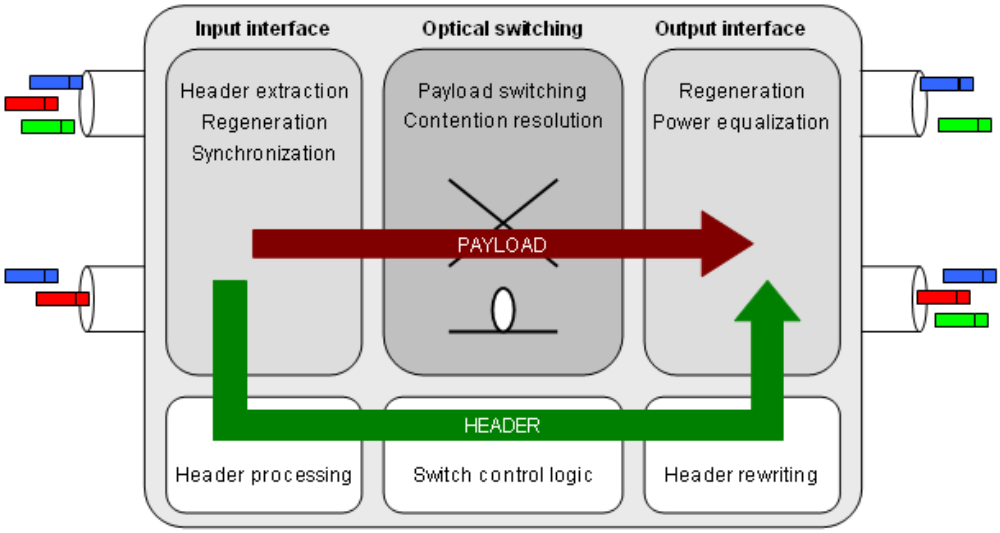

Figure 2.9: A generic OPS node architecture

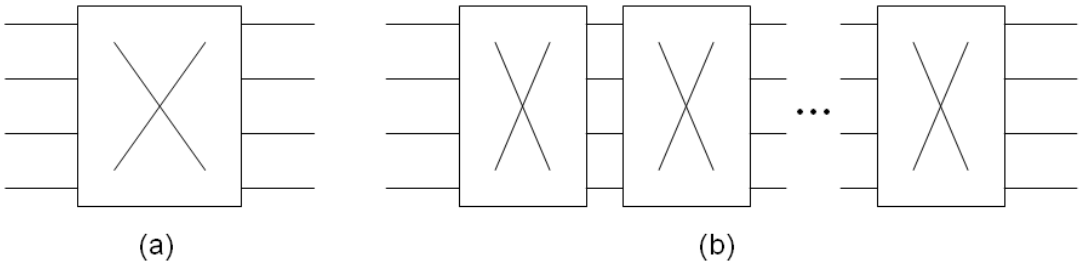

Figure 2.10: Schematic example of the (a) single-stage and (b) multi-stage node architecture

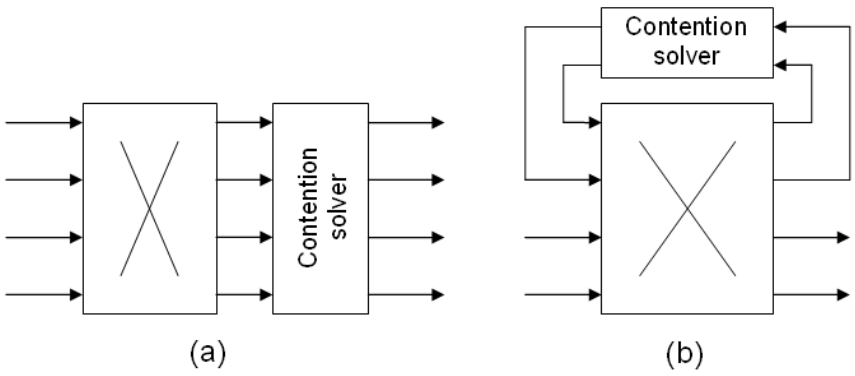

Figure 2.11: Schematic example of the (a) feed-forward and (b) feedback node architecture

Table 2.1: Example of node architectures

	Single stage	**Multi-stage**
Feed-forward	OASIS [60]	SLOB [59], WASPNET [61]
Feedback	SMOP [60], DAVID [43]	

of stage. In a feedback configuration (see Figure 2.11(b), the packets can back to the input of the same stage in order to solve again the contention if the previous one was not sufficient; this implying that the delay generally differs between packets.

In table 2.1 we indicate some example of node architectures proposed in literature according to the classification described above.

Since this thesis is not focussed on architecture for optical packet switching we do not enter into specific technological aspects but deal with the contention resolution problem, which is a mandatory task for whatever architecture.

2.4.2 Techniques for contention resolution

Since OPS is based on statistical multiplexing, packet contentions may arise at each node. Indeed, when a packet enters a node and, according to its header, must be forwarded to a given output fiber and wavelength, it may happen that the requested resource is unavailable because occupied by another packet. Therefore, a contention resolution policy must be applied to reduce the packet loss probability and make the statistical multiplexing more efficient. Contention resolution techniques typically adopted by OPS networks exploit space, time and wavelength domains.

Space domain

Deflection routing is ideally suited to switches that have little buffer space. When there is a conflict between two packets, one will be routed to the correct output port, and the other to any other available output port. In this way, little or no buffer is needed. However, deflection packet may end up following a longer path to its destination. As a result, the end-to-end delay for a packet may be unacceptably high. Also, packets will have to be reordered at the destination since they are likely to arrive out of sequence.

Time domain

The implementation of optical buffering is the major problem. Optical RAM does not exist and in order to implement optical buffering, resort must be made to optical fibre delay-line (FDL) [48]. The alternative, of course, is to convert the optical packet to the electrical domain and store it electronically. This is not an acceptable solution, since electronic devices cannot keep up with the speeds of optical networks. Other solution is to use a hybrid approach; where both electronic and optical buffers are used in cooperation [76]. This solution relaxes the O/E/O bottleneck but introduce

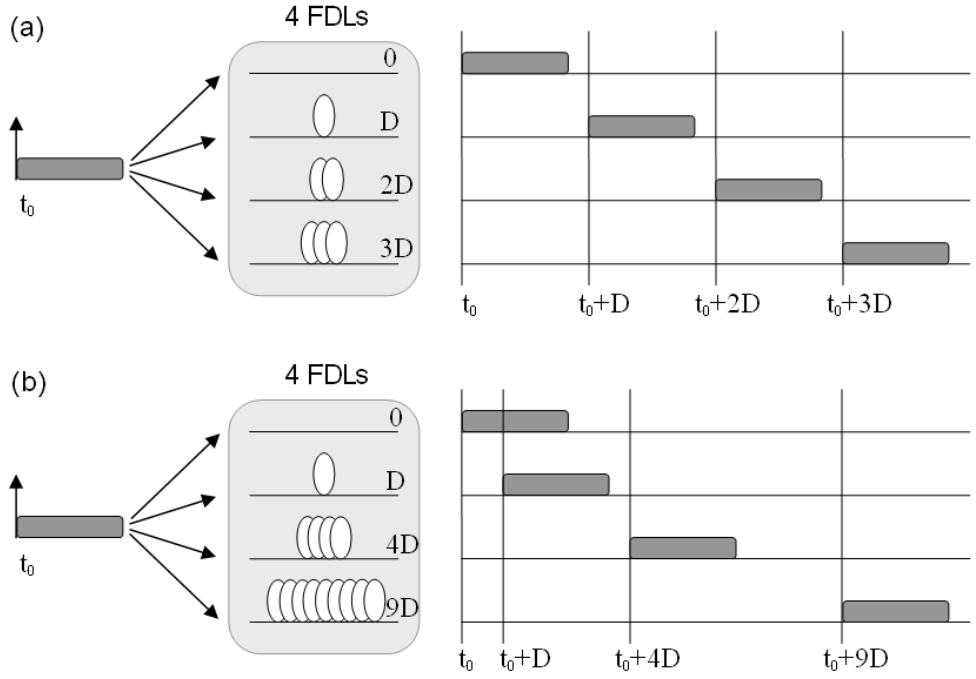

Figure 2.12: Buffer configurations with 4 FDLs: (a) Degenerate $k_j = j - 1$, $\mathbf{Q}_4 = \{0, D, 2D, 3D\}$, (b) Non-degenerate $k_j = (j-1)^2$, $\mathbf{Q}_4 = \{0, D, 4D, 9D\}$

more problems than benefits since both very high speed electronic devices and FDLs are required.

Therefore, in general an optical buffer is made by a pool of B FDL with different lengths providing delays which are multiple of a basic delay unit D, also known as *buffer granularity*. The set of possible delays are denoted by $\mathbf{Q}_B$ and the delay provided by the j-th delay line is $D_j = k_j D$ with $j \in \{1, 2, \ldots, B\}$. A measure of the buffer capacity is given by the *maximum available delay* $D_M = \max_{j \in \{1,2,\ldots,B\}} \{k_j D\}$.

Two examples of optical buffers are shown in Figure 2.12. The first configuration (Figure 2.12(a) assumes $k_j = j - 1$ and corresponds to delays which are consecutive multiples of the delay unit D: the first fibre does not introduce any delay, the second one provides a delay equal to D, the third to $2D$ and so on until the B-th fibre which offers the maximum available delay $D_M = (B-1)D$. Therefore, $\mathbf{Q}_B = \{0, D, 2D, \ldots, (B-1)D\}$. This uniform delay distribution is the simplest case and can be considered a sort of reference. Such configuration has usually referred as *degenerate* buffer. Another possible configuration is the *non-degenerate* buffer where non-uniform delays are used and k_j can assume any value. Figure 2.12(b) shows an example of a non-degenerate buffer where $k_j = (j-1)^2$, and therefore $\mathbf{Q}_B = \{0, D, 4D, \ldots, (B-1)^2 D\}$.

It is clear that D is a key parameter in the buffer design, and it must be accurately chosen. It strongly depends on the format of the optical packets and on the network operation. In the case of FLP-SO and VLP-SO, the value of optimal D is related with the length of the packets. In the case of VLP-AO, the value of optimal D depends

on the average length of the packets and on the contention resolution algorithms as demonstrate for example in [17].

Finally, as buffering tool, FDLs are bulky and not scalable. Compared to electronic buffers and their role in current packet networks, FDLs can only offer limited buffering capabilities (few tens of possible delays at maximum [45]).

Wavelength domain

In WDM, several wavelengths run on a fibre link that connects two optical switches. This can be exploited to minimize the contention probability as follows. Let us assume that two packets are destined to go out of the same output port at the same time. Then they can be still transmitted out, but on two different wavelengths. So then, wavelength converters are used to change the wavelength of the packets in contention in such a way that multiple packets can be sent simultaneously to the same output port. As a result, wavelength conversion is the most effective contention resolution, not incurring additional latency while maintaining the shortest path or minimum hop distance. This method may have some potential in minimizing contention, particularly since the number of wavelengths that can be coupled together onto a single fibre continues to increase.

The recent works on the contention resolution algorithms and QoS mechanisms are discussed in Chapters 7 and 9, respectively.

2.4.3 Example of OPS-based backbone network prototypes

In this section, we briefly describe two OPS-based backbone networks developed in different research/university centers. These examples demonstrate the world-wide interest and the viability of such solutions.

WASPNET [61] is a packet-based optical WDM transport network proposing a node design based on wavelength router devices (Figure 2.13). The network operation is synchronous and the packets are fixed-length. All the inputs are wavelength demultiplexed to a number of parallel planes, one for each wavelength channel, each plane containing an optical packet switch. A combiner merges the wavelength channels from all planes, allowing dynamic wavelength allocation at each plane. The AWG in each plane routes packets to the space switch, which then switched each of them to the correct output port at the correct wavelength via the final stage of wavelength conversion. The contended packets are routed to the recirculating loops for buffering, while the straight-through packets are routed to switch outputs. Contention resolution is implemented assuming a FIFO algorithm and ensures that no more than one packet may exit from each combination of FDLs, a demultiplexer and a multiplexer; otherwise, it will be router to other alternative AWG outputs or, if necessary, to longer-delay lines.

DAVID [43] proposes a broadcast-and-select architecture, which ensures nonblocking performance using SOA technology, as illustrated in Figure 2.14. A shared recirculating FDL buffer is used to help solving contention, where exploitation of the wavelength domain does not suffice. The switch operates synchronously and both

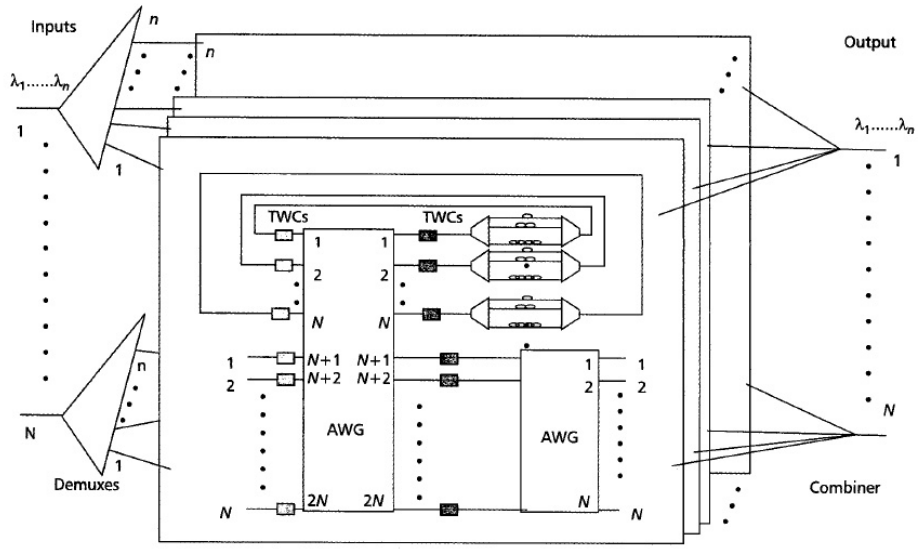

Figure 2.13: Structure of the WASPNET switch node

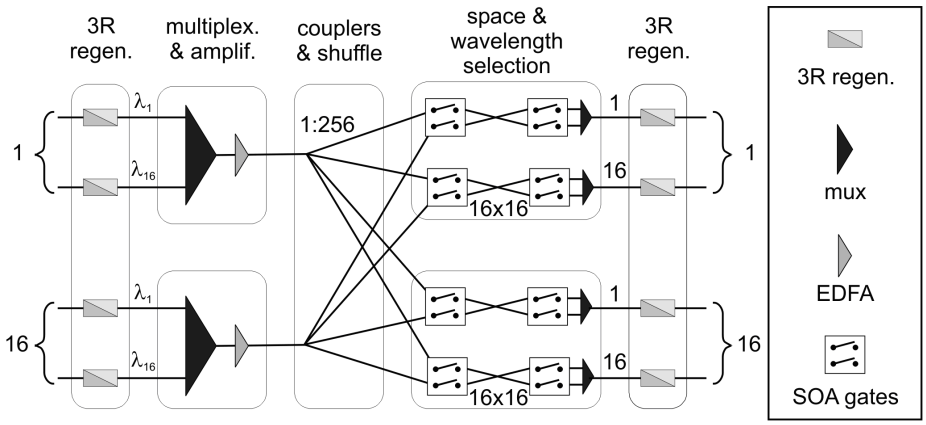

Figure 2.14: Structure of the DAVID switch node

FLP and VLP packet formats alternatives are studied. For the former, a simple first-fit algorithm is proposed and applied to two different buffer structures. For the latter, several algorithms are suggested showing that the optimization of the usage of the FDL must be the main aim to reach acceptable performance.

PART II

OPS-based metro area networks

Chapter 3

Introduction to the OPS-based metro area networks

Some traffic estimates for the UK network over the next few years [39] indicate that when access is primarily over copper-based technology the total traffic volume will be a few Terabit/s. On the other hand, in the event of a mass take-up of FTTH these traffic estimates would reach the order of tens of Terabit/s. In this eventuality, the DAVID project has been developed two advanced network architectures based on composite topologies able to achieve more than 1 Terabit/s throughput, namely multi-PON (on the left side in Figure 3.1) and multi-ring (on the right side in Figure 3.1). They are also known as interconnected WDM PONs and interconnected WDM rings, respectively.

In this thesis we exclusively focus on these architectures. They have common features that we discuss below; the description concerns to the multi-PON architecture while between brackets we refer to the multi-ring case. In the following chapters we separate the studies: at first we focus on the multi-PON and thus on the multi-ring architecture describing for both cases the state-of-the-art and our contributions.

Multi-PON network consists of several uni-directional slotted optical WDM PONs [ring in the case of multi-ring] interconnected in a star topology by Hub. The Hub connects the PONs [rings] in a metro area to at least one optical packet router in the backbone. Nodes are connected to PONs [rings] and provide an electro/optical interface to edge routers/switches at the end of access networks via a variety of legacy interfaces (e.g., IP routers or Ethernet switches). Nodes belonging to the same PON [ring] share the same set of resources (i.e., a given fixed number of wavelengths) in such a way that a MAC protocol is required to arbitrate the access.

Nodes use a statistical time/wavelength/space division multiple access scheme. Indeed, time is divided into slots (Time Division Multiple Access, TDMA), lasting 1 μs (1250 bytes at 10 Gbit/s) since previous studies [30] [13] have shown this is a good compromise between filling ratio optimisation (sufficient level of aggregation in the electrical domain) and delay. Several slots are simultaneously transmitted on different wavelengths (Wavelength Division Multiple Access, WDMA) on the same PON [ring], and PONs [rings] are disjoint in space (Spade Division Multiple Access, SDMA). The network is synchronous and time slots are aligned on all wavelengths

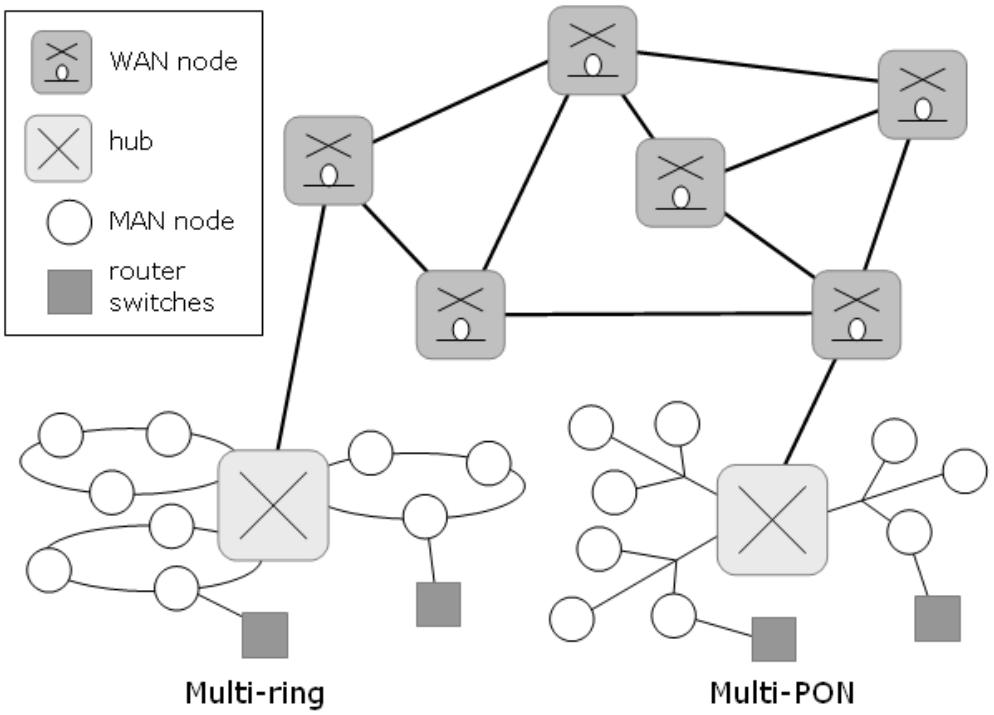

Figure 3.1: Metro network architectures considered in this thesis

of the same PON [ring]; thus, a *multislot* (i.e., a set of slots, one per wavelength) is available at each node in any time slot. The information on the status and availability of network resources is transported out-of-band in a control channel, i.e., a dedicated additional wavelength on each PON (ring). This means that each multislot includes one *control* slot and several *data* slots.

No packet buffering is available at the Hub; therefore buffering is done electronically at nodes. The Hub behaves as a switch between PONs [rings]. PON [ring] interconnections are dynamically modified following a scheduling algorithm. It also arbitrates the allocation of network resources and nodes' access to slots. It is typically run in a centralized fashion at the Hub, although some decisions can be decentralized at network nodes. Nodes must have access opportunities in proportion to a request matrix while avoiding collisions (i.e., transmission in busy slots) and contentions (i.e., sending in different slots at the same time more than one packet addressed to the same node). The request matrix is created at the Hub and can be calculated using explicit bandwidth requests issued by nodes, estimated with traffic measurements or pre-defined by the network management operator.

For these scenarios it is reasonable to think that multimedia and interactive applications will take an important share of the bandwidth, and hence techniques to provide QoS must be designed. The simplest way to do that is by introducing at least two traffic classes having different priority [97]. A drawback of such strategy is that if high-priority traffic is not controlled by any form of Call Admission Control (CAC), the traffic fluctuations can cause two different undesirable situations: 1) high-priority

traffic with strict priority over low priority traffic can prevent the transmission of the latter and 2) it may not be possible to guarantee neither delay nor bandwidth bounds even to high-priority traffic. To avoid this situation in public networks, centralized bandwidth management functions (i.e., traffic engineering) are required. Network operators need to have the possibility to control the amount of high priority traffic injected in the (optical) MAN segments.

To attain this objective, at first we need to determine which services are required in a metro environment. According to what developed in the electrical MAN, we can consider that 3 different classes of service must be supported. These classes are:

- *Guaranteed service.* It refers to data that have real-time constraints that require a guaranteed bandwidth and low jitter. This service must have absolute priority over the other types of services, and must be shaped at the ingress. To provide such services, a reservation mechanism is usually adopted.

- *Priority service.* It is dedicated to near-real-time applications that are less delay and bandwidth-sensitive. In contrast to the guaranteed service, the bandwidth for priority service is not statically allocated but some forms of priority mechanism over best-effort service are usually used.

- *Best-effort service.* It refers to data that can be sent at the leisure of the node. This service is usually weighted by a fairness mechanism in order to ensure that each node gets its fair share of the bandwidth available.

In the following chapters, we present the related work inside of the DAVID project and our contributions which complete the investigations on these network architectures.

Chapter 4

Multi-PON architecture

4.1 State-of-the-art

The network concept (Figure 4.1) is based on a star topology presented in [24], which has been adapted to a metro environment with a target throughput of 1 Terabit/s per second. An AWG Hub interconnects several nodes through Passive Optical Networks (PONs). Each PON has a number of nodes attached to it that may be routers, access networks' head-ends, gateways to/from core networks, or any other kind of network node with a proper interface to this optical packet network.

The switching functions (time switching and lambda switching) are distributed between the nodes and the AWG. The nodes have electronic buffers, where electrical packets are stored and aggregated into longer packets before entering the optical domain, and rapidly tuneable transceivers (Tx/Rx). At a bit rate of 10 Gbit/s the tuning speed should be a few nanoseconds. Nodes buffer the incoming packets electronically. For every packet, the node sends a request to the Network Controller (NC). The NC is hence responsible of allocating the network resources scheduling the requests. In this way, the network can be thought of as a single, big, distributed optical packet router where the complexity sharing with the nodes allows the most use of edge buffers and Tx/Rxs. The central node of the star network uses an AWG; an $N \times N$ multiplexer based on a waveguide grating providing static space routing dependent on the input (port, wavelength)-tuple and which offers frequency periodicity. The same pool of wavelengths is available to each PON in the network. The AWG has Wavelength Converter (WC) arrays, positioned between extra dummy ports, which, in combination with the tuneable Tx/Rxs in the nodes, in effect can provide a fully non-blocking, distributed 3-stage switch fabric between nodes if required. Each WC array has the same number of converters as the product of the number of PONs and the number of grating frequency periods used. The number of WC arrays required depends upon the traffic level and pattern.

4.1.1 MAC protocol

The network employs resource sharing based on WDMA/TDMA, where the time is further organized in frames, each frame containing F slots, as illustrated in Figure 4.2.

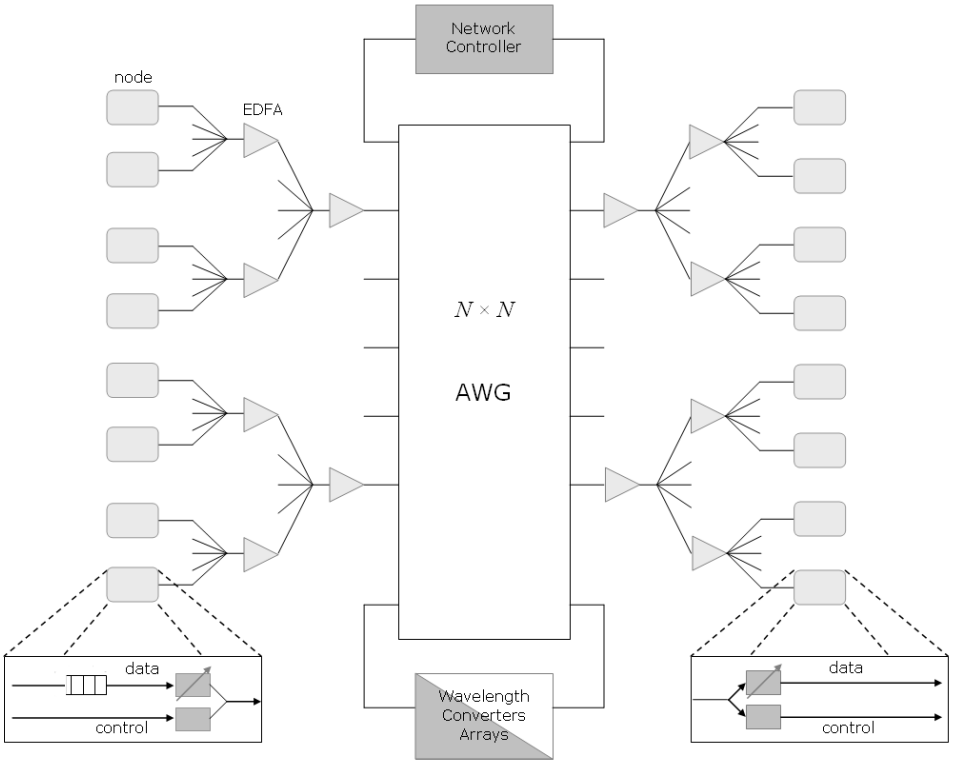

Figure 4.1: Multi-PON architecture

Slots run in parallel and are synchronized among all wavelengths in the same PON, so that a *multi-slot* (a slot per each wavelength plus the control slot) is available at each node in each time slot. The control slot is further partitioned into M *mini-slots*.

As a consequence of this scheme, the nodes see the network as a set of dynamic TDMA sub-networks (one per wavelength) in a way that a MAC protocol is required to avoid transmission collisions (multiple transmission in the same time slot) and receiver contention (more than one node trying to transmit to the same receiver at the same time). The MAC protocol is based on a bandwidth demand scheme. A node wanting to transmit sends a request (using a mini-slot) to the NC indicating the address of the destination node and the amount of slots required to accommodate the packet. A static TDMA protocol is used in the control channel, i.e., the node i has m_i reserved set of mini-slots available in each frame for allocating its requests (as shown in Figure 4.2). Therefore, the NC schedules the requests end-to-end between nodes. When a request has been successfully scheduled, the NC advises the node of the time slot and wavelength channel it has allocated to the packet. The following section discusses this issue.

4.1.2 Scheduling algorithms

The high targeted throughput of this network and the large number of nodes make impractical the implementation of optimum algorithms; it is proved in [6] that the

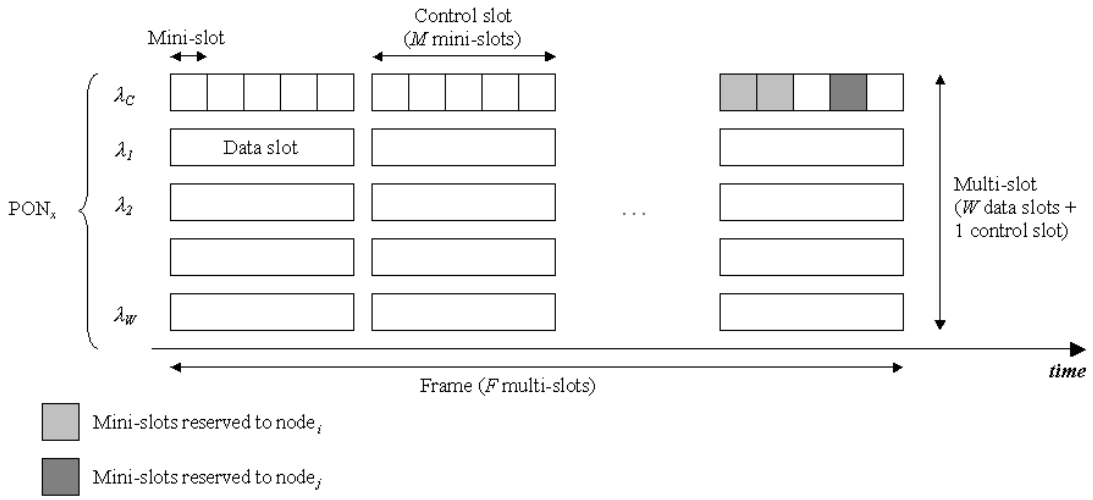

Figure 4.2: Timing structure of the wavelength channels

scheduling problem is NP-hard. Among all possible solutions, two heuristic scheduling algorithms have been considered:

- *Greedy algorithm* [6].
- *Frame-based algorithm* [31].

The Greedy algorithm is based on the separation of the time domain from the wavelength domain using two different algorithms:

- The *Slot Allocation* (SA) algorithm that operates always on a given fixed wavelength topology configuration based on a set of matrices, which provide a map of the network's available resources. According to the contents of these matrices, the algorithm selects the requests to be accepted based on a priority rotation scheme.
- A *Logical Topology Design* (LTD) algorithm that aims to obtain an efficient allocation of available wavelengths in order to reduce the slot allocation failure rate.

The Frame-based algorithm requires two steps for each frame: F-matching and Time Slot Assignment (TSA). The F-matching problem consists in finding the maximum subset of admissible packets that can be scheduled in a frame of length F slots. The TSA problem consists of scheduling through the switch the set of accepted non-conflicting slots in the frame and allocating wavelength channels to them.

For an $N \times N$ switch, the F-matching algorithm selects, at the end of the current frame, a set of up to $N \times F$ non-conflicting packets that will be transmitted in the slots belonging to the next frame. The matching algorithm accepts the set of non-conflicting packets always satisfying the following two non-overbooking properties:

- the number of accepted packets from each input port cannot be higher than F,
- the number of accepted packets to each output port cannot be higher than F.

According to these constraints, the matching algorithm runs through the well-known three phases [77] adapted to the frame length [8]:

1. *request*: each queue requests a number of slots from the corresponding appropriate output port in the frame,
2. *grant*: each output port issues up to F grants distributed amongst the queues destined for that output,
3. *accept*: each input port accepts up to F of the grants received at the port, where each acceptance received by a queue gives the right to use one slot in the next frame.

The selection of requests to be granted, and of grants to be accepted is based upon a rotating priority scheme, which is implemented using two sets of pointers, one for each input and one for each output. Several pointer use/update rules were defined in [7] that led to the definitions of different variants of the frame-based matching algorithm. Since the adoption of any variant of the frame-based matching algorithm does not affect the techniques we propose, we use the rules called NOB8, where the input (output) ports move their pointer to the output (input) port following the last one to which it gave an acceptance (grant).

For what regards scheduling a set of non-conflicting requests in a time frame (TSA problem), it is possible to see that the problem is equivalent to the routing of circuits in a Clos interconnection network, for which several algorithms are available in the literature; some of these algorithms are suited for parallel implementations (see, e.g., [72]), leading to complexities which can be smaller than the complexity of a typical sub-optimal F-matching algorithm. In this study we used the well-known sequential TSA algorithm proposed in [63].

4.1.3 QoS provisioning

Two different classes of services are considered in previous works: a *persistent* and *non-persistent* service [6]. The former is a connection-oriented approach which provides isochronous service: a node requiring it sends a persistent request to the NC indicating the amount of slots required in each frame (connection setup). Therefore, resources are allocated until one of the two involved terminals signals that it wants either release some slots for the connection or close it. The latter is based on a best-effort approach which provides asynchronous service; the nodes send a nonpersistent request and the resources are allocated only in a single frame; in the next frame, the NC will release all the resource previously allocated to best-effort traffic.

4.2 Contributions

Our contributions include: (i) the performance evaluation of these architectures and the identification of the weaknesses and of the open issues, (ii) the optimization of the proposed architectures and MAC protocols, and finally (iii) the proposal of a QoS mechanism to support priority and best-effort services.

4.2.1 Simulation scenario

The performances of the proposed mechanisms are evaluated in order to assess their merits. The simulation results presented in the following sections have been obtained by means of an ad-hoc event-driven simulator reproducing a real scale configuration of the multi-PON network. The parameters of the network are:

- P indicates the number of PONs;
- n indicates the number of nodes per PON;
- W indicates the number of wavelengths per fiber;
- WCA indicates the number of wavelength converter arrays;
- B_w indicates the bit-rate. It is set to 10 Gbit/s in every simulation scenario;
- P_s indicates the duration of the time-slot. It is set to 1 μs;
- F indicates the number of time-slots per frame;
- m_i indicates the number of mini-slots per control slot. It is set to 15;
- L indicates the distance between the Hub and the nodes. It is set to 20 km;
- ρ indicates the offered load;
- $\mathbf{M}$ indicates the PON-to-PON traffic matrix, whose generic element $\mathbf{M}_{i,j}$ is a real number ranging between 0 and 1 representing the percentage of traffic coming from input PON i and going to output PON j with respect to ρ. Four different traffic matrix are defined named: *uniform* $\mathbf{M}^U$, *power-of-two* $\mathbf{M}^P$, *diagonal* $\mathbf{M}^D$, and *dynamic diagonal* $\mathbf{M}^{DD}$. For the case of $P = 4$, the matrices are as follows:

$$M^U = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \qquad M^P = \frac{1}{15} \begin{bmatrix} 1 & 2 & 4 & 8 \\ 2 & 4 & 8 & 1 \\ 4 & 8 & 1 & 2 \\ 8 & 1 & 2 & 4 \end{bmatrix}$$

$$M^D = \begin{bmatrix} 0.55 & 0.15 & 0.15 & 0.15 \\ 0.15 & 0.55 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.55 & 0.15 \\ 0.15 & 0.15 & 0.15 & 0.55 \end{bmatrix} \quad M^{DD} = \begin{bmatrix} 0.4 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.4 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.4 \end{bmatrix}$$

For the dynamic diagonal traffic matrix the values on the columns shift on the right each 200 ms. This means that if the figure shows the situation at time 0, after, for example, 200 ms the new values $\mathbf{M}^{DD}_{i,j}$ are equal to $\mathbf{M}^{DD}_{i,|j-1|_4}$ of the figure.

These matrices represent a good sample of all possible traffic patterns: the classical uniform matric to evaluate a fair situation, the power-of-two matrix which demonstrates performance degradations when applied to the scheduling, the diagonal matrix to evaluate the inter-PON unbalanced situation, and finally a dynamic matrix to evaluate the behavior under fluctuations.

The electrical queues at nodes are considered infinite.

Concerning the traffic model, the packet interarrival time for any traffic class follows a self-similar process implemented as a superposition of 16 strictly alternating independent and identically distributed ON/OFF sources. The duration of each ON/OFF period was assumed to be a random variable with a Pareto distribution with shape $\alpha = 1.2$, which leads to a Hurst parameter of $H = 0.9$ [103]. All packets have the same size and fit in one slot.

Note that we assume an IP self-similar traffic model, since it is reasonable to think that IP will be the predominant traffic in metropolitan networks.

The number of simulated packets is chosen big enough to reach steady-state results and a 95% confidence interval is calculated.

We define the following measures to evaluate the performance of the metro network:

- *Throughput.* It is the usual performance measure and is calculated as the ratio between used and available slots.
- *Average and Maximum end-to-end delay.* They are, respectively, the average and maximum time needed to transmit a packet between a pair of nodes including both the service and transmission time.
- *Average Packet Loss Rate* (PLR). Assuming infinite buffer lengths, a packet is lost when the NC fails to serve the corresponding request within the QoS constraints. It is calculated as the ratio between the lost and generated packets.

In the following evaluation sections, we will only show the most significant measures according to the purposes of the study.

4.3 Performance evaluation

For this architecture, we have firstly performed an exhaustive performance evaluation in order to determine possible drawbacks of both the architecture and MAC protocol.

As an example, we show the results of the multi-PON network using the Greedy algorithms and considering $P = 4$ PONs, $n = 250$ nodes, $W = 32$ wavelengths, $WCA = 17$, and $F = 100$ slots per frame.

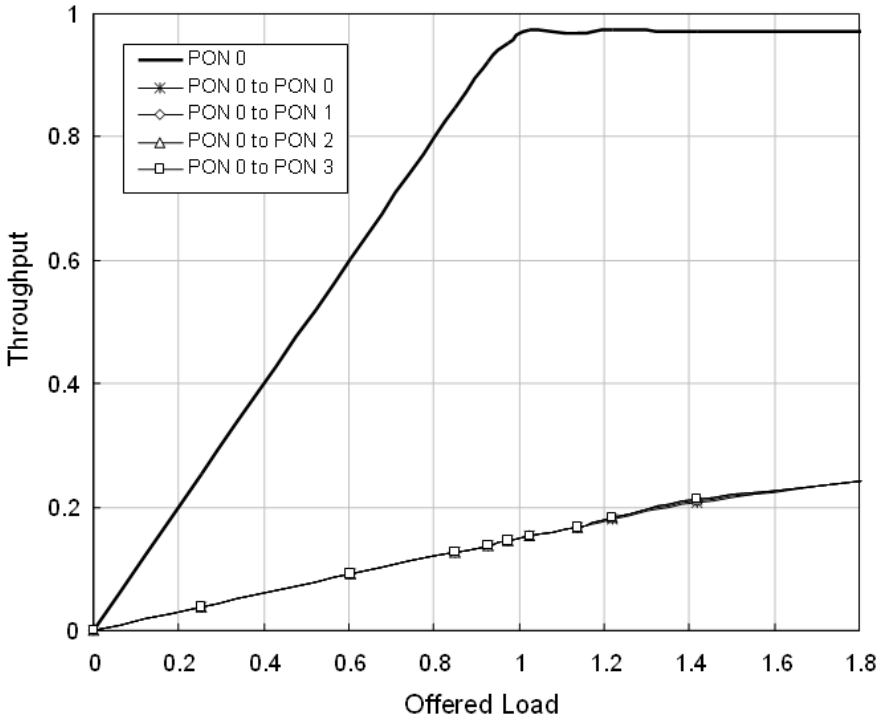

Figure 4.3: Throughput as a function of the offered load under uniform traffic matrix

Figure 4.3 shows the throughput per each destination PON reachable from PON 0 as a function of the offered load using the uniform traffic matrix. Although we report the throughput for a single PON, the same behavior holds for all other PONs due to the traffic symmetries. We can observe that the throughput increases with the offered load until it reaches the saturation. The PONs fairly use the resources as the overlapped curves indicates.

In Figure 4.4, we report the throughput per each destination PON reachable from PON 0 as a function of offered load under diagonal traffic matrix. From Figure 4.4, we can see that some degree of fairness between competitive traffic is enforced. For low value of offered load (ranging from 0.1 to 0.95), the throughput of each PON grows proportionally to their level. As soon as the offered load increases to values that create congestion, the network treats the PON-to-PON connections fairly according to a max-min like fairness criteria; i.e. the amount of admitted PON 0-to-PON 0 traffic is decreased to equalize the amount of traffic transmitted from each PON.

Next experiment is set up to evaluate the capacity of the network to cope with the rapid traffic fluctuations using the dynamic diagonal traffic matrix.

In Figure 4.5(a), we examine the throughput from PON-0 to PON-0, and from PON-0 to each of the others possible destinations (PON-1, PON-2 and PON-3), as a function of time under dynamic unbalanced traffic near the saturation point ($\rho = 0.92$).

Figure 4.5(a) shows that the throughput per destination PON is always close to optimal one. That is, when the load of PON i-to-PON j interconnection is $\mathbf{M}_{i,j}^{DD} = 0.4$, the throughput is $0.4 \times \rho$; also when the $\mathbf{M}_{i,j}^{DD} = 0.2$, the throughput is $0.2 \times \rho$.

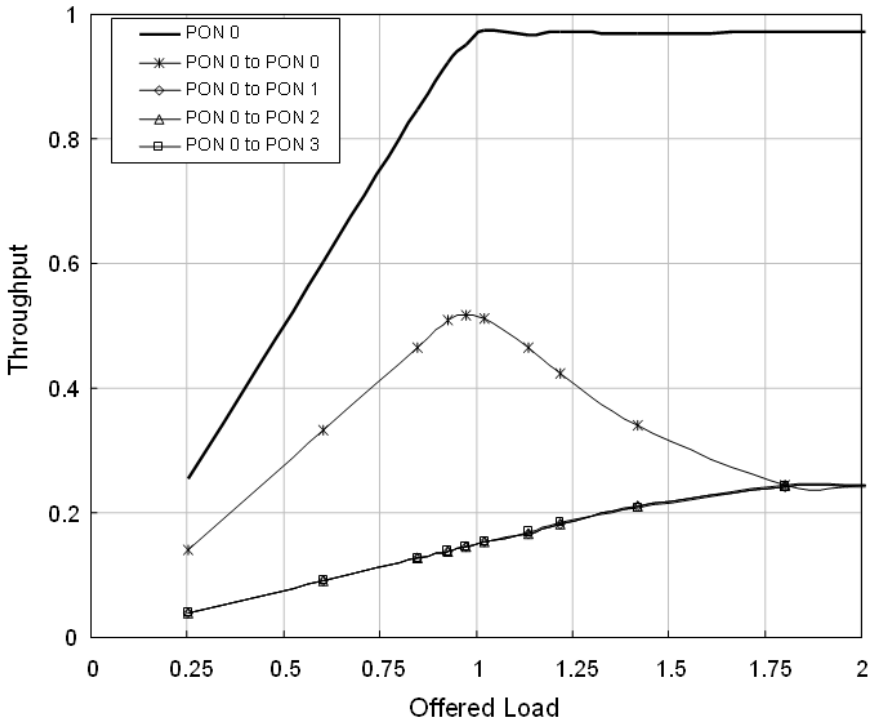

Figure 4.4: Throughput as a function of the offered load under diagonal traffic matrix

Moreover, It shows that the throughput per destination PON reacts very quickly to the traffic fluctuations. To confirm this point, Figure 4.5(b) reports an enlargement around 200 ms where a fluctuation occurs; with this figure, we can evaluate a transient behavior of less than 5 ms.

In next figures we compare the performance of the Greedy algorithm and the Frame-based algorithm considering a network with $P = 4$ PONs, $n = 32$ nodes, and $W = 32$ wavelengths. The power-of-two traffic matrix is used for this study.

The comparison is made in terms of throughput (Figure 4.6), and average delay (Figure 4.7). All traffic is treated as Best-Effort and packets are discarded only when requests fail to be served regardless of the buffer backlog, i.e. there are zero losses within the network and no packets are dropped within the nodes because of buffer overflow as lengths are assumed infinite. The frame length that maximizes the matching performance is the mean burst length times the number of switch ports [8]. As this figure can be very large sometimes, the chosen frame length is usually a compromise between throughput performance and maximum delay. For this study we have chosen a frame length of $F = 100$ slots, which provides a good performance with a reasonable delay below $1ms$ (1000 slots) for traffic loads under 80% (Figure 4.7). Figure 4.6 shows that the Frame-based algorithm yields a better performance even with just one WC array, reaching 90% throughput.

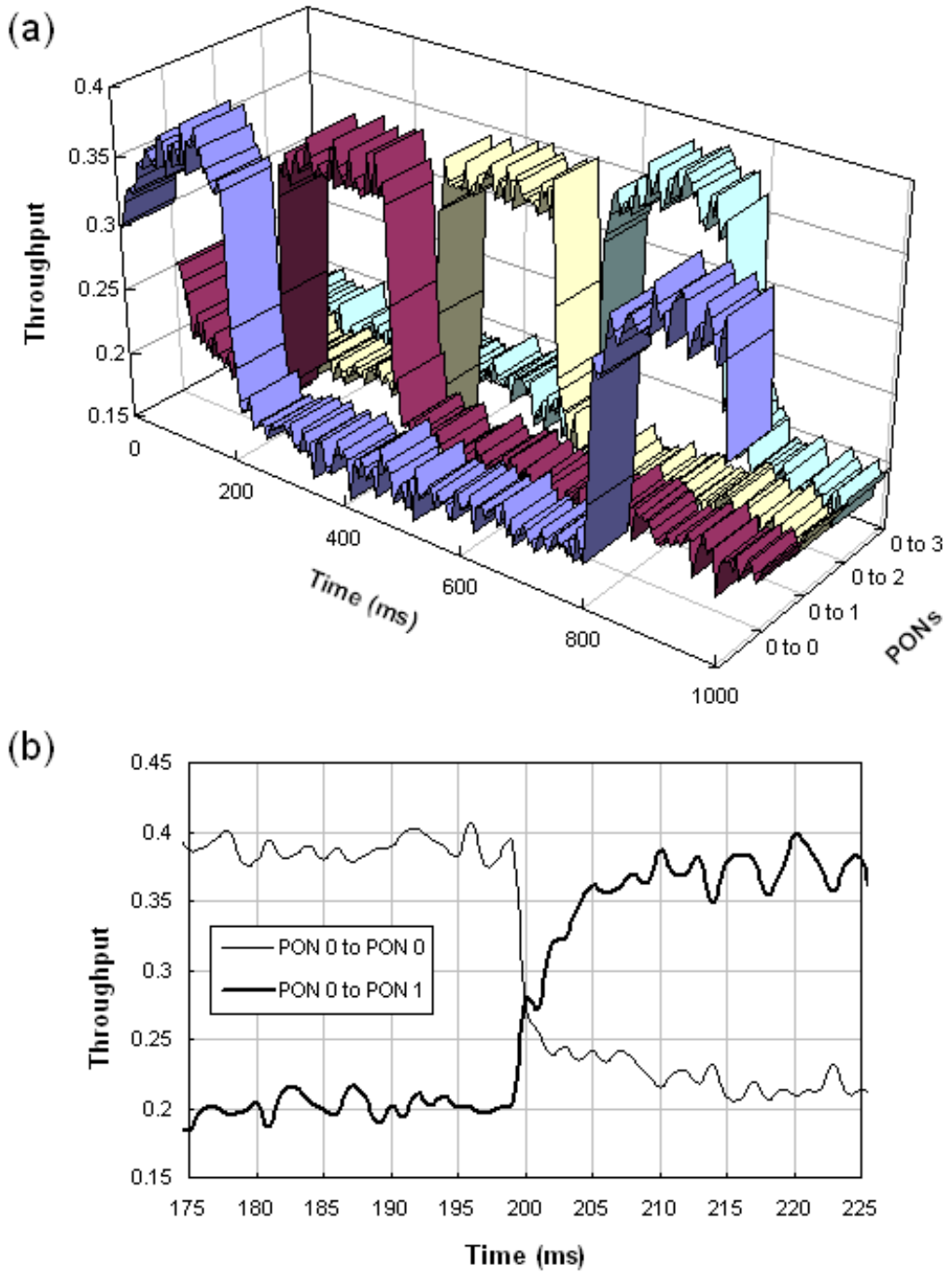

Figure 4.5: Throughput as a function of the offered load under dynamic diagonal traffic matrix

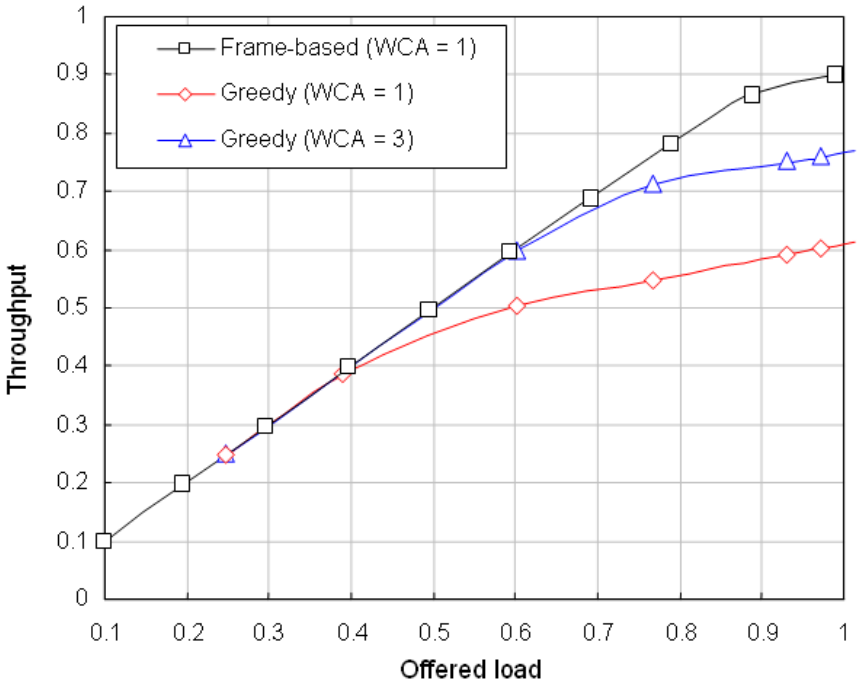

Figure 4.6: Throughput as a function of the offered load comparing the Greedy and the Frame-based algorithm

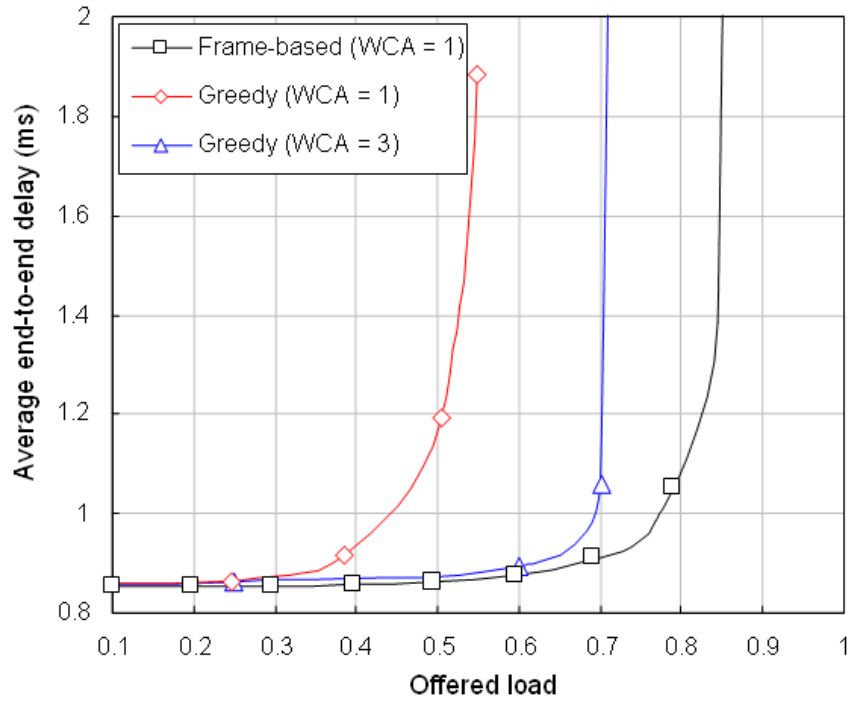

Figure 4.7: Average end-to-end delay as a function of the offered load comparing the Greedy and the Frame-based algorithm

4.4 Optimization

Two main problems can be identified from the performance evaluation section.

The first one regards the waiting time in the queuing stage of the node due to the Head of the Line (HoL) blocking [77]. The HoL blocking is typical of FIFO queues: a packet at the head of the FIFO queue that cannot be transmitted to avoid collisions or contentions on the PON may prevent a successful transmission of another packet following in the FIFO order. To avoid this undesirable situations, the optimized queueing stage is composed by several queues, one per each destination PON. In this way, the queue architecture is very similar to the VOQ (Virtual Output Queue) architecture used in Input Queued (IQ) switches [77], where, at each input port, packets are stored in separate queues on the basis of the destination port they should reach.

The second one regards the retransmission of the nodes' requests after having been received a negative allocation answer from the NC. The physical distance between nodes and NC adds important latency which lead to a drastic increase of the waiting time in the queues. To cope with this problem, the requests are pipelined to the NC, i.e., the node does not wait for the corresponding grant to arrive, but sends new requests in every frame. In such a way, if a node sends a request to the NC and receives a negative answer, it does not require to send it again during the next frame. The NC maintains information on the non-allocated requests and tries to schedule them in next frame together with the new incoming requests.

The optimized solution has been compared with the original solution using the Frame-based algorithm and a network with $P = 4$ PONs, $n = 32$ nodes, $W = 32$ wavelengths, $WCA = 1$, and $F = 100$ slots per frame. The power-of-two traffic matrix is used for this study.

Figure 4.8 and Figure 4.9 plot the throughput and the maximum end-to-end delay, respectively, as a function of the offered load comparing the original and the optimized solution. It is evident that both measures indicate that the optimized solution achieves better results reaching very good performance.

4.5 QoS provisioning

4.5.1 Problem formulation

In Chapter 3 we discuss that we are interested to support 3 different services: guaranteed, priority and best-effort. Since work in [6] proposed a method to support guaranteed service and best-effort, we concentrate our contribution in proposing a method to support the priority service. Since the priority class only requires better treatment than best-effort without precluding the guaranteed service and/or increasing the control complexity, the mechanism to provide it must be as simple as possible. Following this directive, we design a distributed scheme that need some additional functionalities at the nodes while little changes are required for both the scheduling and the MAC protocol.

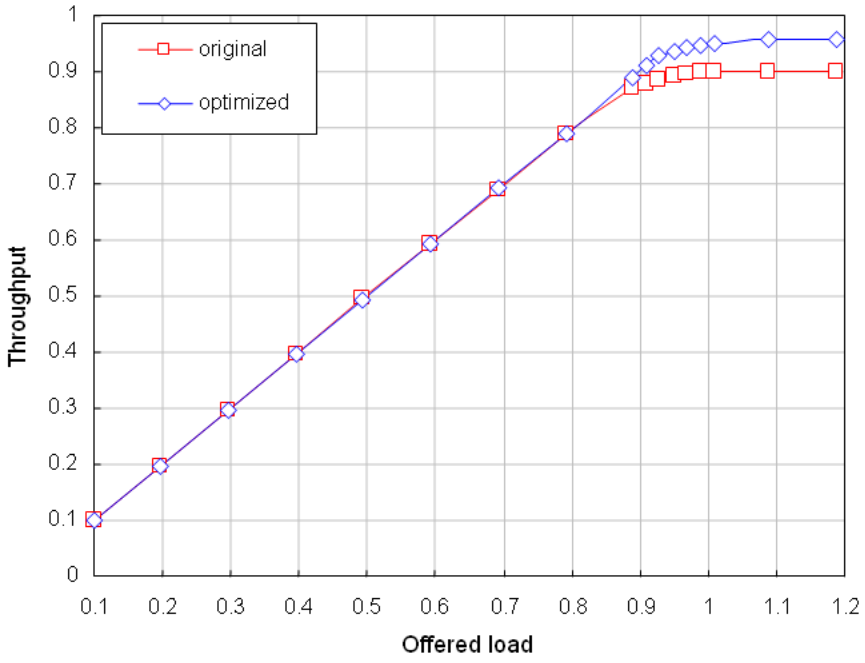

Figure 4.8: Throughput as a function of the offered load comparing the original and the optimized solution

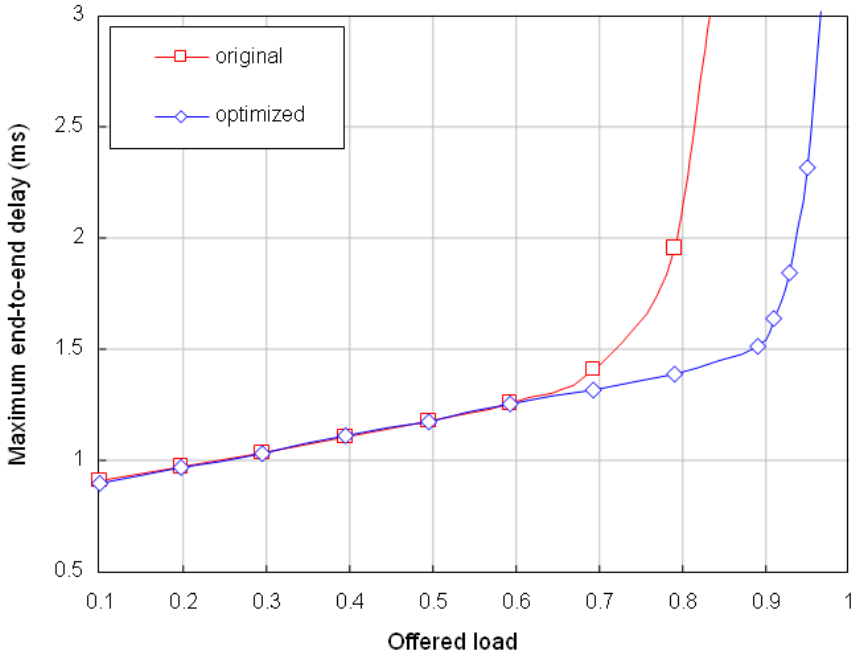

Figure 4.9: Maximum end-to-end delay as a function of the offered load comparing the original and the optimized solution

4.5.2 QoS strategy: the Limited Attempts (LA) technique

This optical packet network has no optical buffers and its switching functionality can be non-blocking for any traffic matrix with sufficient WC arrays. Therefore we assume that the scheduling algorithm along with the adopted QoS strategy are the only factors determining the maximum delay and the packet loss rate because of failing to serve the requests within the QoS constraints. While still yielding the maximum possible throughput, the scheduling algorithm and the nodes must ensure that the PLR required by the traffic class is achieved and that the maximum delay reaches acceptable levels. Hence we are not interested in controlling the delay or in controlling its variation -as opposed to other propositions [2]- as a QoS strategy. Two traffic classes are considered, namely a Best-Effort (BE) class as low priority traffic and a High-Priority (HP) class as high priority traffic; the latter offering lower PLR and limited maximum delay.

The novelty of the new QoS strategy presented in this work lies in the combination of two different mechanisms, distributed between the scheduling algorithm (matching + TSA) in the central node and more importantly the way switching requests are issued in the nodes. The multi-class-adapted, scheduling algorithm works in the central node giving different priorities to different types of traffic. On the other hand, nodes give switching requests further opportunities in subsequent frames when they fail to be served in the current frame being scheduled; the number of attempts depending on the type of traffic. The number of attempts is pre-determined (h for BE and k for HP traffic) and when requests fail to be served after (h, k) times respectively they are dropped regardless of the buffer backlog. Varying the pair (h, k) is a compromise between achieving low losses and limiting the maximum delay. We call this approach the Limited Attempts (LA) technique.

Various possibilities exist for applying the scheduling algorithm to different classes of traffic. Here we present two methods for handling HP and LP traffic:

1. The first method tries to maximize the overall throughput simultaneously for both HP and BE traffic requests. HP traffic takes precedence over LP traffic in the granting and acceptance phases of the matching process, which is run only once for both types of traffic. We refer to this technique as Throughput Maximization (TM);

2. The second method maximizes the throughput of the HP traffic class. In this case, the HP requests are scheduled first, then the BE requests afterwards using the remaining resources. Hence the matching algorithm must be run twice and this is the 'traditional' technique [100]. We refer to this technique as High-Priority Maximization (HM).

4.5.3 Performance evaluation

Although we only show results regarding the Frame-based algorithms, the same QoS strategy can be applied to the Greedy algorithms as shown in [29].

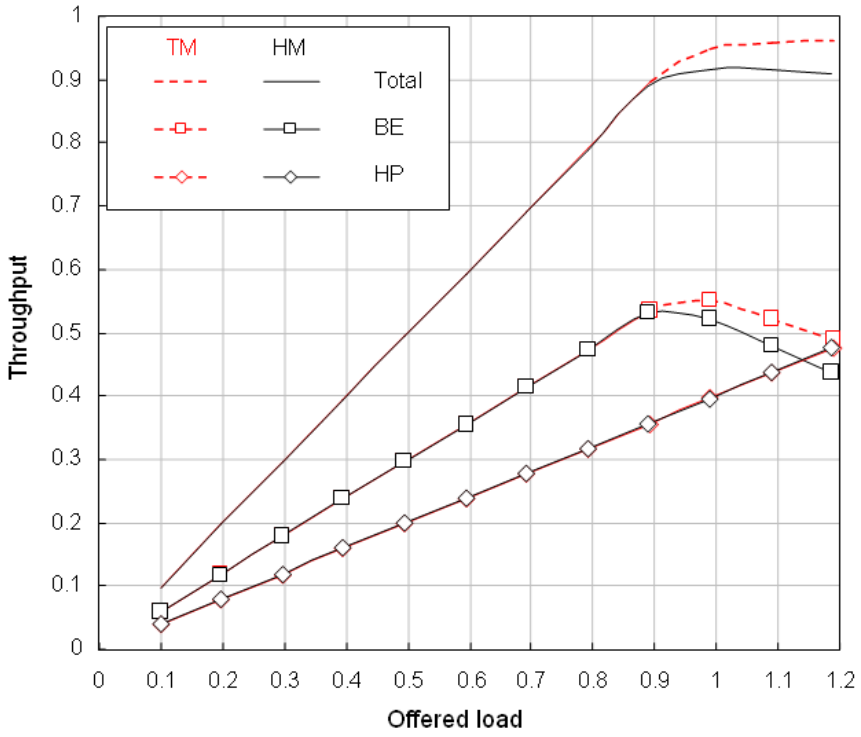

Figure 4.10: Throughput as a function of the offered load comparing TM and HM techniques and considering ($h = 3$, $k = 7$)

In next figures we consider a network with $P = 4$ PONs, $n = 32$ nodes, $W = 32$ wavelengths, and $WCA = 1$. The power-of-two traffic matrix is used for this study. Results shown look at a node pair within the PON belonging to the anti-diagonal links in power-of-two traffic matrix. As already stated, packets are lost only when requests fail to be served after (h, k) request attempts have been given.

In Figure 4.10–4.13 we assume a bullish scenario for high quality traffic; 40% of the traffic is HP and 60% is BE. This balance is changed in Figure 4.14 by varying the HP traffic percentage from non-existent to 100%, always at 100% load. Figure 4.10 compares the TM and the HM techniques considering ($h = 3$, $k = 7$), while Figure 4.11, Figure 4.12 and Figure 4.13 show curves for different values of (h, k) using the TM technique. Figure 4.14 shows the throughput attained by each traffic type and overall traffic when the percentage of HP traffic changes.

Although very similar, Figure 4.10 shows that at very high loads and at congestion levels, the overall throughput and BE throughput are higher using the TM method than the HM. HP traffic is served as requested with either of the two techniques at the expense of BE traffic when the overall demand cannot be satisfied and therefore the HP throughput is maintained by the multi-class matching technique. As we are interested in maximizing the overall throughput, and for the sake of brevity are only considering this result and not other issues such as delay, we have chosen the TM technique for the rest of the QoS strategy evaluation. Hence this choice is not intended to demonstrate a superiority of one technique over the other, as more detailed studies

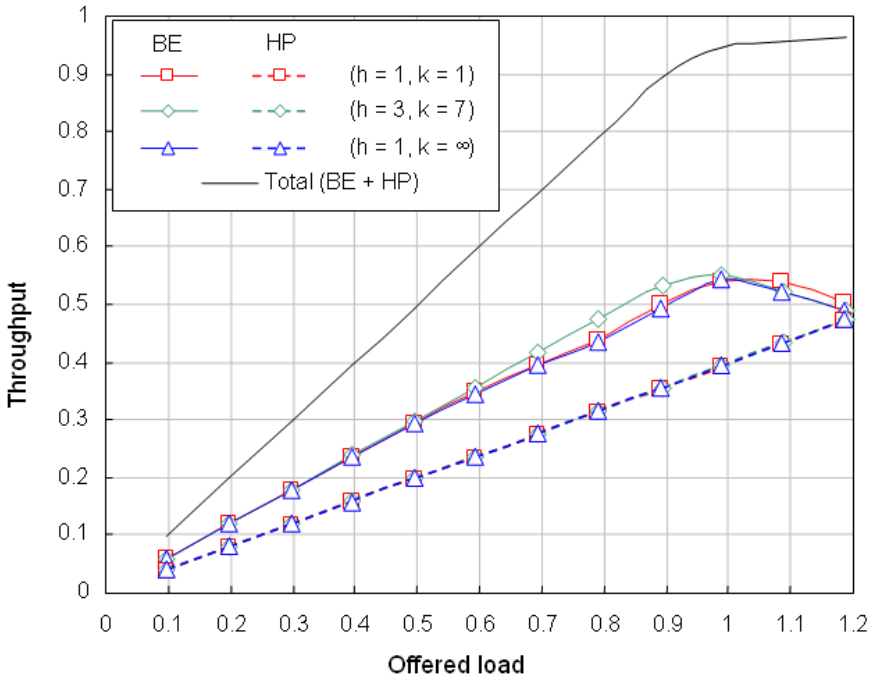

Figure 4.11: Throughput as a function of the offered load comparing different values of (h, k) and using the TM technique

would be necessary and most likely it would rather depend on particular scenarios and strategies.

Figure 4.11 shows the class-relative and total network throughput as a function of the offered load. The different tested (h, k) values have little effect on the network throughput. For low values of offered load, both BE and HP traffic grow proportionally to their level, and above 50% load differences are small. At congestion levels, i.e., for total loads higher than the network capacity, the amount of admitted BE traffic decreases to ensure the transmission of HP traffic, which means that this QoS strategy enforces priority to HP traffic. Losses are small and the graph resolution does not show them until very high loads are reached.

Figure 4.12 shows PLR as a function of offered load. For any value of h, HP traffic has the highest and the lowest losses for $k = 1$ and $k = \infty$ respectively. The losses for the latter case do not show up in the graph meaning that they are lower than 10^{-8} and therefore practically non-existent (they are not measurable within the simulation time). For the other simulated case, i.e. $k = 7$, losses are very small, always lower than 10^{-5}. Note that the results for the HP traffic may change for different values of h. On the other hand, BE losses for $h = 3$ are remarkably lower than for $h = 1$, almost two orders of magnitude at high loads. These results show the good performance of this QoS strategy. However, we do not only need a reasonable average delay, but also a bounded maximum delay within reasonable values.

Figure 4.13 shows the maximum end-to-end delay (worst case) as a function of offered load. Obviously, it is not possible to simulate all possibilities, and therefore

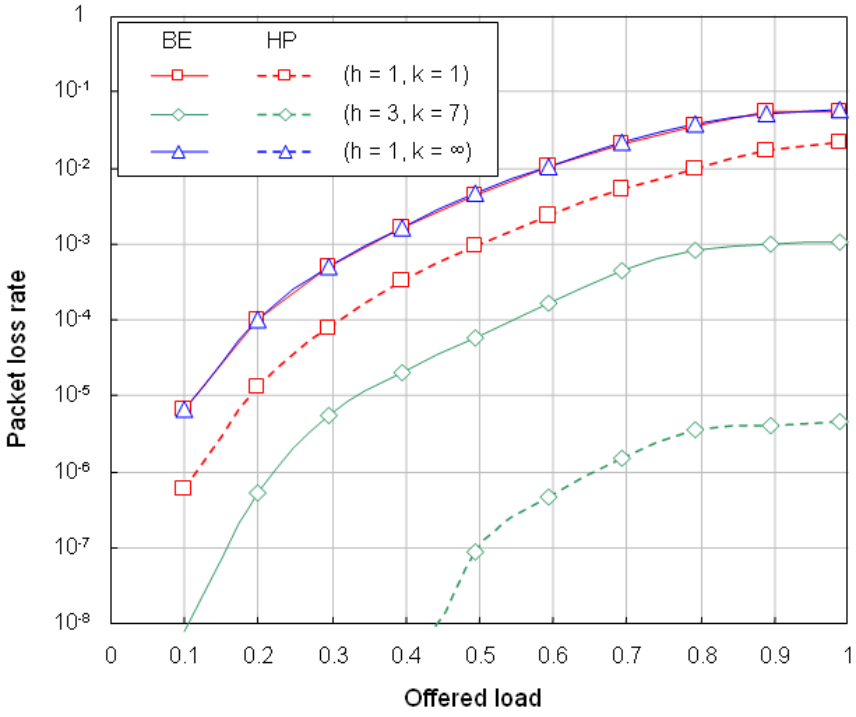

Figure 4.12: Packet loss rate as a function of the offered load comparing different values of (h, k) and using the TM technique

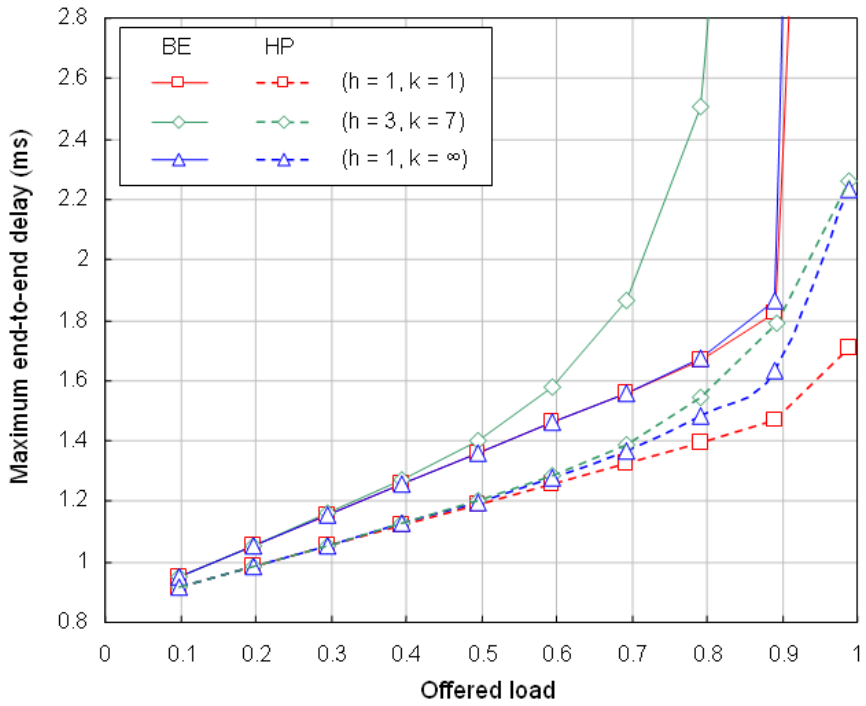

Figure 4.13: Maximum end-to-end delay as a function of the offered load comparing different values of (h, k) and using the TM technique

the obtained maximum delay must be understood as an approximation. Nonetheless, the high number of simulation runs brings confidence that the delays shown are good approximations and can be reasonably expected. Buffer lengths are assumed to be infinite and therefore packets are not dropped because of buffer overflow, but only when packets have used up all their opportunities to send switching requests. Although Figure 4.13 shows almost vertical curves, the delays are huge but still finite, which for the sake of clarity are not shown. Even though the (h, k) values are finite, and it would seem every packet eventually should be either successfully switched or dropped, in some cases the backlogged packets are unable to send the switching requests because of high traffic load, remaining in the infinite queue, and hence the possibility of huge delays. Undoubtedly, in a real case where the buffers are finite, once a maximum delay is reached packets that have not been able to send switching requests would be discarded and losses in Figure 4.12 for BE traffic would be higher at very high loads. If we were to use other variants of the frame-based matching algorithm these delays would be shorter so the PLR with finite buffers would come close to these results.

Loosely speaking the higher the (h, k) values the higher the delays, but particular values for each traffic type affect the delay of the other class. Although this study does not attempt to find the optimal values of (h, k), we see that h has a stronger effect on the delays than k. This is due to the higher priority of HP traffic true for whatever value is given to (h, k) and thus it is the BE traffic that suffers the most. Delay and PLR are mutually dependent and each one can only improve at the expense of the other. For example, with ($h = 1$, $k = 1$) at 100% traffic load BE and HP traffic experience a PLR of 6×10^{-2} and 2×10^{-2} respectively, and the maximum delay for HP traffic is 1.7 ms (1700 slots). However, whilst with ($h = 3$, $k = 7$) BE and HP traffic experience lower PLR of 10^{-3} and 4.5×10^{-5} respectively, the maximum delay for HP traffic has now increased to 2.26 ms (2260 slots).

Finally, Figure 4.14 shows the throughput as a function of the relative percentage of HP traffic, always at 100% total offered load (HP+BE). The dotted line represents the relative bandwidth not used by HP traffic left for BE traffic. The achieved throughput for HP traffic perfectly matches the relative load percentage increase, until it reaches a value of 95% at 100% relative percentage of HP traffic. This result shows the robustness of the QoS strategy in terms of throughput, ensuring the service of the HP traffic with low losses, up to very high loads.

4.5.4 Comparison with other QoS techniques

Other similar QoS methods are present in the literature. For instance Absolute Priority (AP), and more effective Random Early Discard (RED) mechanism [47] are used to provide QoS applying selective packets dropping techniques. Nonetheless, our proposal is more appropriate for the considered environment. If a node sends a request to the NC and receives a negative answer, it does not require to send it again during the next available frame. Indeed, the NC maintains information on the non-allocated requests and tries to schedule them in each frame together with the new requests. The NC also maintains information on the number of attempts

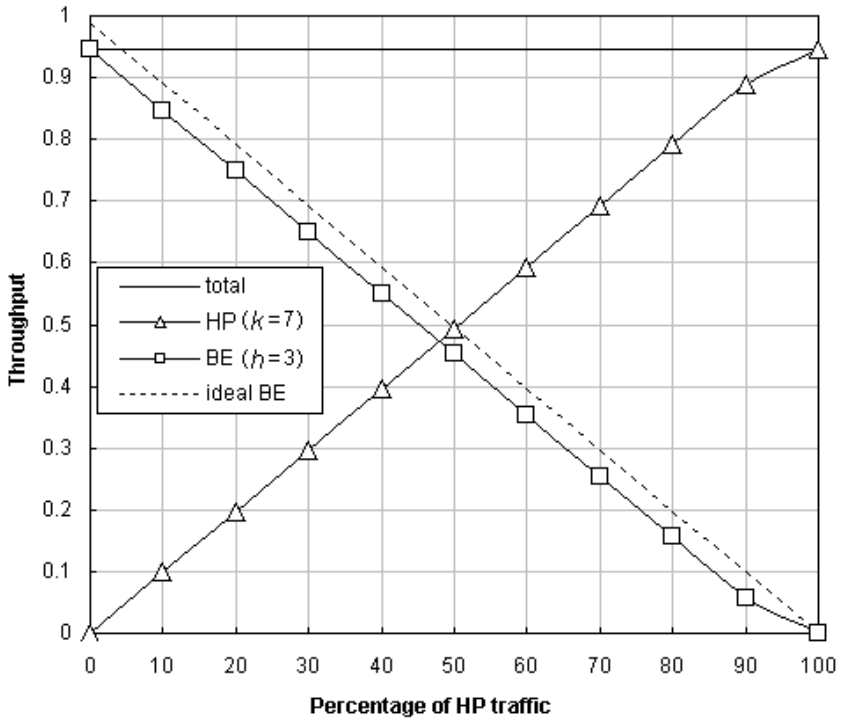

Figure 4.14: Throughput as a function of HP traffic relative load percentage at 100% total load using the TM technique

used per each request and removes those that already spent all opportunities. At the same time, the nodes drop the corresponding packets that receives too many negative answers. The same coordinate strategy is not possible to apply to neither AP nor RED mechanism and therefore the nodes must resend new request for any negative answer.

In Figure 4.15 and Figure 4.16, we compare our mechanism (LA in the figure) with the AP and RED mechanisms in terms of network throughput and maximum end-to-end delay, respectively. The network uses the Frame-based algorithm and consider 40% of HP traffic while the rest is BE traffic. We consider a network with $P = 4$ PONs, $n = 32$ nodes, $W = 32$ wavelengths, and $WCA = 1$. The traffic matrix is the power-of-two matrix while the traffic model is self-similar.

The figures prove the advantageous of our proposal. While the AP presents the worse results in both comparisons, RED and LA show similar PLR but LA outperforms the RED in terms of delay.

4.6 Summary

The performance of the multi-PON architecture has been evaluated considering several simulation scenarios. The performance results have been obtained using a real scale simulator including self-similar traffic model and different traffic patterns between interconnected PONs.

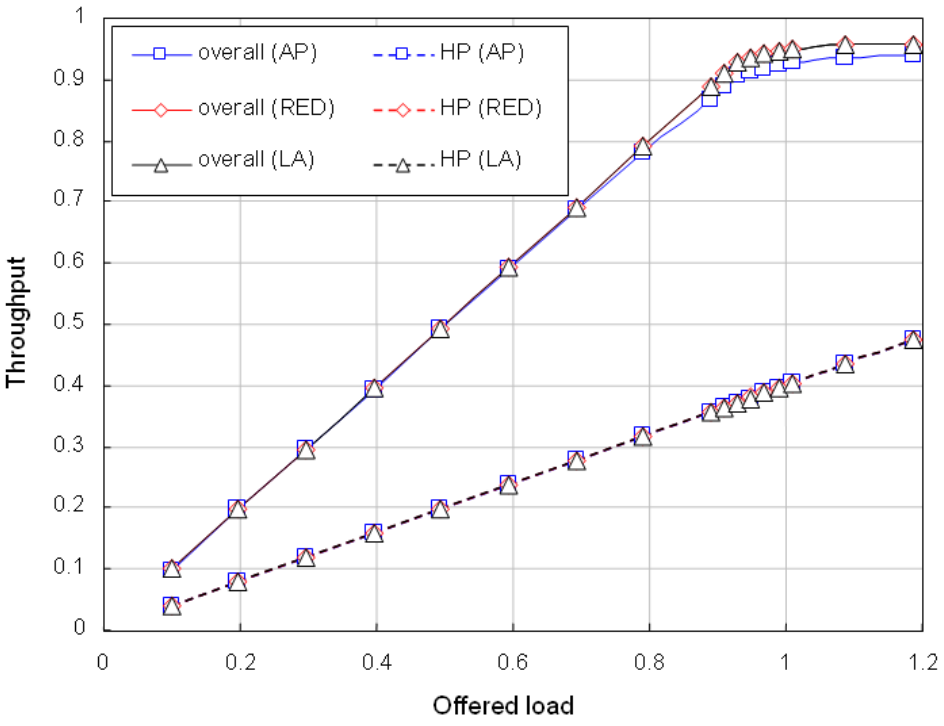

Figure 4.15: Throughput as a function of the offered load comparing the AP, RED and LA

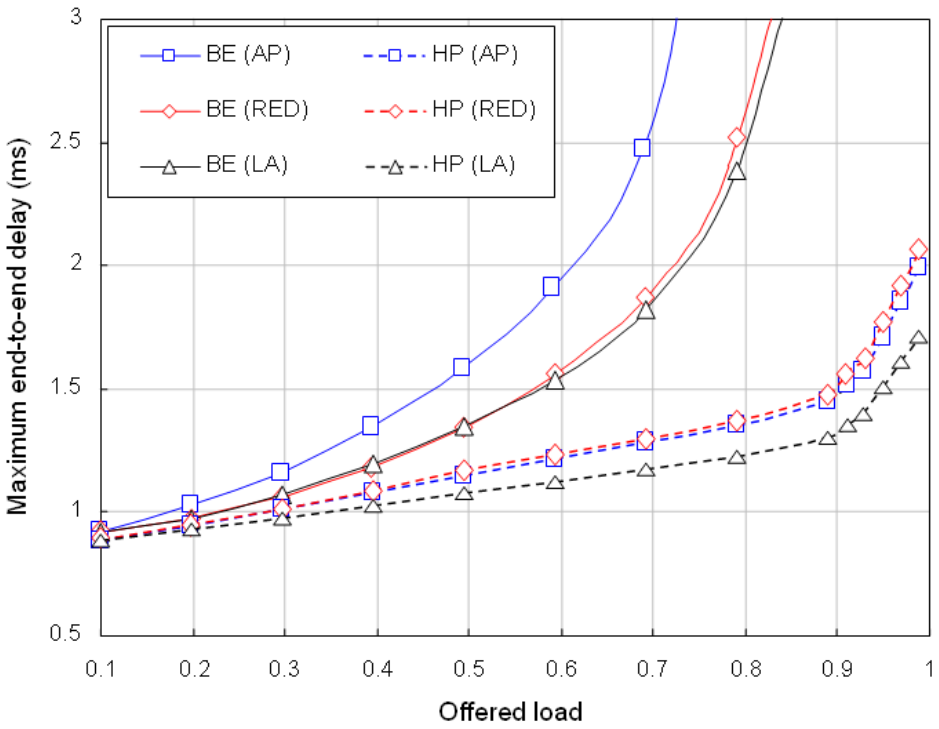

Figure 4.16: Maximum end-to-end delay as a function of the offered load comparing the AP, RED and LA

Two main weaknesses have been identified and overtaken by optimized mechanisms. The validity of the proposals have been demonstrated, indeed both throughput and maximum end-to-end delay measures achieve better results than the original proposal.

A QoS strategy for two classes of traffic has been also suggested. This strategy achieves a clear differentiation between the 2 traffic classes, each complying with their requirements in a very robust way, i.e., with no effect on the network throughput. We have also tested two different techniques for handling the two classes of traffic when using a frame-based matching algorithm. One technique runs the algorithm on a per traffic class basis, i.e. sequentially running the matching algorithm for each traffic class. The other technique runs the matching algorithm once, giving precedence to the high priority traffic in the granting and acceptance phases. Results in terms of overall throughput, tested for a particular traffic balance between two types of traffic, are better when running a single instance of the algorithm. However, neither technique affected the performance of the high priority traffic. The single instance technique was used to compare performance results for different numbers of opportunities in different frames given to each type of traffic.

The proposed novel QoS strategy shows very good performance for the two considered classes, namely Best-Effort and High-Quality traffic. The High-Quality class results show low Packet Loss Probability and bounded maximum delay, whilst acceptable levels are also achieved for the Best-Effort class. These results, obtained for a particular traffic balance, have also been shown, in terms of throughput, to be valid for other traffic balances where the percentage of High Quality traffic was varied from nil to 100%.

Chapter 5

Multi-ring architecture

5.1 State-of-the-art

The multi-ring architecture consists of a number of unidirectional slotted WDM rings of metropolitan dimensions, which collect traffic from several ring nodes. The WDM rings are interconnected to other rings via the Hub, and to a packet switch in the core. The rings can be either physically disjoint, or be obtained by partitioning the optical bandwidth into disjoint portions. The use of a Hub node that is in control of the resources makes the multi-ring different from other optical ring networks like e.g., the HORNET (and without any limiting relation between node counts and the number of wavelength). The Hub node is used to forward optical packets between ring networks, as well as to interconnect the metro area to the backbone through an electronic Gateway. The Hub is an SOA-based optical packet switch capable to cope with a very high level of traffic (Terabit/s). The lack of real optical memories is compensated through the use of an extended MAC protocol. The optical Hub is configured by a controller which exploits the control channels of each connected ring network, in order to calculate the switching permutation. The details of the technological implementation of the Hub architecture are described in [43].

Nodes are composed of an electronic part and an optical part. The electronic part realizes the adaptation with client layers, and packet buffering (electronically). At the optical level, two node architectures are proposed in the DAVID project to propose a progressive introduction of optical packet technologies.

Targeting a short/medium term approach, a first proposal was made to limit the use of advanced optical technologies and use commercial and mature ones instead. Based on passive structures as described in [73], the architecture uses optical couplers and off-line optical filters to minimize physical issues when cascading the nodes (Figure 5.1(a). This passive structure is lacking the packet drop stage, so that nodes keep all packets on the ring, simply copying packets addressed to them. We call this solution *passive multi-ring* (PMR). To allow simultaneous add and drop operation within the same slot, up- and downstream channels are spectrally separated. Nodes access the ring using separate sets of wavelengths: W for transmission, and W for reception. Data are sent by nodes on transmission wavelengths and switched from

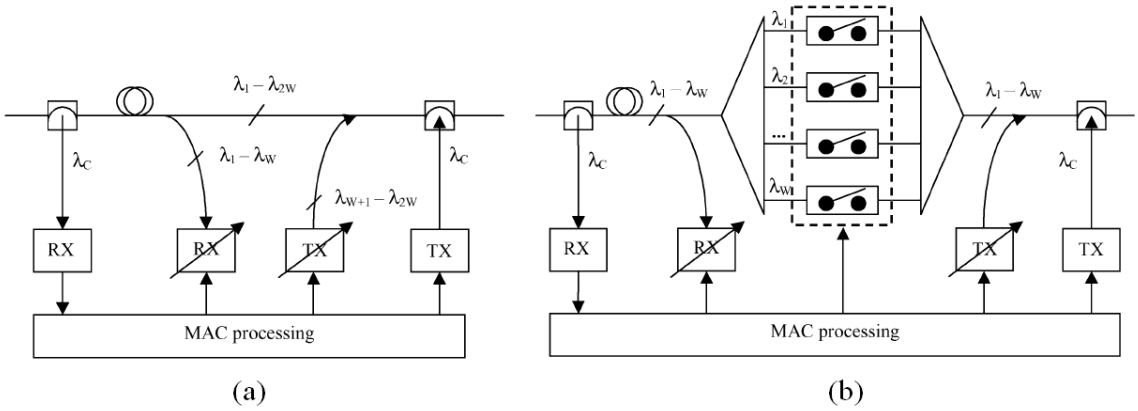

Figure 5.1: Architectures of (a) PMR node with transmission and reception decoupling and (b) MR node with erasure capability

transmission wavelengths to reception wavelengths at the Hub, which must provide wavelength conversion. Packets are received by nodes from reception wavelengths and are dropped when they reach the Hub, which, in turn, generates empty transmission channels for downstream nodes.

A second node structure is also considered as a longer-term approach allowing this erasing capability (Figure 5.1(b) which allows packet removal at the destination, hence wavelength reuse (i.e., packets only circulate along ring spans between source and destination nodes). We call this solution *active multi-ring* or simply multi-ring (MR). In this case, nodes use the same set of W wavelengths for both transmission and reception.

In the following sections we only focus on the active multi-ring architecture. We use the passive multi-ring architecture as a reference in Section 5.5.4 and in the benchmarking analysis of Chapter 6.

5.1.1 MAC protocol

The rings are shared media, requiring a MAC protocol to arbitrate access to its slots, in order to regulate both time and wavelength dimensions. The overall system works as a combined wavelength/time/space distributed multiplexer. Contention and collision is avoided by an allocation algorithm and intelligent operation in the ring nodes using the control channel. Only the control channel is converted to the electrical domain for processing at each ring node, while the bulk of user information remains in the optical domain until its final destination in the end ring. The control slot information includes the state (empty or used) of the data slots, and the destination address of the corresponding data packets.

The inlet-outlet Hub allocation algorithm works as follows: a measurement cycle is defined (the length of which is denoted by F), during which the Hub monitors the use of the slots allocated to any ring pair. The monitoring of ring-to-ring traffic can be based either upon measurements of the load on the different rings at the Hub,

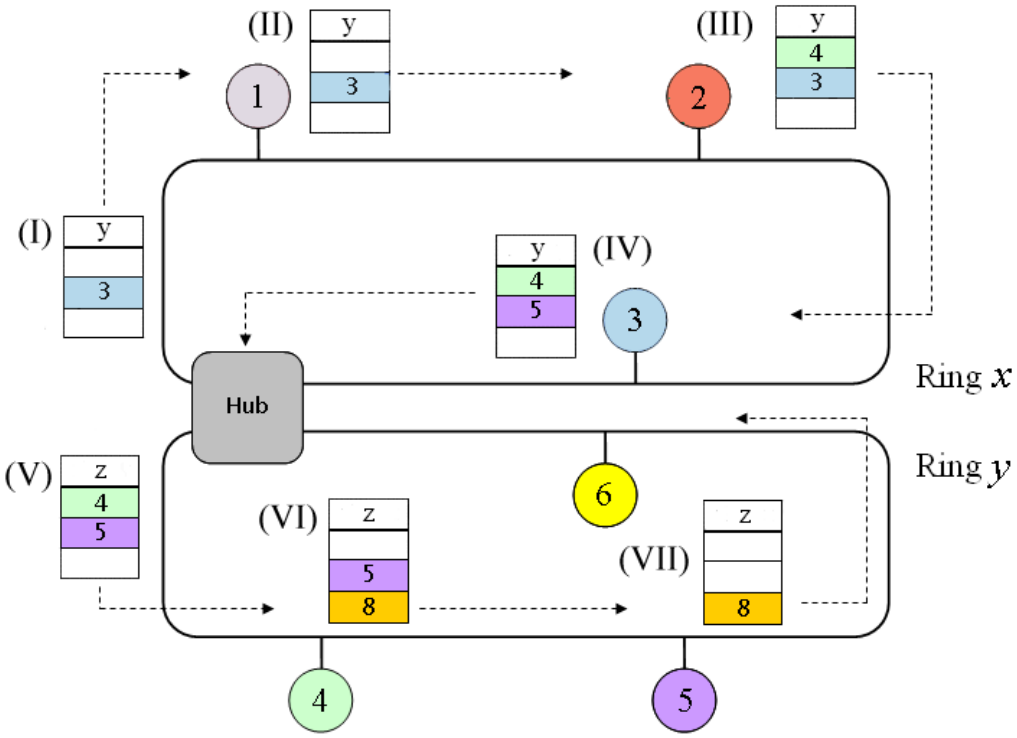

Figure 5.2: Example of multi-slot forwarding in the multi-ring. Colors in slot represent packet destinations

or upon explicit reservations issued by ring nodes. At the end of the measurement cycle, the Hub issues a new set of switching permutations, to be used for the coming measurement cycle. The Hub acts as a nonblocking switch that is reconfigured in every time slot and can exploit wavelength conversion to solve contention. In every time slot, the Hub operates a permutations from input rings to output rings. This permutation is the same for all wavelength of each ring and is known for each time slot in each ring: each multi-slot is labeled by the Hub with the identity of the ring which packets transmitted in the multi-slot will be forwarded by the Hub.

Given this behavior, each multi-slot traverses a sequence of ring, e.g., as illustrated in Figure 5.2. Nodes of ring x transmit data to be received by nodes of ring y (steps II to IV). Ring x can be viewed as the "upstream" ring, where transmission occur, while ring y can be viewed as the "downstream" ring, where receptions occur. When the considered multi-slot traverses the downstream ring y (steps V to VII), it gathers transmissions for the next ring z.

5.1.2 Scheduling algorithm

The computation of the sequence of permutations operated by the Hub is a scheduling problem as shown in Figure 5.3. The Hub scheduler is driven by a ring-to-ring request matrix $\mathbf{R}$, each element $\mathbf{R}_{i,j}$ contains the number of multi-slots that must be transmitted from input ring i to output ring j. The request matrix $\mathbf{R}$ is decomposed

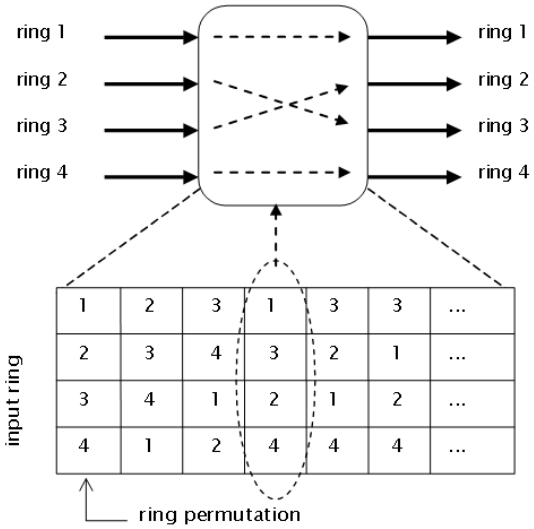

Figure 5.3: Scheduling at the Hub

into F switching matrices $\mathbf{P}$ through iterated application of the critical maximum matching algorithm [63].

5.1.3 Traffic measurements

The request matrix $\mathbf{R}$ is computed at the end of each measurement cycle F, as the sum of 3 contributions:

$$\mathbf{R} = \lceil \mathbf{SM} + \beta\mathbf{IC} + \gamma\mathbf{EC} \rceil$$

with β and γ positive constants, where:

- **SM** is a measure of the average number of multi-slots transmitted among ring pairs during each OW. To smooth out measurement errors, SM is passed through an exponential filter. This is an absolute throughput measure.
- **IC** is the percentage of filled slots which indicates potential congestion situation. This is a relative throughput measure.
- **EC** takes into account explicit congestion signals sent by nodes. This congestion signals are triggered at nodes by SAT: if a node receive the SAT and the length of its queue exceeds the quota Q, the node sends a congestion signal to the Hub.

In order to perform traffic measurement at the Hub, the slot reuse capability (spatial reuse) is not exploited; all traffic is forced to pass through the Hub before being removed from the ring.

5.1.4 Fairness control

The empty-slot operation can exhibit fairness problem under unbalanced traffic. This is particulary true in the ring topology, in which, as already mentioned, upstream nodes have generally better access chances than downstream nodes.

Credit-based schemes is used to enforce throughput fairness. A control signal called SAT is circulated in store-and-forward mode from node to node along ring; a node forwarding the SAT is granted a transmission quota Q and can transmit up to Q packets before the next SAT reception. When a node receives the SAT, it immediately forward the SAT to the next node on the ring if it is satisfied (hence the name SAT), i.e., if no packets are waiting for transmission or if Q packets were transmitted since the previous SAT reception. If the node is not satisfied, the SAT is kept at the node until one of the two conditions above is met.

5.1.5 QoS provisioning

A QoS strategy providing 2 priority classes (i.e., 2 asynchronous services) is proposed in [2]. This is a connection-less approach, therefore it does not guaranteed neither the delay nor the bandwidth. The nodes requiring to send a first priority class packet but no free slot is found, mark a busy slot using an additional flag available in the control channel. This reservation is not meant for the use of the specific node which did the marking but it is at the disposal of any node with first priority packet to be sent. Since the marking is done on already busy slot it is likely that the slot will travel around the rings. In order to not proliferate the marks inefficiently, every node keeps track of its action so that it also removes a reservation mark for every first priority packet sent in an unmarked empty slot. The second priority packets can only use unmarked empty slots.

5.2 Contributions

Our contributions include: (i) the performance evaluation of these architectures and the identification of the drawbacks and of the open issues, (ii) the optimization of the proposed architectures and MAC protocols, and finally (iii) the proposal of different QoS mechanisms to support guaranteed and best-effort services.

5.2.1 Simulation scenario

The performances of the proposed mechanisms are evaluated in order to assess their merits. The simulation results presented in the following sections have been obtained by means of an ad-hoc event-driven simulator reproducing a real scale configuration of the multi-ring network. The parameters of the network are:

- R indicates the number of rings;
- n indicates the number of nodes per ring;

- W indicates the number of wavelengths per fiber;
- B_w indicates the bit-rate. It is set to 10 Gbit/s in every simulation scenario;
- P_s indicates the duration of the time-slot. It is set to 1 μs;
- F indicates the number of time-slots per frame;
- RTT indicates the ring round trip time;
- L indicates the length of the ring;
- ρ indicates the offered load;
- $\mathbf{M}$ indicates the ring-to-ring traffic matrix, whose generic element $\mathbf{M}_{i,j}$ is a real number ranging between 0 and 1 representing the percentage of traffic coming from input ring i and going to output ring j with respect to ρ. Four different traffic matrix are defined named: *uniform* $\mathbf{M}^U$, *diagonal-x* $\mathbf{M}^{Dx}$, *power-of-ten* $\mathbf{M}^P$, and *very unbalanced* $\mathbf{M}^V$. For the case of $R = 4$ and $x = 7$, the matrices are as follows:

$$M^U = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \qquad M^{D7} = \frac{1}{10} \begin{bmatrix} 7 & 1 & 1 & 1 \\ 1 & 7 & 1 & 1 \\ 1 & 1 & 7 & 1 \\ 1 & 1 & 1 & 7 \end{bmatrix}$$

$$M^P = \frac{1}{1111} \begin{bmatrix} 1 & 10 & 10^2 & 10^3 \\ 10 & 10^2 & 10^3 & 1 \\ 10^2 & 10^3 & 1 & 10 \\ 10^3 & 1 & 10 & 10^2 \end{bmatrix} \quad M^V = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{10} & \frac{1}{3} & \frac{1}{15} \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 \end{bmatrix}$$

  Other traffic matrices used to evaluate specific behavior will be directly introduced in the evaluation description.

The electrical queues at nodes are considered infinite.

Two types of input traffic were considered for the best-effort traffic: the classical model with Poisson distribution of the interarrival times (*Poisson model*) and a self-similar process (*Self-similar model*) implemented as a superposition of 16 strictly alternating independent and identically distributed ON/OFF sources. The duration of each ON/OFF period was assumed to be a random variable with a Pareto distribution with shape $\alpha = 1.2$, which leads to a Hurst parameter of $H = 0.9$ [103]. All packets have the same size and fit in one slot.

Instead, guaranteed traffic is connection-oriented; each node generates connection requests occupying one slot per frame. Both the connection interarrival time and the connection duration are geometrically distributed.

The mean value of the interarrival times for best-effort and guaranteed traffic is selected accordingly to generate the required offered load ρ.

The number of simulated packets is chosen big enough to reach steady-state results and a 95% confidence interval is calculated.

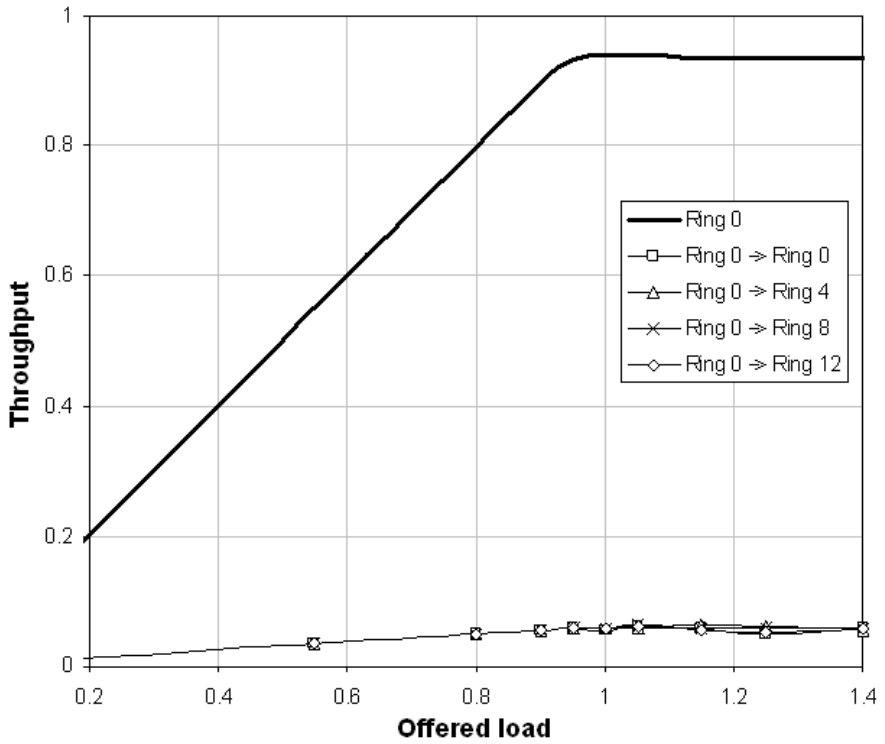

Figure 5.4: Throughput as a function of the offered load under uniform traffic matrix

To evaluate the performance of the multi-ring network, we use both absolute and relative throughput measures. The former calculated as the ratio between used and available slots, the latter as the ratio between used and required slots.

5.3 Performance evaluation

The main issues to address regarding the performance of the multi-ring metro network are the achievable throughput and the fairness. The more the throughput is close to one (the ideal value) the better the MAC protocol. This is not enough because we also want the MAC protocol to share as evenly as possible the bandwidth between the nodes.

In Figure 5.4, Figure 5.5, and Figure 5.6 we set up a network with $R = 16$ rings, $n = 10$ nodes, $W = 4$ wavelengths, $Q = 500$ packets, $L = 100$ km which means that a slots needs $RTT = 0.5$ ms (500 slots) to circulate around a ring, and $F = 10000$ slots per frame. The traffic model is the Poisson one.

Figure 5.4 shows the throughput per destination ring 0, 4, 8 and 12 on ring 0, and the overall throughput of ring 0 (solid line) as a function of the offered load under uniform traffic matrix. Although we report the throughput for some rings, the same behavior holds for all other rings due to the traffic symmetries. We can observe that the throughput increases with the offered load until it reaches the saturation. The total network utilization is close to 0.935 and each destination ring is treated fairly.

In Figure 5.5, the throughput per node is plotted against the number of nodes

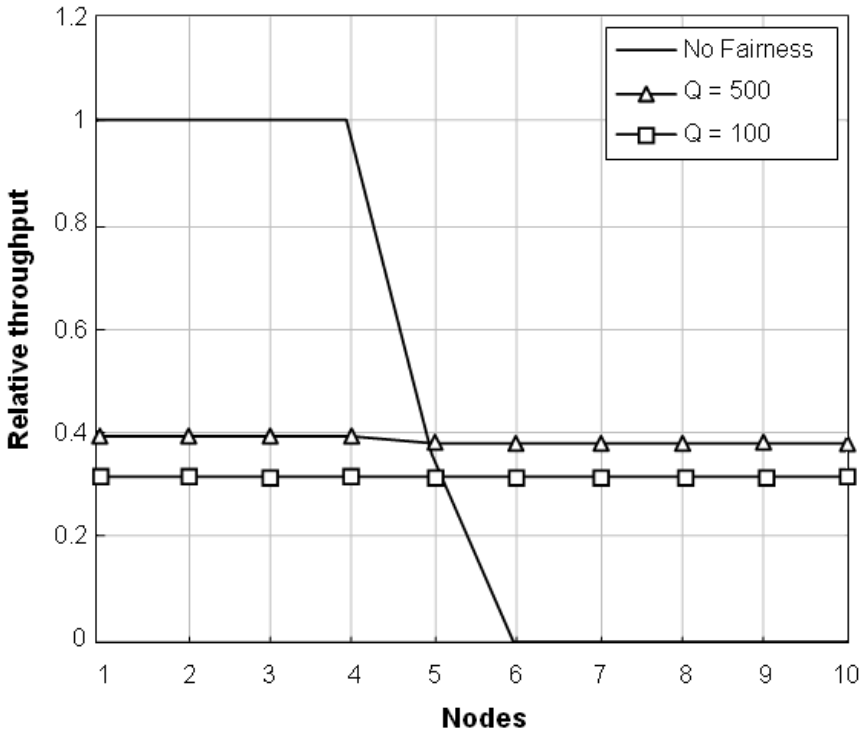

Figure 5.5: Relative throughput per node for total load on the ring of 0.7, without fairness control (solid line) and with SAT for two values of Q

active on the ring. Two cases have been simulated with the SAT quota set at $Q = 500$ and $Q = 100$. The traffic is uniform with a load of 0.7 per ring, and node unbalanced: all nodes send packets only to node 0 of ring 0, except node 0 of ring 0 that sends packets uniformly to the rest of nodes of ring 0. As expected without fairness control (solid line), the bandwidth utilization is unfair: upstream nodes (node 1, 2, 3, 4, and partially the node 5) use all empty slots (the relative throughput is one) and the downstream nodes (node 6, 7, 8, 9, and 10) cannot transmit (the used bandwidth is zero). By introducing the fairness control the bandwidth utilization becomes fair both for the case $Q = 500$ (triangle markers) and $Q = 100$ (square markers).

The fairness problem for inter-ring communications is addressed in Figure 5.6. Here, the diagonal-3 traffic matrix is considered, where about 30% of the traffic is intra-ring while the remaining 70% is evenly spread among the remaining rings, and per-ring uniformly distributed among the nodes. We can see that for low values of the offered load (not congesting the network), the throughput is proportional to the traffic matrix weights. For higher values of the offered load, the rings are treated according to a max-min fairness criteria: the intra-ring throughput decreases to give a fair bandwidth portion to the inter-ring connections.

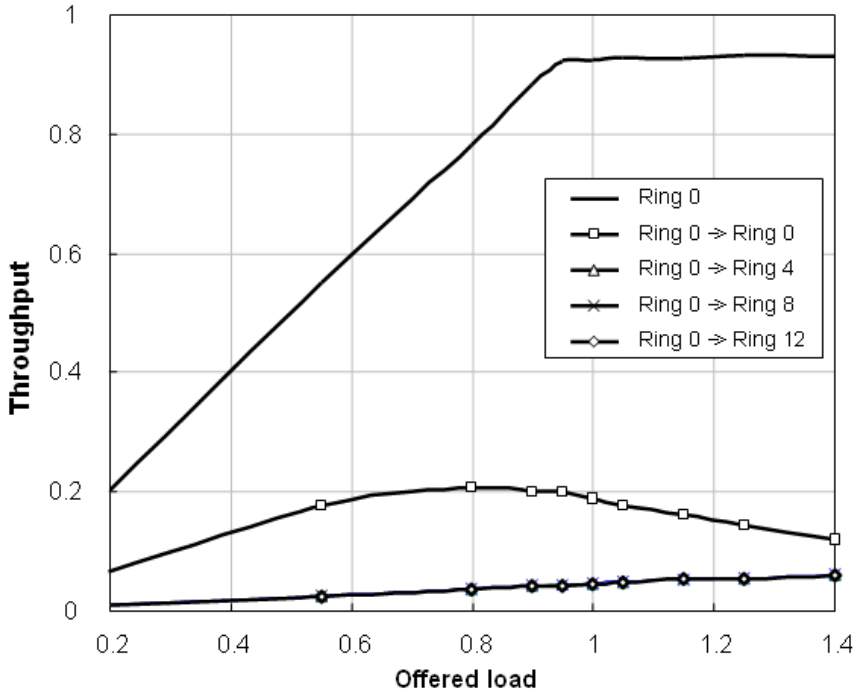

Figure 5.6: Throughput as a function of the offered load under diagonal-3 traffic matrix

5.4 Optimization

From the performance evaluation section, we identify two main weaknesses.

The first one regards the impossibility of exploiting the spatial reuse capability which is one of the main advantageous of the ring topology. Indeed, in the original proposal, the Hub needs to measure every transmitted packet to estimate the traffic request matrix and generate the ring-to-ring permutations.

To perform traffic measurement in presence of spatial reuse, we introduce an additional field in the control channel which we call SR. If any node receives a slot previously transmitted by a node on the same ring before it passes through the Hub, the node marks the SR field. So then, the Hub has to react according to the following three possible actions: 1) when the Hub receives empty both the slot and SR, the Hub count 0 slots; 2) when the Hub receives an empty slot but SR is marked, the Hub measures one slot; and 3) when the Hub receives a full slot, the Hub measures one slot independently of the status of SR.

The benefits of exploiting the spatial reuse in the MAC protocol is studied in Figure 5.7 which shows the throughput with (dashed line) and without (solid line) exploiting the spatial reuse for a metro with $R = 16$ rings and $W = 4$ wavelengths (Figure 5.7(a), and with $R = 4$ and $W = 16$ (Figure 5.7(b); the diagonal-3 and diagonal-7 traffic matrices are used respectively. For both figures, the interarrival traffic follows the Poisson model. The rest of parameters is set as in the previous study.

As expected, the higher the percentage of intra-ring traffic the more the gain of performance due to the spatial reuse. This is more evident in small networks (low number of rings). This result is quite obvious if we consider that the spatial reuse is exploited only in presence of intra-ring traffic.

The second weakness regards the notification of the explicit congestion signals. In the original MAC protocol, the SAT token has been used for providing two functions: 1) controls the fair access by limiting the number of transmitted slots, and 2) triggers the notification of the congestion situation from nodes to the Hub. Under particular traffic conditions, the Hub may get the urgencies provided by SATs with considerable delay.

In order to overcome this problem, a new mechanism is introduced in such a way that multi-slots operate as triggers for notifying the explicit congestion signals from nodes to the Hub independently of the SATs. Hence, the Hub will take into account the signals to generate the new set of ring permutations.

As an illustrative example, in Figure 5.8 we study the capacity of the network to overtake congestion situation comparing the original and the optimized solution. The parameter of the network are $R = 4$, $W = 16$, and $n = 10$. The rest of parameters remains unchanged with respect to the previous study. The Self-similar traffic model and diagonal-7 traffic matrix are used. To enforce the congestion, at time 700000 slots, the first and second columns of the traffic matrix are interchanged which simulates a drastic traffic fluctuation. The offered load is set to $\rho = 0.85$.

Both figures show a transient behavior due to the time required for adapting the network to the new traffic situation. The original proposal lasts around 200000 time-slots to return to a stable condition, while it is only 100000 time-slots for the optimized solution. This result indicates that the modification really improve the capacity of the network to solve congestion situations.

5.5 QoS provisioning

5.5.1 Problem formulation

In Chapter 3 we discuss that we are interested to support 3 different services: guaranteed, priority and best-effort. Since work in [2] proposed a method to support priority and best-effort services, we concentrate our contribution in proposing a method to support the guaranteed service.

We hence consider a guaranteed service (GS) traffic class with guaranteed bandwidth, and a best effort (BE) traffic class. The management of BE traffic is a relative easy task as has been studied in [9] and briefly summarized in Section 5.1. The Hub can schedule the permutations according to both traffic measurements and congestion signals issued by nodes allowing the nodes to decide whether to transmit and/or to receive by checking on the control slot of the multi-slot what has been already transmitted by upstream nodes. On the other hand, a connection-oriented approach is necessary to guarantee the GS traffic requirements, where the Hub establishes connections between the nodes reserving the required resources along the rings.

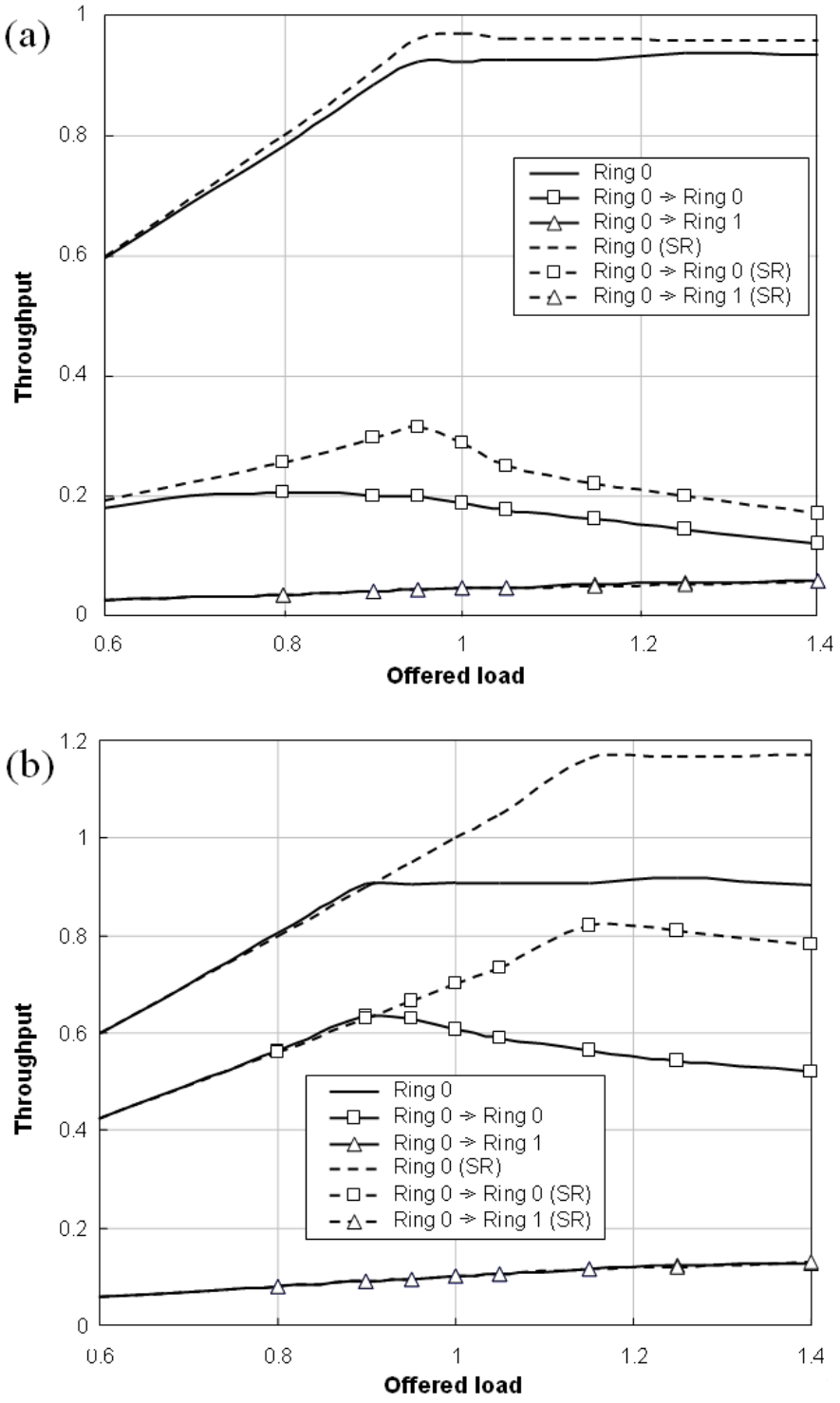

Figure 5.7: Throughput as a function of the offered load under diagonal traffic matrix with spatial reuse (dashed line) and without spatial reuse (solid line). (a) Network with 16 rings and 4 wavelengths per ring, (b) Network with 4 rings and 16 wavelengths per ring

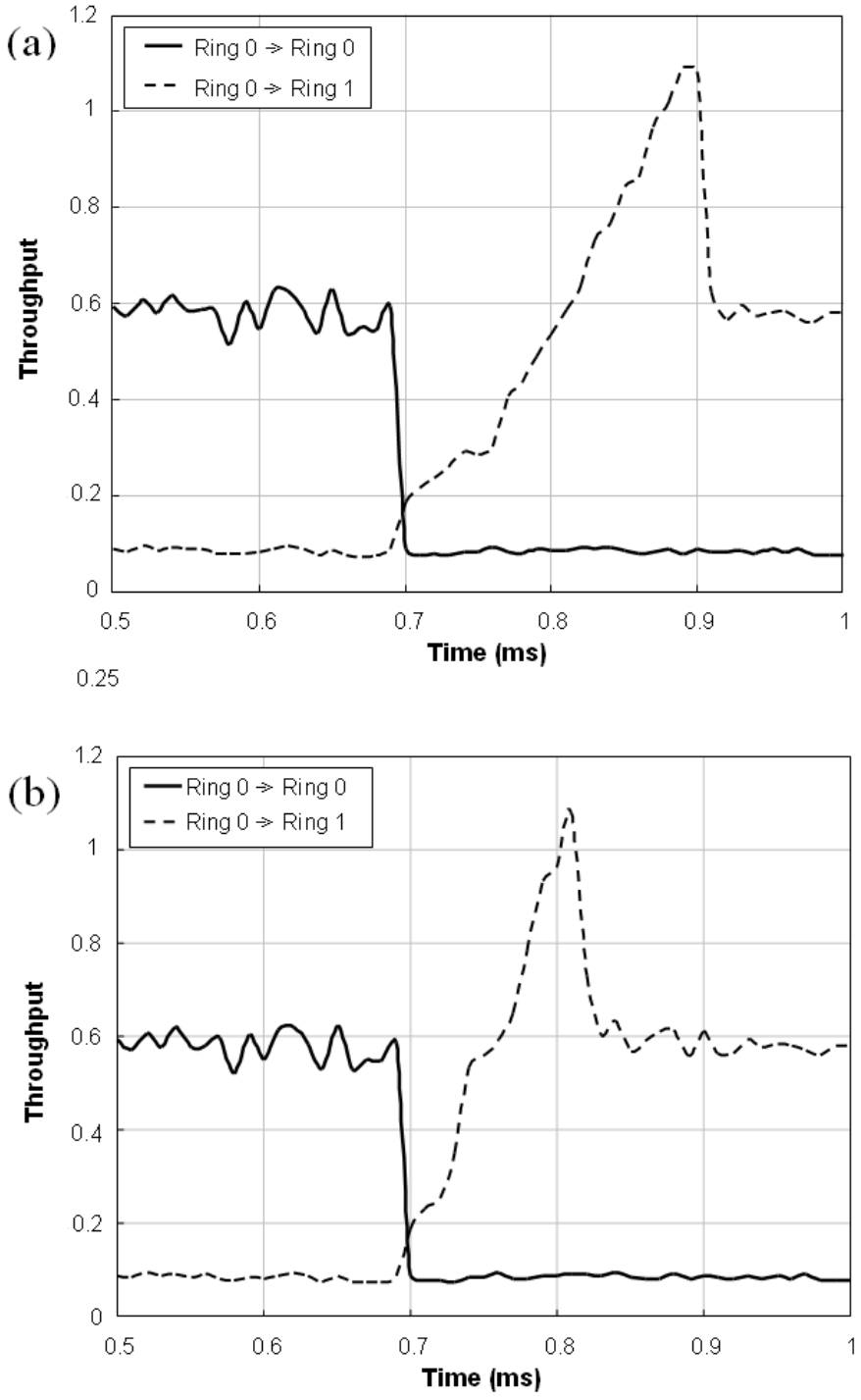

Figure 5.8: Throughput as a function of the offered load under diagonal-7 traffic matrix with traffic fluctuation comparing (a) the original and (b) the optimized solution

Therefore, the scheduling algorithm performs two tasks: (1) maintaining the virtual, time-slotted connections between the nodes, and (2) fairly sharing the unreserved bandwidth to the BE traffic (traffic subject to fairness control).

To set up the virtual connection, the nodes send explicit GS reservation requests to the Hub which collects them in a node-to-node connection request matrix **A**. At the same time, the Hub estimates the BE traffic requirements by measuring the BE traffic load and calculates a ring-to-ring matrix **B***.

The scheduling is based on a fixed-size frame of length F slots, which is considered the most suited solution for a system with guaranteed bandwidth allocation. Indeed, with a fixed-size frame, a reservation issued in terms of bit rate can easily be translated into an equivalent number of slots per frame. The frame length F must be chosen trading the allocation granularity (asking for longer frames) for the access delay and scheduling complexity (asking for shorter frames). The selection of optimal values for F is outside the scope of this thesis, but we typically envision several thousand slots in the frame.

Therefore, the problem of scheduling multi-class traffic can be stated as follows:

- GIVEN
  the node-to-node GS connection request matrix **A**
  the ring-to-ring BE request matrix **B***
- FIND
  a conflict-free slot allocation within a frame of length F
- SUCH THAT
  the number of allocated GS connection is maximized
  then, the number of transmitted BE packets is maximized

Satisfying the following constraints:

1. *Priority of GS traffic over BE traffic.* GS requests must be satisfied before serving BE traffic.
2. *Persistent allocation of GS connections.* The allocation of new GS and BE traffic must not affect currently established GS connections.
3. *Atomic allocation of GS requests.* GS requests are accepted only when they can be fully satisfied; otherwise, they must be refused. Atomic allocation typically makes sense when requests correspond to single real-time user connections, whereas it does not apply to elastic BE traffic, nor to sources that multiplex several data flows.
4. *No contentions.* Since each node is equipped with only one tunable data transceiver, it can transmit and receive at most one packet in each multi-slot.
5. *Avoid in-transit collision at the Hub.* A packet collision may occur at the Hub between the packets injected in the downstream ring (after having been switched at the Hub) and the slots reserved by the Hub in the upstream ring for the transmission of GS traffic.

The multi-class scheduling problem is NP-hard because it is generalization of the well-known knapsack optimization problem, which is NP-hard [83]. In fact, GS traffic requests can be considered as a set of objects of different sizes that must be fit in a knapsack of capacity F. The problem can be solved in polynomial time if we adopt heuristic solutions, which decrease the complexity of the scheduling accepting some degree of worse performance.

5.5.2 Heuristic solution

Given the node-to-node matrix **A** which contains the GS connection requests and the ring-to-ring matrix **B*** which contains the BE traffic load measurements, the heuristic approach is an incremental algorithm, which consists of three steps:

1. At first, all the slots that were allocated in the previous frame to BE traffic, as well as those corresponding to ended GS connections are released, so that only persistent GS connections remain allocated.

2. New GS requests are scheduled scanning the resources in a round-robin way. This steps ends when either all requests have been satisfied or all slots in the frame have been considered.

3. The remaining slots are used to allocate BE traffic contained in the matrix **B*** which is scheduled independently of GS traffic through an iterated critical maximum size matching [83]. The set of ring-to-ring permutations obtained must be fit in the unused part of the frame by selecting the permutations that allocate the maximum number of slots.

The complexity of this algorithm is $O(N^2F)$.

This algorithm satisfies all the constraints described above except the in-transit collision. Different methods can be used to fulfill this requirement.

For example, the Hub can take into account nodes' positions along the rings, while computing the scheduling and can inform all the nodes about the GS reservation before issues the new set of permutation matrices. In such a solution, which we call *Full Shared* (FS), the in-transit collision constraint can be satisfied since the nodes know the slots reserved for the other nodes.

The problem can be made easier by separating transmissions and receptions either in frequency or in time. The first approach leads to the *Frequency Decoupling* (FD) solution where the nodes use $\frac{W}{2}$ wavelengths to transmit, and the rest to receive packets. This method precludes the spatial reuse capability since all packets must be switched at the Hub before being received. In the second approach, dubbed *Time Decoupling* (TD), the same set of wavelengths is shared in the time domain and used only one RTT (Round-Trip Time) out of two to transmit the packets. In this case, the spatial reuse capability is available since the node can receive in every moment. For both solutions, the Hub can schedule the matrices without knowing the nodes' position along the rings.

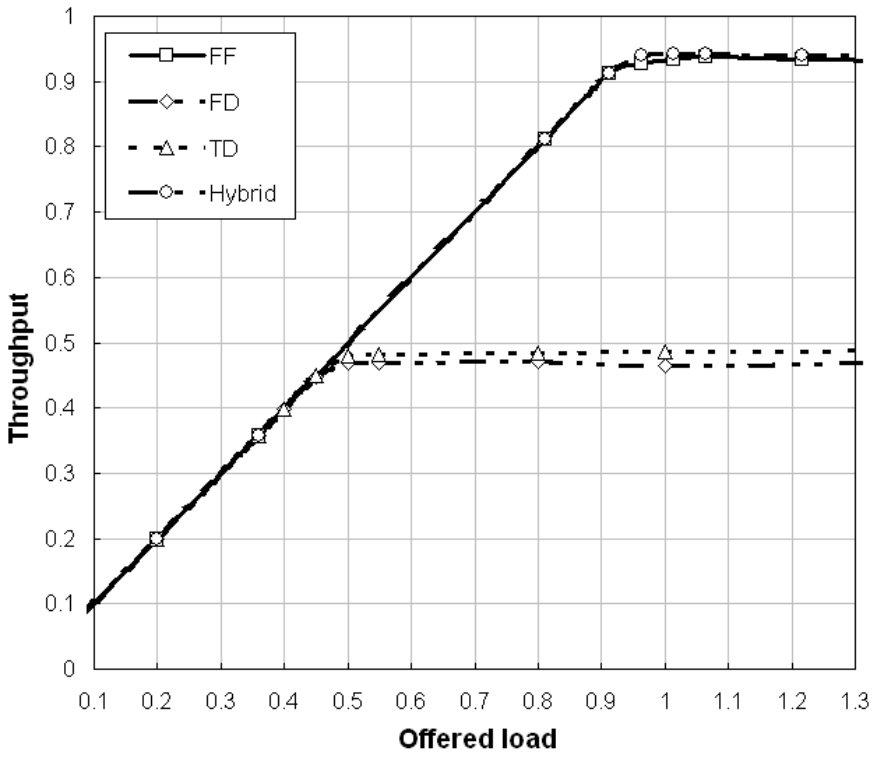

Figure 5.9: Throughput as a function of the offered load under uniform traffic matrix. GS load is fixed to 30%

Nonetheless, both FD and TD have the drawback that the available bandwidth is halved by the fixed resource partitioning. This problem can be overtaken if the decoupling is only applied to the GS traffic, and not to the BE traffic. In this case, the Hub can schedule the matrices without knowing the nodes' position along the rings but must inform the nodes about the GS reservation as in the FS approach.

5.5.3 Performance evaluation

The simulated network consists of $R = 4$ rings and $n = 10$ nodes per ring, each ring with $W = 4$ wavelengths (plus 1 for the control channel). The length of the ring is $L = 100$ km and the quota for the BE traffic is $Q = 500$. The GS traffic is not subject of fairness control. The frame is $F = 10000$ slots. The traffic is the self-similar model.

Figure 5.9 shows the total (BE+GS) throughput of the four solutions as a function of offered load (with GS = 30% of total load) considering uniform traffic distribution. The Hybrid solution obtains the best performance.

Figure 5.10 shows the throughput as a function of the HP relative load percentage assuming 100% total load and unbalanced traffic distribution; 70% of traffic is intra-ring (where the space reuse is possible), the rest is uniformly distributed among the inter-rings. The figure depicts the BE and total (BE+GS) throughput. The space reuse exploitation makes TD better than FD, while Hybrid still achieves the best results as far as GS traffic requires less than 50% of total resource.

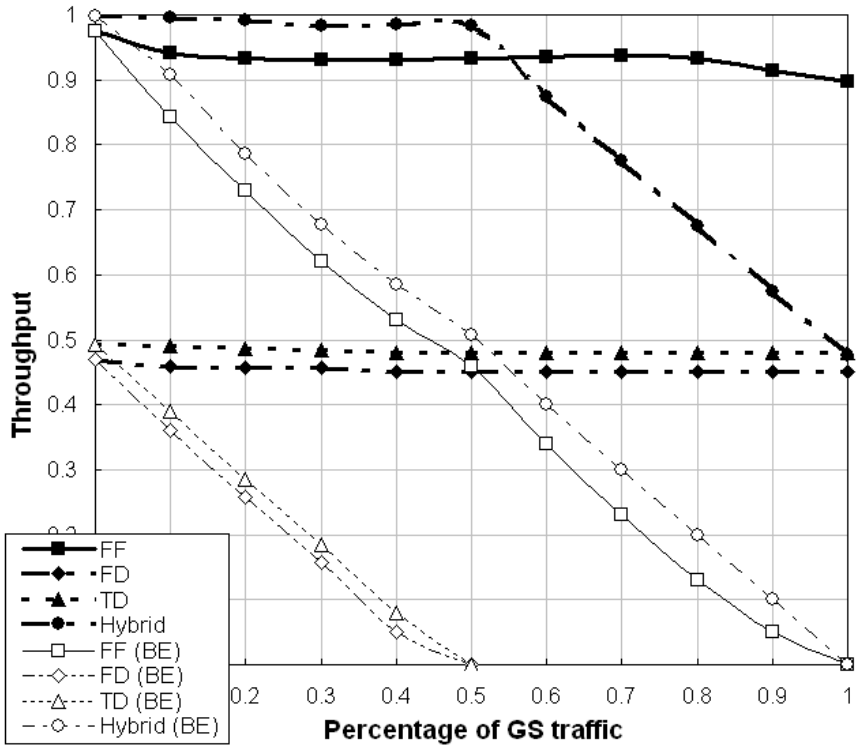

Figure 5.10: Throughput as a function of GS traffic relative load assuming 100% total load under diagonal traffic matrix

Table 5.1: Average running times for the four solutions

	FS	**FD**	**TD**	**Hybrid**
Time	2298 s	1354 s	1298 s	1711 s

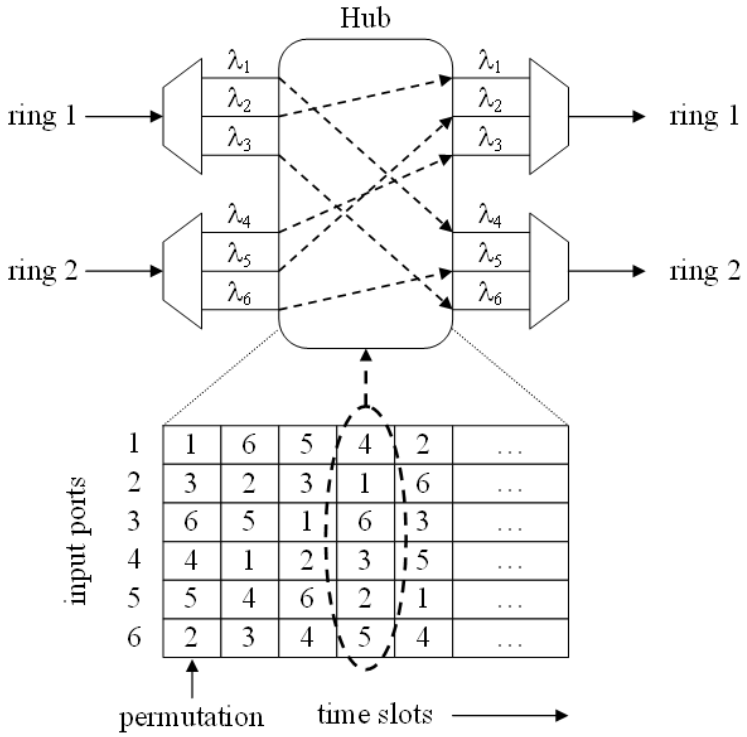

Figure 5.11: Scheduling wavelength-to-wavelength permutations at the Hub

Table 5.1 compares the complexity of the four solutions showing the average running time to obtain a point of the previous figures.

The simulation results show that the Hybrid solution is preferable when the GS traffic is less than 50% of total resources. TD and FD solutions can be adopted when the bandwidth on network links is not a bottleneck.

5.5.4 Optimization of the QoS mechanism

Scheduling can aim at different levels of performance guarantees. In general, an allocation of slots in the frame should provide average rates to node pairs in accordance with the traffic matrix. The scheme proposed in [9] for BE traffic allocates ring-to-ring rates at the Hub, and access decisions are decentralized at nodes, which, however, do not have guaranteed access. Advantages of that approach are the very small amount of information in the control channel and the good scalability properties.

However, if higher node-to-node guarantees are required, as it is the case for GS traffic, the scheduler must allocate slots to single node-to-node requests and perform wavelength-to-wavelength permutations among rings. The resource allocation problem becomes mostly centralized, and the amount of information in the control channel increases. It must be noted that a centralized scheduling is often desired by network operators, who may want to apply different control policies for different nodes (e.g., introducing an admission control policy for GS traffic requests).

Figure 5.11 shows a metro network comprising two rings conveying three wave-

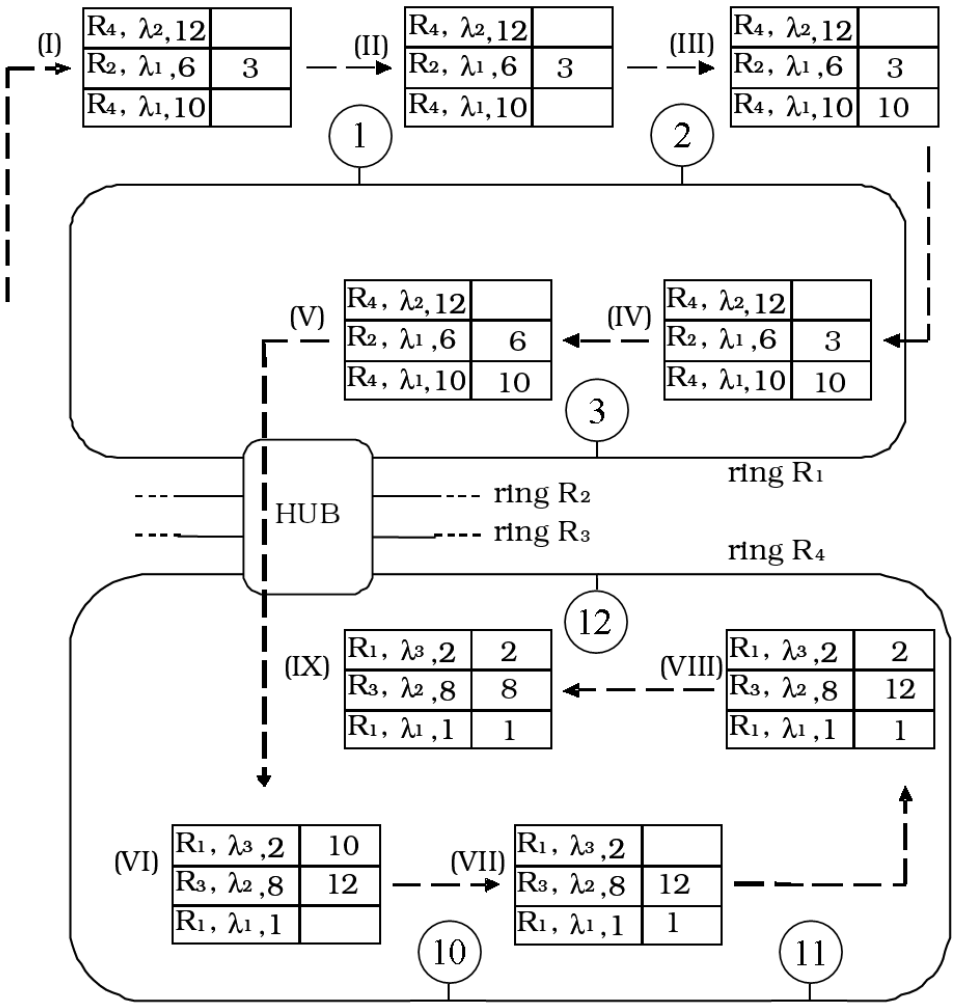

Figure 5.12: Example of multi-slot forwarding in the multi-ring network with wavelength-to-wavelength permutations

lengths each, where the Hub operates wavelength-to-wavelength permutations.

Figure 5.12 illustrates the operation of a network when the Hub performs wavelength-to-wavelength permutations. For simplicity, we consider a network with $R = 4$ rings, $W = 3$ wavelengths per ring, and $n = 3$ nodes per ring. Control slots are on the left-hand side of multi-slots, whereas data slots are on the right-hand side. Data slots are reserved for packets addressed to nodes identified in the triple (ring, wavelength, node) inserted by the scheduling algorithm running at the Hub in the corresponding control slot. Numbers in data slots denote the destination node to which the packets they convey are addressed. We describe the slot forwarding process in successive steps showing the path followed by a multi-slot as it travels from ring R_1 to ring R_4.

1. Control slots indicate that the first data slot is reserved for sending data to node 12 on ring R_4 using wavelength λ_2, the second data slot is reserved for a packet to be delivered on λ_1 to node 6 on R_2, and the third slot is booked for

transporting on λ_1 a packet addressed to node 10 on R_4. The second data slot is carrying a packet headed to node 3.

2. Node 1 leaves the multi-slot untouched.

3. Node 2 has a packet for node 10 on ring R_4 and inserts it in the corresponding data slot.

4. Node 3 drops the packet transported in the second data slot and reuses it to send a packet to node 6 on ring R_2.

5. The multi-slot reaches the Hub.

6. The Hub switches data slots to their intended output ring/wavelength as indicated by the triple contained in the relevant control slot. Therefore, the third slot containing a packet headed to node 10 is switched to wavelength λ_1 (first slot) of ring R_4, and a packet coming from another ring and with destination node 12 is inserted in the second slot. Finally, the scheduling algorithm inserts new destination triples in the control slots.

7. Node 10 drops the packet sent from node 2 and transmits a packet to node 1 using the third data slot.

8. Node 11 transmits a packet to node 2 on ring R_1 using the first data slot.

9. Node 12 drops the packet in the second slot, and transmits a new packet to node 8 on ring R_3.

In the presence of multi-class traffic, the Hub collects bandwidth requests issued by nodes for both GS and BE services, builds two node-to-node traffic request matrices (**A** and **B**), and schedules them in a sequence of wavelength-to-wavelength permutations. Both request matrices store the number of slots that must be transmitted from a node to any other node within a frame of F slots.

Now, the Hub knows all the packets that will be transmitted during the next frame, therefore it is easy to satisfied all the constraints described above using the following heuristic scheduling algorithm which is an adaptation of the previous one:

1. All the slots assigned to BE traffic and all the slots reserved for ended GS connections are released.

2. New GS connection requests are scheduled scanning the resources in a round-robin way. This step ends when either all requests have been satisfied or all slots in the frame have been considered.

3. BE requests are scheduled using wavelength-to-wavelength permutations scanning the resources in a round robin way taking into account the in-transit collisions.

The complexity of this algorithm is $O(N^2FW)$.

In order to evaluate the performance of the optimized solution, we also consider two frequency decoupling solutions proposed in [10] for the passive multi-ring node architecture. Indeed, the passive multi-ring well adopts the FD method since no erasure stage is required. These solutions comprise an optimum polynomial algorithm with no atomicity constraint of the GS connection requests and a faster greedy algorithm based on the same round-robin method adopted in our heuristic algorithm.

We study by simulation a network configuration comprising $R = 4$ rings and $n = 16$ nodes per ring, each node sharing $W = 4$ wavelengths. The latter means that for the multi-ring architecture each ring conveys 5 wavelengths (4 for data and 1 for control), whereas, for the passive multi-ring architecture, the wavelengths carried on each ring are 9 (4 for upstream traffic, 4 for downstream traffic, and 1 for control). The ring round-trip time to $RTT = 512$ μs, which means that the propagation delay on each ring is 512 times the slot duration, and the frame duration is $F = 10240$ slots (20 RTTs).

All figures consist of three plots showing the performance of the optimum algorithm for the passive configuration and the heuristics for both the passive and the active multi-ring architecture.

All plots show the throughput as a function of the amount of GS traffic present in the network, when the total BE offered load is exactly 1. In other words, when the GS traffic load on the horizontal axis of the figures is 0.2, the total network load is 1.2. Note, however, that both BE and GS traffic are distributed among different rings according to the chosen ring-to-ring matrix $\mathbf{M}^x$.

The plots in Figures 5.13-5.15 show the throughput for each destination ring on source ring 1 for GS traffic (white markers), the total GS throughput (black square markers), the total BE throughput (dashed line without markers), and the total throughput on ring 1 (solid line without markers). Although we plot the throughput for a single ring, the same behavior holds for all the other rings due to traffic symmetries.

Figure 5.13 compares the three solutions under uniform traffic. Figure 5.13(a) shows that GS throughput increases with the offered load until it reaches the value of 1 in overload. The overall throughput is constantly equal to 1, as it is always possible to fill with BE traffic the slots left free by GS connections. Figure 5.13(b) presents the same behavior as Figure 5.13(a) except when the offered load exceeds 1. In this case, the heuristics is not capable of maintaining the overall network throughput to 1.

In the active multi-ring configuration shown in Figure 5.13(c), the network behaves differently. Indeed, in this case, we have half wavelengths and more contention probability due to shared transmission and reception channels; therefore, when only BE traffic is present in the network the overall network throughput is 0.97. As the GS traffic increases, the network throughput increases as well, reaching values even higher than 1. Note that in this case, the amount of BE traffic never drops to 0. Indeed, since we measure throughput as the ratio between the number of transmitted packets and the number of available slots, it may happen that a slot can be used more than once to transport different packets during one single round trip. For instance, a node

can transmit a packet to a neighboring node on the same ring; the destination node can reuse the same slot to transmit another packet to a different node on the same ring, and so on. This is an implicit consequence of the wavelength reuse capability, and increases network throughput. Nevertheless, this gain can be exploited only for intra-ring traffic, i.e., when the transmitter and the receiver belong to the same ring. In any other case (i.e., for inter-ring traffic), packets must be switched at the Hub from their source ring to their destination ring and, therefore, slots containing such packets cannot be reused. For this reason, when traffic is mostly inter-ring, the gain obtained from wavelength reuse is lower: in Figure 5.13(c), it reaches 1.05, and in Figure 5.15(c), it is not even noticeable. In Figure 5.14(c), instead, it becomes more evident, reaching 1.17, because the diagonal traffic pattern has a higher percentage of intra-ring traffic than the other scenarios.

In Figure 5.14(a), obtained using the optimum algorithm for the passive multi-ring architecture, the overall throughput is always 1, and GS throughput proportionally increases with the offered load. It is interesting to note how intra-ring traffic, after reaching a value of about 0.7, starts to decrease and leaves resources to inter-ring traffic when the total GS traffic equals to the network capacity. This is due to the fact that the scheduler in overload tends to equalize the load on the rings, according to maxmin throughput fairness. As before, the heuristic algorithm in Figure 5.14(b) behaves similarly to the optimum one, except when the offered load exceeds 1. Moreover, the heuristics does not equalize rings' load.

In Figure 5.15(a), we observe the optimum behavior of the algorithm with the power-of-ten traffic pattern. All GS connections are allocated optimally and total throughput remains equal to 1. This is not true for both heuristic algorithms. The heuristics for the passive configuration in Figure 5.15(b) is not able to allocate all GS requests, reaching a throughput around 0.98 independent of the amount of GS traffic injected. The heuristics for the active configuration in Figure 5.15(c) performs even worse and the total network throughput does not rise above 0.9.

Finally, in Figure 5.16, we analyze the case of the very unbalanced traffic pattern: Since it is not symmetric, in each subplot, we show the throughput of GS traffic on each ring, the total network throughput of GS and BE traffic, and the overall network throughput. The offered load is normalized with respect to ring 2, where the traffic is higher. This means that when GS traffic load in the horizontal axis of Figure 5.16 is equal to 1, ring 1 is 50% loaded, ring 2 is 100% loaded, and rings 3 and 4 are 33.3% loaded; therefore, the total network load is about 0.54. Figure 5.16(a) shows that while the allocation of GS traffic remains very close to the optimum, the very unbalanced traffic pattern causes total traffic allocation to be suboptimal. This is due to the non optimality of a double matching with two traffic classes. Indeed, when only one class is present in the network, the total throughput is always equal to 100%. Furthermore, heuristic algorithms in Figure 5.16(b) and (c) are not able to allocate all traffic.

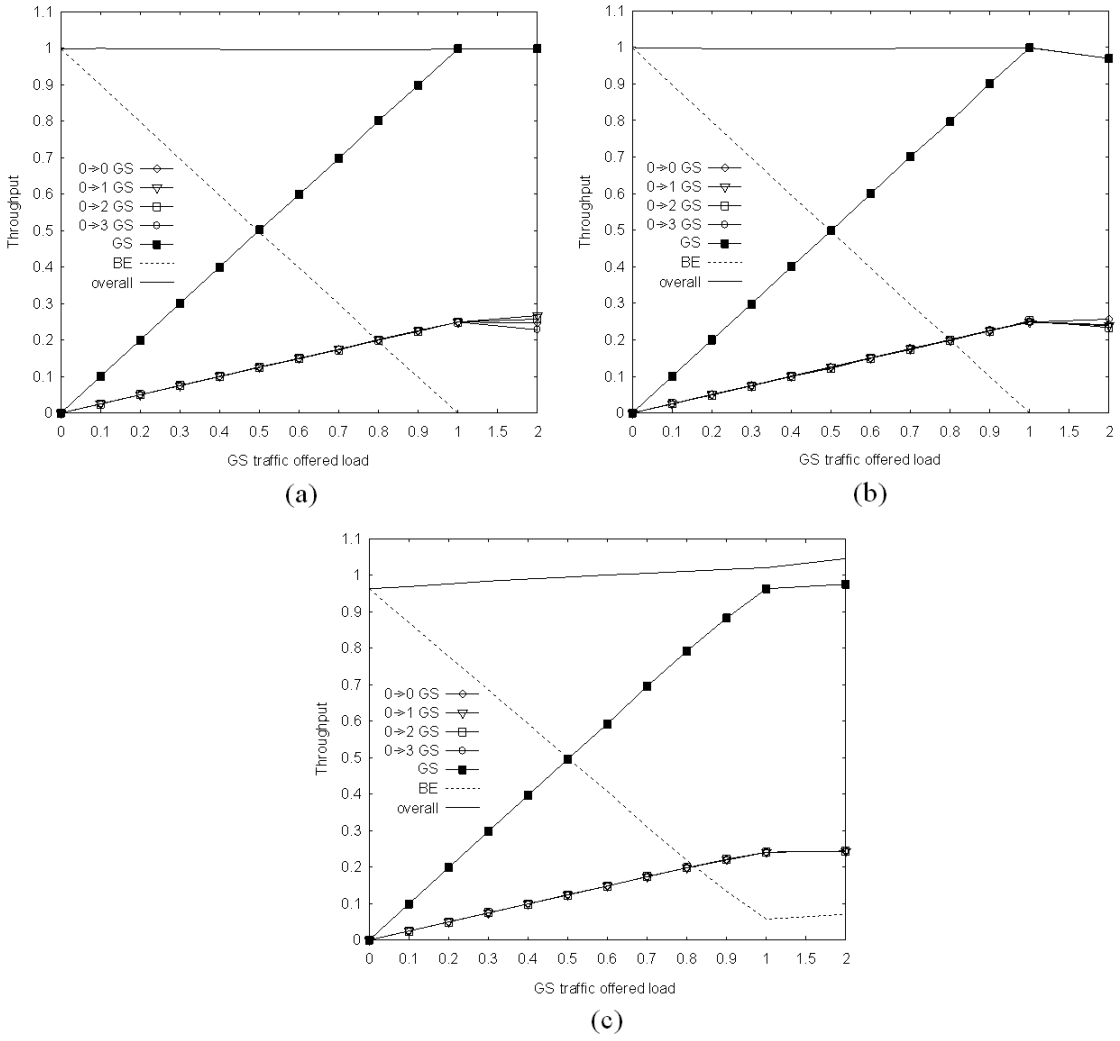

Figure 5.13: Throughput as a function of GS traffic relative load under the uniform traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration

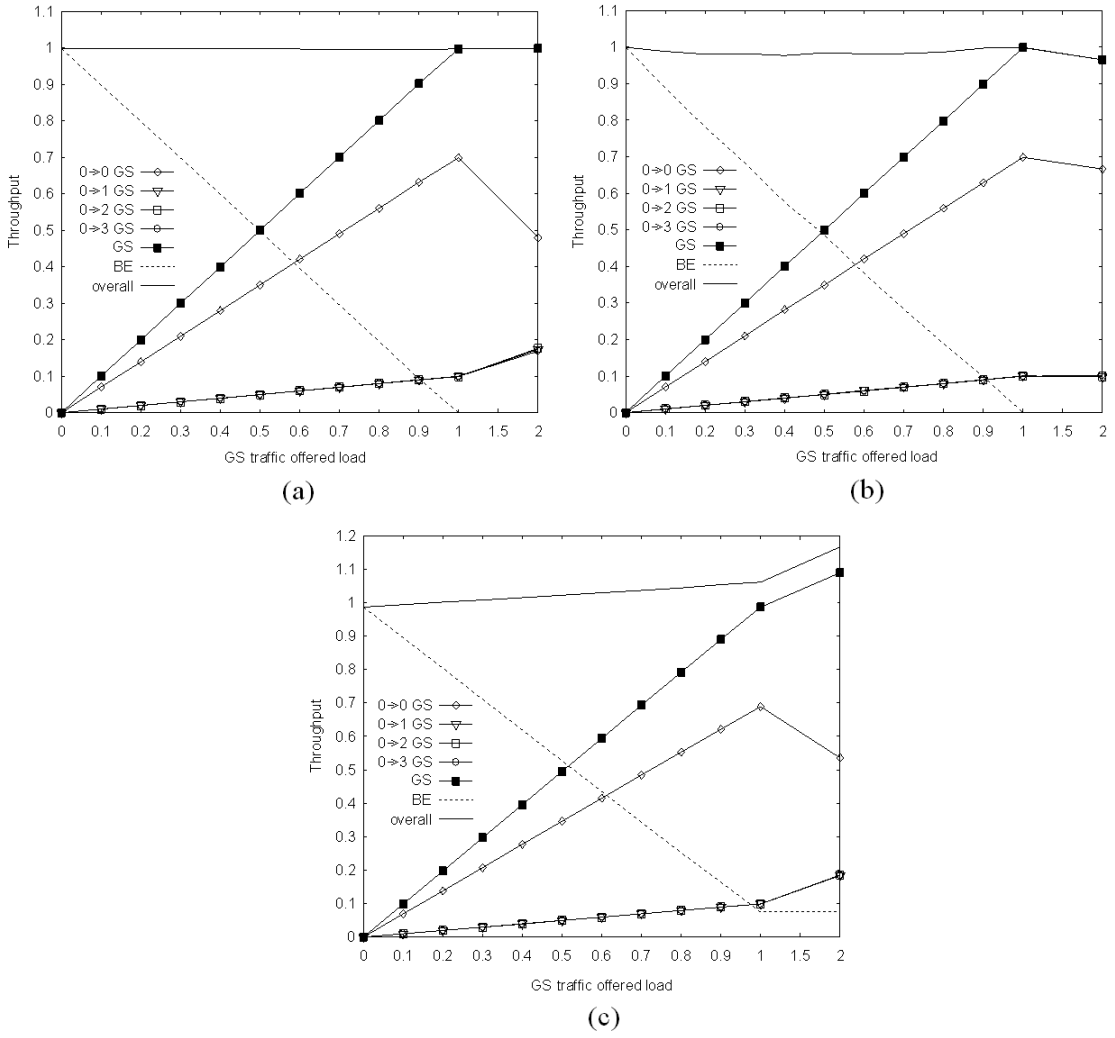

Figure 5.14: Throughput as a function of GS traffic relative load under the diagonal traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration

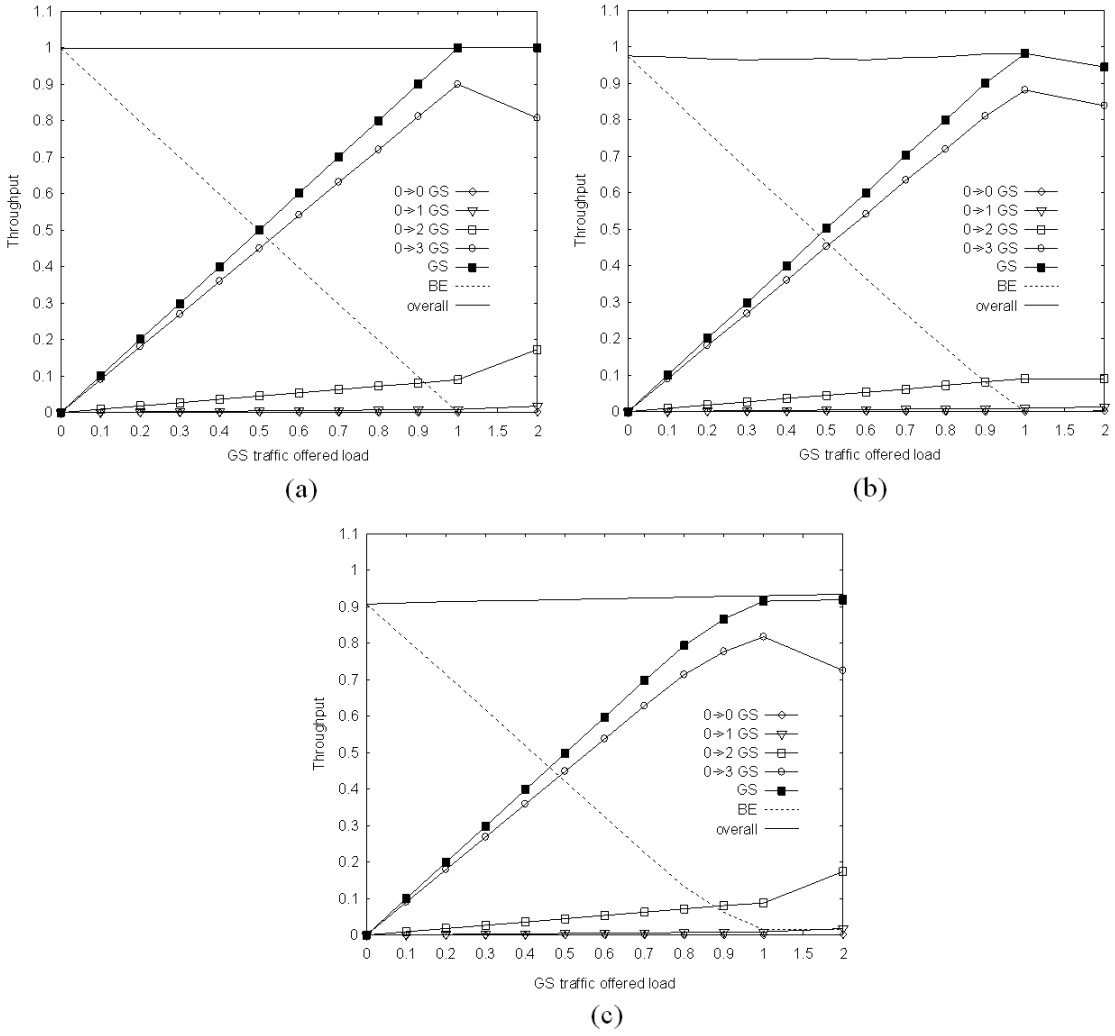

Figure 5.15: Throughput as a function of GS traffic relative load under the power-of-ten traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration

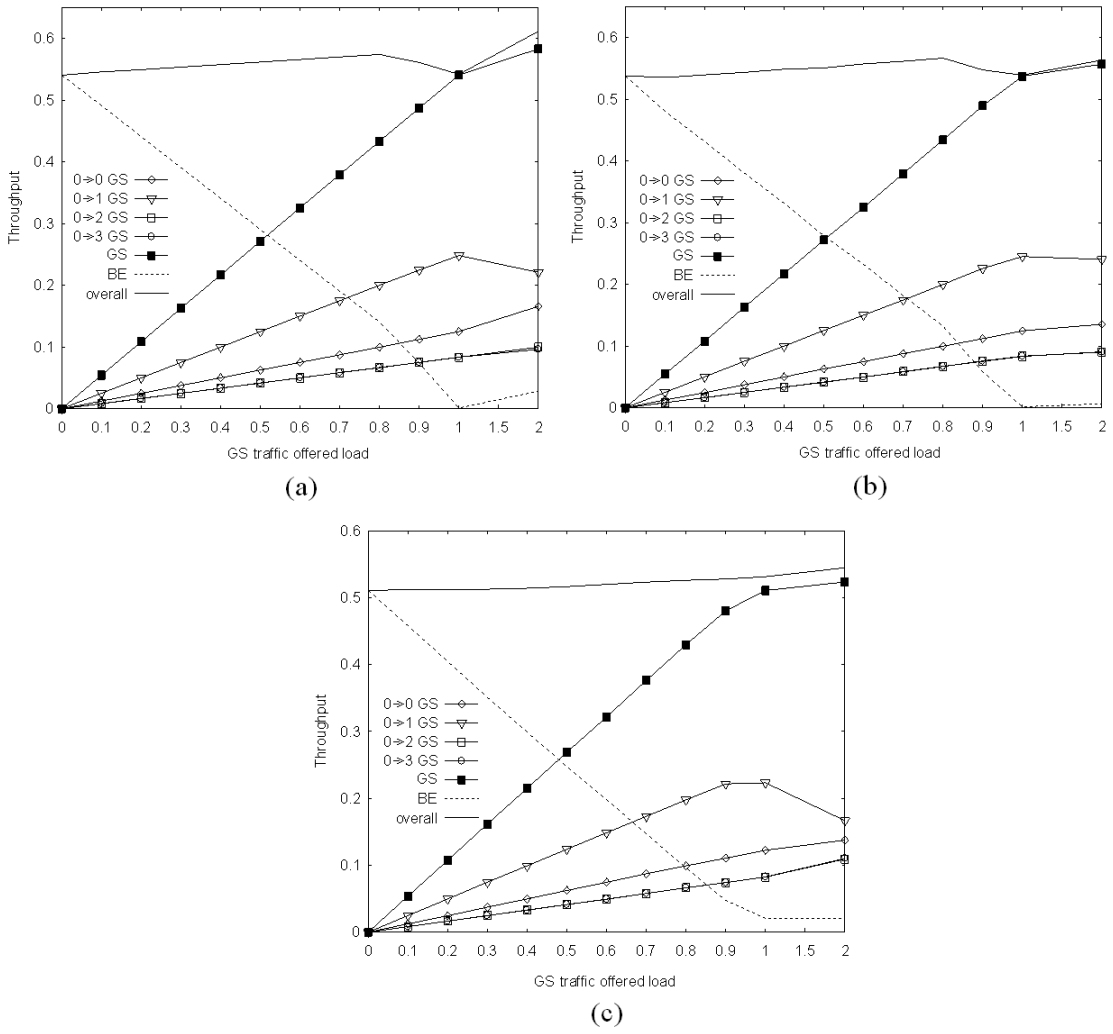

Figure 5.16: Throughput as a function of GS traffic relative load under the very unbalanced traffic matrix: comparison of (a) the optimal and (b) the heuristic solution for the passive multi-ring configuration, and (c) the heuristic solution for the active multi-ring configuration

5.6 Summary

In this part of the thesis, we focussed on the multi-ring network architecture. Its performance has been evaluated by simulation considering several scenarios. The performance results have been obtained using a real scale simulator including self-similar traffic model and different traffic patterns between interconnected rings.

Two main weaknesses have been identified and overtaken by optimized mechanisms. The spatial reuse capability is exploited adding a special field in the control slot. The results indicate that, as expected, the higher the percentage of intra-ring traffic the more the gain of performance due to the spatial reuse. This is more evident in small networks (low number of rings). This result is quite obvious if we consider that the spatial reuse is exploited only in presence of intra-ring traffic. The congestion notification mechanism has been improved using the multi-slots as triggers for notifying explicit congestion signals from nodes to the Hub. The validity of the proposals have been demonstrated by numerical results.

Finally we discussed the problem of allocating resources to provide guaranteed and best-effort services in different configurations of the DAVID metro network. We discussed architectural alternatives, and different formulations of the resource allocation problem. The latter is solved in a mostly centralized fashion at the Hub, but some access decisions may be de-centralized at network nodes, depending on the network configuration, and on the desired level of performance guarantees. Tradeoffs between optimality and complexity of the allocation schemes were observed by simulation. Our study shows the flexibility of the network architecture, and the effectiveness of the proposed strategies in accommodating very diverse traffic patterns.

Chapter 6

Benchmarking

In order to compare the multi-ring and multi-PON solutions described previously, we perform a dimensioning and benchmarking study in terms of cost/effectiveness.

In addition, any OPS-based solutions, when mature for commercial deployment, will naturally have to compete with SONET/SDH, and with other recent metro technologies such as Ethernet (IEEE 802.3) or RPR (IEEE 802.17) [96]. Therefore, we do not restrict ourselves to detailing the multi-PON and multi-ring performance, but also compare them to non-OPS technologies. Results in the following sections received important inputs, in terms of traffic scenarios and of network architectures, by manufacturer and operator members of the DAVID project, which made available their internal confidential information to all partners.

Our contributions on this task deal with the resource dimensioning of multi-PON, multi-ring and RPR solutions. To complete the benchmarking environment, in this chapter we also include the contributions of the other partners which focus on defining the network scenario (network operator partners), dimensioning the passive multi-ring, SDH and Ethernet solutions and evaluating the CAPEX and OPEX costs (both tasks performed by the manufacturer partners). Therefore, the use of *we* in the following sections refers to all DAVID project partners.

6.1 Methodology and network scenario description

The methodology consisted of fixing an initial traffic matrix and applying it to the different network architectures. Through computer simulation and analytical models, we determine the resources required in each network architecture (number of transceivers, number of wavelengths, number of optical amplifiers, etc.) to have similar performance (packet loss rate, delay, jitter). Figure 6.1 schematically illustrates this methodology.

The study is restricted to a common network scenario with one Hub and 16 nodes distributed over a 100-km ring network. Four different node types are considered: 1 server node, 2 big nodes, 4 medium nodes, and 9 small nodes. We also considered three different mean traffic volumes: 20 Gbit/s (20G), 40 Gbit/s (40G) and 80 Gbit/s (80G). In addition we fixed the ratio between the up- and downstream traffic in the

```
Objective:

  • GIVEN a traffic matrix and a number of nodes
  • FIND the minimum dimension of the nodes and Hub
  • SUCH THAT the traffic matrix is successfully delivered
```

```
Iterative method:

  1. Calculate the network dimension based on mean traffic requirements;
  2. Set up a realistic scenario and run a simulation;
  3. If at least one queue is congested (for example above 1 ms of waiting
     time);
       (a) Search the node with the more congested queue and increase its
           capacity by adding a wavelength;
       (b) Repeat step 2;
  4. End.
```

Figure 6.1: The network dimensioning methodology

Table 6.1: Node types and traffic assumptions

Node type	**Quantity**	**Upstream traffic**	**Downstream traffic**
Server	1	20.0%	2.4%
Big	2	3.2%	8.4%
Medium	4	1.6%	4.8%
Small	9	0.8%	2.4%
Total	16	40.0%	60.0%

network and the number of nodes per type on the ring. This is summarized in Table 6.1. Finally, we considered 55% of the total generated traffic coming from the backbone through the gateway, while 80% of the traffic generated at the nodes was destined to the gateway. The network characteristics considered in this study were chosen to reflect typical metro scenarios encountered by operators. The ring length of 100 km was chosen to be compatible with the node cascadeability constraints derived in [98], while the number of nodes was chosen to match the limiting size of SONET/SDH rings. The diversity in the node types and their respective traffic volumes are believed to be representative for mid term metro networks.

Two dimensioning studies have been considered. The first one regards the comparison between the multi-PON and the multi-ring architectures. The second one regards the comparison between the multi-ring architecture and the passive multi-ring architecture, SONET/SDH, Ethernet and RPR.

Table 6.2: Major components quantities for mean traffic (80G scenario)

Device	**multi-PON**	**multi-ring no space reuse**	**multi-ring**
Fiber	423 km	400 km	400 km
Multiplexer port	1144	1456	1376
10 Gbit/s Txs	160	160	144
10 Gbit/s Rxs	44	44	40
Wavelength converter	8	18	17

6.2 Multi-PON versus multi-ring

To perform this evaluation, we firstly analyze the optical hardware dimensioning based on mean traffic requirements for both multi-PON and multi-ring networks. For the latter, we also consider the possibility to remove the space reuse capability of the ring topology. Packet-level medium access control protocol simulations, using various scheduling algorithms are then run to compare component and network requirements under more realistic statistical traffic fluctuations, using a maximum packet delay criterion of 1000 time slots.

The scenario assumes 80 Gbit/s total traffic capacity to and from the 16 nodes of a metro network. All 16 nodes can be supported in a single PON, but must be partitioned into two 8-node rings. This is due to the power budget limitations of the active node structure studied in [98]. The Hub traffic goes via its own ring or PON to and from the core network. This means that the architectures comprises:

- 1 PON of 16 nodes + 1 PON for the connection to the core network;
- 2 rings of 8 nodes + 1 ring for the connection to the core network.

To minimize the complexity of this study, and to enable comparisons between multi-rings and multi-PON to be made analytically, the physical topology of the metro network is taken to be a circle of radius R km, with all nodes plus the Hub switch equally spaced around it. Such a perfect structure is certainly not normal, but it is no more unique nor less meaningful for comparing multi-ring and multi-PON than any other arbitrary, real-life node distribution.

The quantities of the most costly components are summarized in Table 6.2 for PONs using frame-based scheduling and for multi-ring with and without space reuse, including the PON and ring connections to the core network.

Traffic-dependent component quantities are calculated from the mean values of the traffic matrix in table 6.1. Because all Txs/Rxs are assumed to be 10 Gbit/s, this therefore requires 3 wavelength channels from the Hub. The total downstream traffic from Hub to node PON is 48.0 Gbit/s (Table 6.1), requiring 5 channels. So 8 channels are needed downstream from the Hub in total, hence requiring 8 wavelength converters. The numbers of Txs and Rxs needed by each node are similarly derived

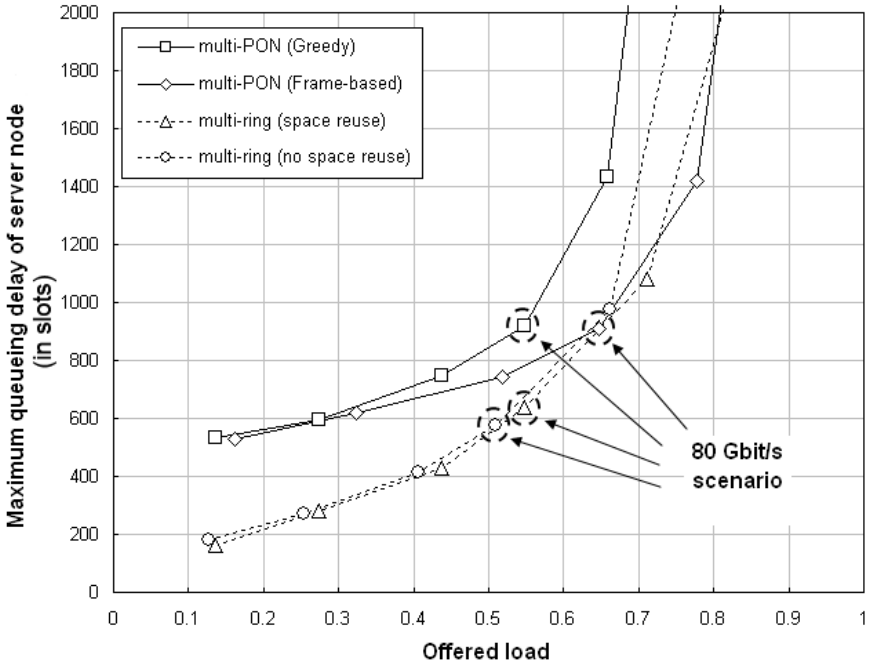

Figure 6.2: The considered switch architecture

from the Tx and Rx capacities per node. The numbers of Txs and Rxs needed for the core connection are derived from the total capacities to and from the Hub.

For multi-ring more wavelength converters are required partly because the traffic is fragmented between two 8-node rings but mainly because 3R regeneration/wavelength conversion are needed at the Hub input as well as the output. Spatial reuse provides modest component savings. The major component savings between multi-PON and multi-ring are in wavelength multiplexers.

The packet scheduling simulations begin with the component resources that would be installed to support the mean traffic capacities. But packet delays caused by self-similar traffic sources and matching algorithm inefficiencies, which particularly affect the server node, require more resources (channels, Txs and Rxs) to be added, in discrete increments, until a maximum queueing delay below 1000 time slots is obtained. Figure 6.2 shows the maximum delay vs. offered load curves (i.e. delay vs. traffic capacity) for the 4 resulting multi-PON and multi-ring networks, each now requiring different installed capacities. The loads corresponding to the 80 Gbit/s scenario are ringed. These are the only points for which each network supports precisely 80 Gbit/s mean traffic with minimum resources to guarantee less than 1,000 slot maximum delay. Multi-PON with frame-based scheduling (and a Hub with slot-by-slot switching between channels) require the smallest of the 4 network capacities and hence achieve the highest load for the allowable 1000 slot queueing delay. 11 downstream channels and hence converters are needed in total, instead of 8, which increases the network capacity to 110 Gbit/s. Multi-ring and PONs using a greedy scheduling algorithm all require greater network capacity, and hence provide lower loads, than multi-PON using frame-based scheduling.

The results indicate that the multi-PON solution can provide cost reduction over multi-ring. This is mainly due to the partitioning of the nodes into two 8-node rings. New physical layer simulations carried out in [42] indicate that, using the components and system parameters measured in the corresponding subsystems of the DAVID demonstrator, 16 nodes can be cascaded in a single ring. In this eventuality, the multi-ring approach seems comparable or superior than multi-PON.

6.3 Multi-ring versus passive multi-ring, SDH, Ethernet, and RPR

6.3.1 Benchmarked solutions

To compare the multi-ring approaches (both passive and active node structure, PMR and MR acronyms respectively) with the classical Ethernet, RPR and SDH approaches, the node structures shown in Figure 6.3 are adopted. For the Ethernet solution (Figure 6.3(a), we considered a star topology where each access node was connected directly to a central Hub through an unshared point-to-point fiber connection (doubled for protection). For both the SDH (Figure 6.3(b) and RPR (Figure 6.3(c) cases, we considered an opaque structure: optical Multiplexers (MUX) and Demultiplexers (DMUX) filter the optical channels which correspond to parallel rings terminated at each node. In the SDH approach, a single Cross-Connect (XC, switching at the STM-1 or STM-4 granularity) allows to connect to multiple rings as well as to provide add/drop access. The Hub in this case also is an SDH Cross-Connect (again switching at the STM-1 or STM-4 granularity) terminating/generating all wavelengths of the rings and of the gateway. To achieve protection capability, this structure is doubled. By nature, RPR relies on a single physical ring topology. To provide access to multiple wavelengths, multiple RPR chips are provided. Interconnection between the various RPR rings is achieved through an IP/MPLS Router, which also provides add/drop access to each of the thus stacked wavelength-rings. At the Hub, RPR interfaces are needed for all wavelengths and for connecting the gateway. The RPR architecture inherently has protection capabilities, since each physical ring is in fact composed of two counter-rotating rings. All node architectures include DABs to aggregate the data traffic coming from/going to the client layer.

6.3.2 Resource dimensioning

Taking into account the functionality and limitations of each network architecture, we performed benchmarking studies dimensioning the capacity required in each node and at the Hub to obtain similar performance. For this study we did not include any consideration of protection.

In Figure 6.4, we show the node capacity (in Gbit/s) required in each metro solution considering the three traffic volumes, while Table 6.3 illustrates the needs in terms of transport resources: fibers (including the connection between the Hub and

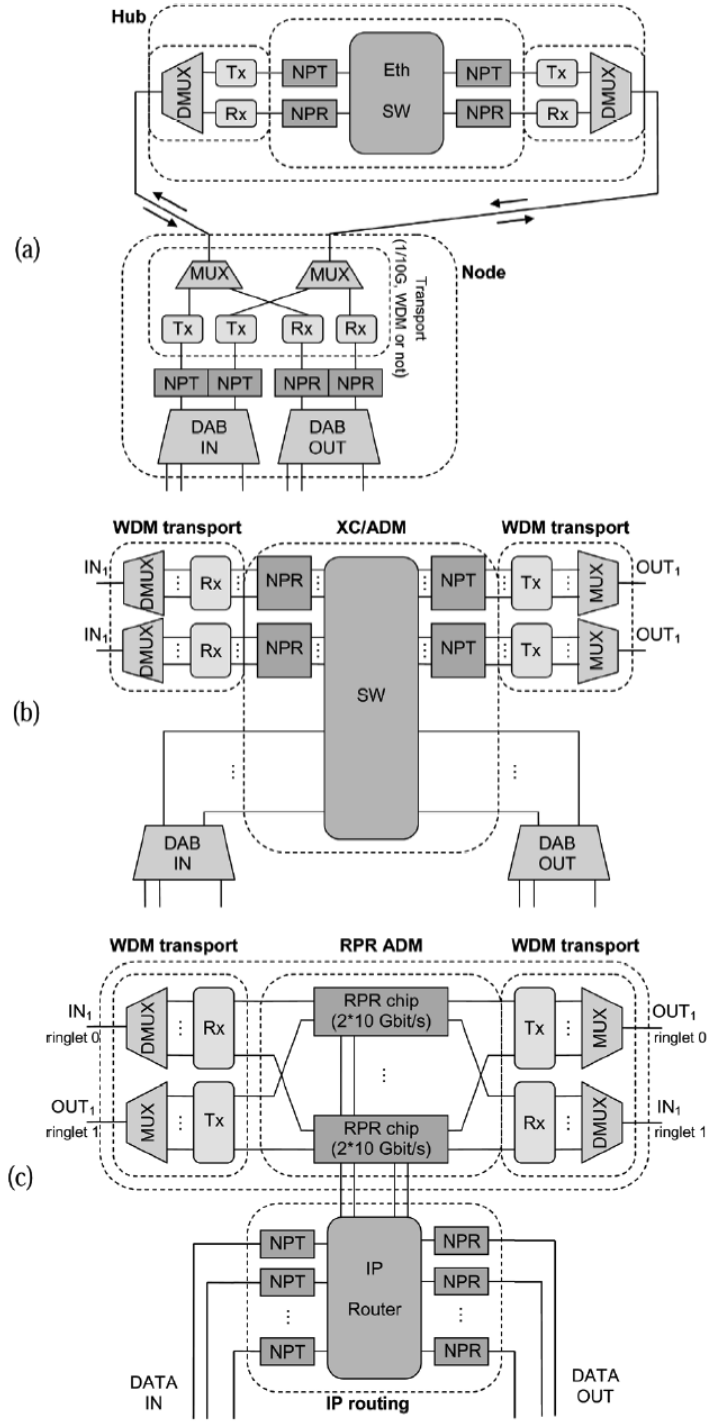

Figure 6.3: Node structures. (DMUX: wavelength demultiplexer; MUX: wavelength multiplexer; NPR: network processing receiver; NPT: network processing transmitter; SW: STM-1/STM-4 switch; Eth SW: Ethernet switch; XC: cross-connect; DAB: data aggregation board. (a) Point-to-point Ethernet Hub + Node, (b) SDH node, and (c) RPR node.

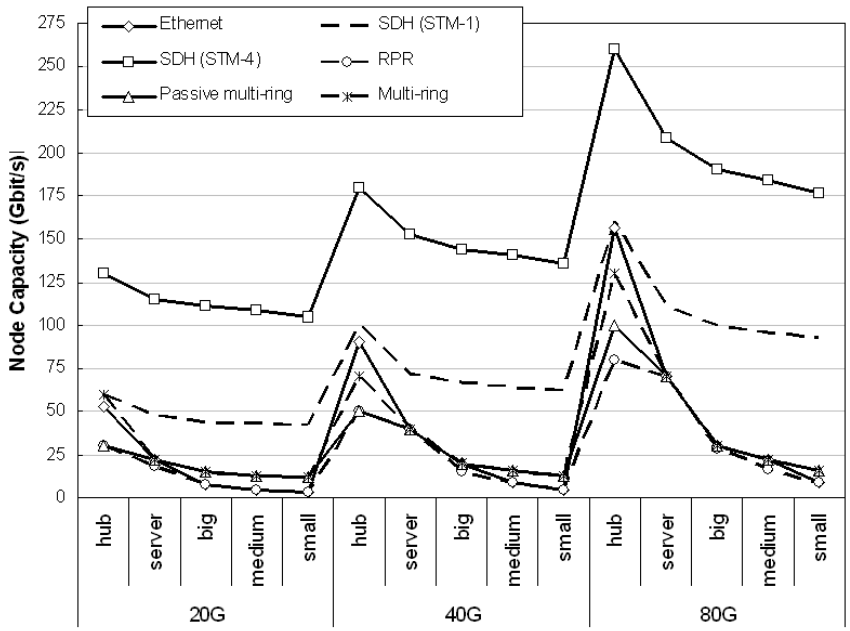

Figure 6.4: Node capacity (in Gbit/s) required in the different network architectures for the three traffic volumes

the gateway), wavelengths (either 1 Gbit/s or 10 Gbit/s channels), and transceivers (either 1 Gbit/s or 10 Gbit/s TRx).

Figure 6.4 offers important results concerning the required resources as well as the scalability of the architecture when the traffic increases. In the SDH case, the major part of the node size is used for transit traffic (hence the smaller relative differences in required capacity between node types) which causes over-dimensioning. This stems from the fact that at least one circuit must be established between each source-destination pair in the network. This effect can be considerably reduced by using SDH circuits just between nodes and the Hub in a star topology rather adopting the ring-approach. In contrast, for the active multi-ring architecture, the required node capacity is nearly optimal thanks to the flexible design and the optical by-pass capability. Nonetheless, the waveband concept (which avoids the need of a full 32-wavelength selector in each ring node) imposes an over-dimensioning of the Hub. The packet-based passive multi-ring, RPR and Ethernet solutions are very similar in terms of dimensioning. Ethernet has a slight gain in nodes due to the possibility to use low bit rate interfaces, but a drawback at the Hub due to the non-shared transport resources and the star topology. For instance, it requires 44 fibers for the 40G scenario. From Figure 6.4, the RPR solution seems the better one since all nodes, as well as the Hub require less capacity with respect to the other solutions. Nevertheless, the opaque structure of the RPR forces a high number of transceivers as show in Table 6.3.

Table 6.3: Transport resources required in the different architectures

Scenario	**Device**	**Ethernet**	**SDH (STM-1)**	**SDH (STM-4)**	**RPR**	**PMR**	**MR**
20G	Fiber	31	2	2	3	2	2
	1G channel	46	0	0	0	0	0
	10G channel	6	6	13	3	6	6
	1G TRx	46	0	0	0	0	0
	10G TRx	6	72	176	21	24	33
40G	Fiber	44	2	2	3	2	2
	1G channel	60	0	0	0	0	0
	10G channel	12	10	18	5	10	7
	1G TRx	60	0	0	0	0	0
	10G TRx	12	110	231	40	29	36
80G	Fiber	72	2	2	3	2	2
	1G channel	64	0	0	0	0	0
	10G channel	26	16	26	8	19	13
	1G TRx	54	0	0	0	0	0
	10G TRx	26	167	307	61	41	50

Table 6.4: Example of CAPEX analysis: cost relative to the passive multi-ring

Scenario	**Ethernet**	**SDH (STM-1)**	**SDH (STM-4)**	**RPR**	**PMR**	**MR**
20G	-15%	+10%	+135%	-28%	0%	+58%
40G	+14%	+38%	+167%	-13%	0%	+50%
80G	+23%	+65%	+189%	+19%	0%	+45%

6.4 Example of CAPEX analysis

An extensive CAPEX analysis based on the resource requirements highlighted in the dimensioning studies of each architecture is performed within the DAVID project based on component costs obtained by confidential means and market survey. Since the results provided here are not part of our contribution, we only show an example of the CAPEX analysis (Table 6.4).

The costs are counted relative to the CAPEX for the passive multi-ring (PMR) architecture. RPR is the cheapest solution only for the initial capacity: when increasing network capacity, the optical transparency provided by the passive optical architecture enables to obtain lower CAPEX. Indeed, the PMR solution is quite competitive even for low traffic volumes, being the second cheapest solution behind RPR for the 40G traffic scenario and the cheapest solution for the 80G traffic scenario. The Ethernet and SDH solutions are in most cases not highly competitive, in both cases due to the non-sharing of resources.

The active multi-ring solution pays, with the initial assumptions (limited capacity), for the complexity of the nodes. In addition, the traffic matrix with a high proportion of extra-ring traffic (80%) is clearly a disadvantage for the active multi-ring solution which can not strongly exploit the optical space re-use mechanism.

It is important to note that the limited capacity penalizes the use of an optical Hub for both passive and active multi-ring networks. An architecture similar to the passive multi-ring solution but using an Ethernet switch at the Hub (like DBORN [73]) can be more appropriate for a first introduction of optical packets in metro networks.

Extending the initial traffic matrices up to 1Tbit/s scenarios (with 160G or 320G on 4 or 2 rings, where each ring should be further doubled for protection), the passive multi-ring shows the best CAPEX value. On the other hand, implementing correcting factors to the initial component cost assumptions to take into account some optics cost reduction (foreseen at production of higher volumes for optical components), the active multi-ring solution becomes an interesting solution [42].

6.5 Example of OPEX analysis

As for the CAPEX analysis, here we also show an example of the OPEX analysis based on the resource dimensioning requirements.

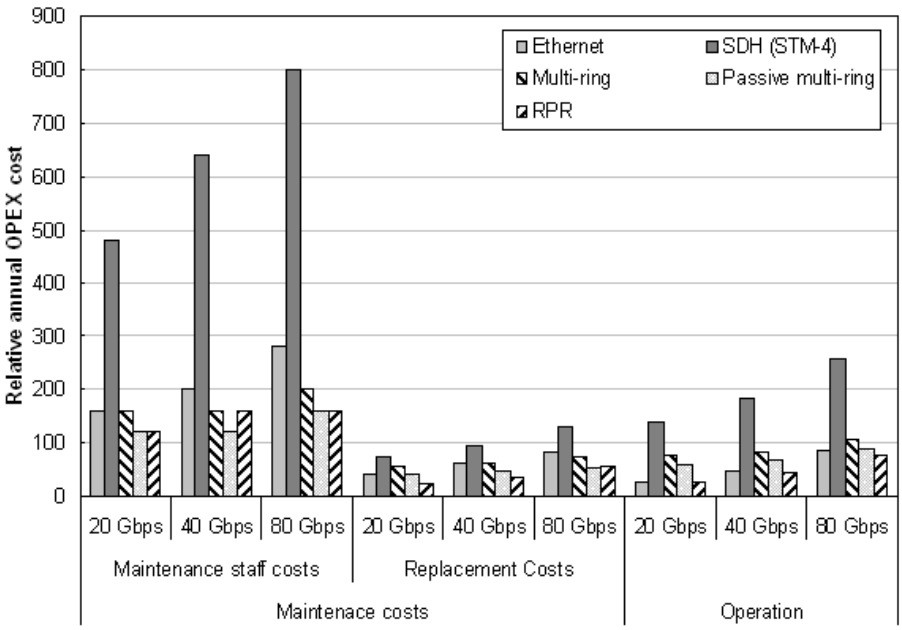

Figure 6.5: Example of OPEX analysis: relative annual OPEX cost comparison in the different network architecture for the three traffic volumes

A common model where annual costs have been calculated as a percentage of the equipment costs is adopted to calculate the OPEX values. OPEX includes various operational costs, ranging from administrative costs, over service development, to network planning costs etc. The comparison is limited to costs related to network operations and maintenance mainly because other cost factors are most likely to not significantly differ between the various architectures.

The maintenance costs have been defined as all the costs related to the resolution of physical problems in the network such as fiber cuts or equipment failure. It can be calculated as the sum of replacement costs and the maintenance staff costs. The first part encompasses the cost of failed network elements and is proportional to its failure probability, while the second includes labour costs and obviously depend on the required amount of personnel. The operational costs include all the recurrent costs which are periodically necessary for undisturbed operation.

OPEX results for the different network architectures are depicted in Figure 6.5. The cost specific values are expressed as relative to the cost of 1 fibre.km

The OPEX costs for the SDH ring are considerably larger than for the other scenarios, since it includes many more network elements, most of them electronics. RPR on the other hand, being the option with the fewest number of network elements, presents the lowest OPEX costs. Yet it is closely followed by the passive multi-ring, Ethernet and active multi-ring solutions.

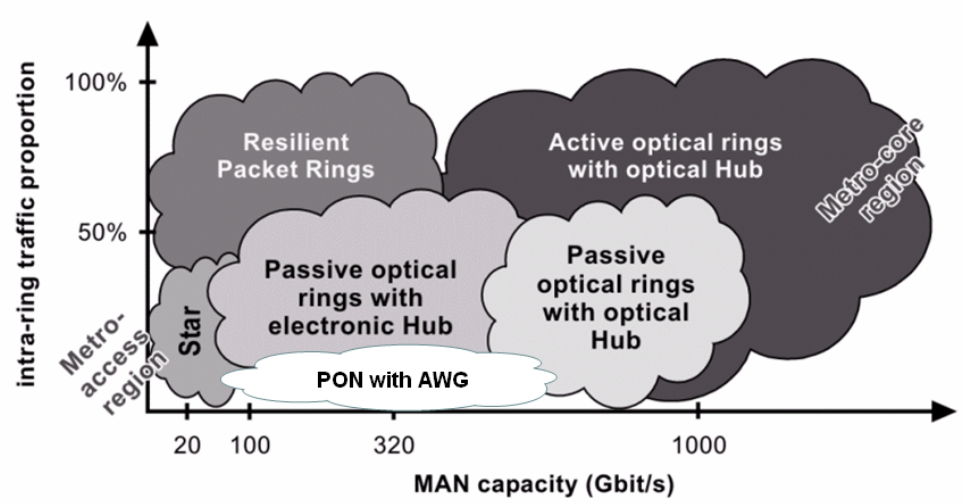

Figure 6.6: Possible introduction scenario of the different metro technologies

6.6 Conclusions and perspectives

From the extensive benchmarking results -whose results have been summarized above- and despite all uncertainties of market analyzes and forecasts, we can foresee a possible introduction scenario of the different metro technologies with respect to the required capacity and the traffic repartition. Figure 6.6 depicts this scenario, whose tendencies could be summarized as follows:

- With low capacity (few tens of Gbit/s), two advantageous solutions can be identified: the star Ethernet (possibly with WDM to share fibre resources) when the ratio of intra-ring traffic is low, whereas RPR appears the most optimized solution thanks to space re-use capability.

- At a short/medium term with increasing access bit rate and resulting metro capacity in the range of tens to few hundreds of Gbit/s, passive optical ring structure with an electrical Hub is well suited, as in the DBORN architecture proposed by Alcatel [73]. Due to the lack of transparency, RPR requires a high amount of transceivers and filtering ports on the ring which makes the solution less competitive. The multi-PON architecture seems a good alternatives when large amount of information is changed between different PONs (i.e., modest intra-ring traffic).

- At a longer term, under the assumption of a strong introduction of high bit rate access networks (FTTx, GPONs), the capacity in the metro can reach hundreds of Gbit/s to 1 Tbit/s. In this case, the two DAVID solutions become competitive, thanks to the optical transparency both at the node and Hub levels.

PART III

OPS-based wide area networks

Chapter 7

Introduction to the OPS-based wide area network

7.1 State-of-the-art

In this thesis, we consider the general switch architecture with full connectivity and wavelength conversion shown in Figure 7.1 and capable to switch asynchronous, variable-length packets [35]. This switch acts as an output queueing switch; it uses a feed-forward configuration [60] and the optical buffer is made by B FDLs. The electronic Switch Control Logic (SCL) takes all the decisions regarding the configuration of the hardware to realize the proper switching actions. When a packet arrives, the SCL examines the header and lookups the forwarding table to determine the output fiber, determining also the network path. Successively, the SCL performs the following functions:

- choose which wavelength of the output fiber will be used to transmit the packet, in order to properly control the output interface;
- decide whether the packet has to be delayed by using the FDLs or it has to be dropped, since the required queuing resource is congested.

These decisions are routing independent and all the wavelengths of a given output fiber are equivalent for routing purposes but are not from the contention resolution point of view. The choices of wavelength and delay are actually correlated, being the need to delay a packet related to the availability of the wavelength selected. This is what we call the *Wavelength and Delay Selection* (WDS) problem. We therefore consider contention resolution policies able to exploit only the *time* and *wavelength* domains and not the space domain.

The technology limitation of the optical queuing motivates significant research efforts in recent years dealing with the design of simple WDS contention resolution policies (see for instance [48] [99] [15]). Almost all of them solve the contentions on a per-packet basis, i.e. the WDS algorithm is executed at each incoming packet (we call this approach *connectionless*). This means that once the forwarding component has

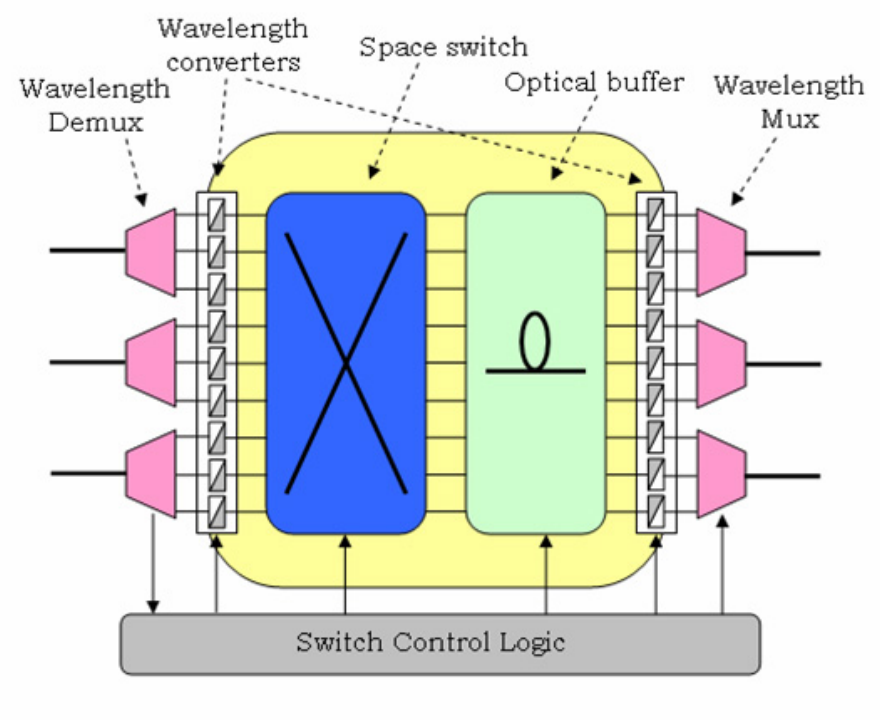

Figure 7.1: The considered switch architecture

decided to which output port p the packet should be sent, the functions performed by the SCL are:

1. Search for the set of wavelength $W \in p$ not busy;
   (a) If $W = \emptyset$ (i.e., all queues are full), the packet is lost;
   (b) If $W \neq \emptyset$, select the wavelength $w \in W$ to transmit the packet on;
2. Determine the delay D_j and select the FDL j to send the packet to.

The choice of the wavelength in step 1.b is the key point and can be implemented by following different policies, producing different processing loads at the SCL and different resource utilizations. In [15] several heuristic WDS algorithms were presented and studied, showing that the choice of the algorithm may significantly change the performance. In fact, when a packet has to be buffered, the choice of the delay is not free, since the number of delays available within a FDL buffer is discrete. As explained in [14], this creates gaps between queued packets that can be considered equivalent to an increase of the packet service time, meaning an artificial increase in the traffic load (*excess load*). It has been demonstrated in [99] that a WDS algorithm (called VOID algorithm) that aims at minimizing those gaps gives best performance with respect to other policies. Nonetheless, the computational complexity of the VOID algorithm is very high since it requires to know the length and the duration of every gap in the queues. A simplification of this algorithm called MINGAP is proposed in

[15]; it selects the FDL with the minimum gap only between the last queued packet and the new one.

Despite the fact that the WDS algorithm can be relatively simple to implement, taking per-packet decisions requires too much computations considering that each switch has several ports, each port several wavelengths and each wavelength transports packets at 10 Gbit/s or above. To overtake this problem, OPS concepts are recently extended to a connection-oriented network scenario [16], for instance based on MPLS. In this scenario, a suitable design of WDS algorithms permits to obtain fairly good performance, by exploiting queuing behaviors related to the connection-oriented nature of the traffic, but with a significant saving in term of processing effort for the switch control with respect to the connectionless case.

7.2 The connection-oriented OPS network

The connection-oriented OPS network comprises several nodes connected in a mesh topology. Based on destination address and quality of service requirements, packets coming from the client networks are classified at the edge nodes into a finite number of subsets such as the *Forwarding Equivalent Classes* (FECs) concept defined in MPLS environment. Each FEC is identified by an additional *label* added to the packets. Edge nodes are in charge of setting up and maintain the unidirectional Optical Virtual Circuits (OVCs) throughout the network. Packets belonging to the same FEC are identical from a forwarding point of view and are transferred from source to destination along the OVC which corresponds to their label. On each core node, a simple label matching operation is performed on a pre-computed OVC forwarding table, thus simplifying and speeding up the forwarding function.

Due to the high number of traffic flows being typically transported by a WDM network, the adoption of a pure MPLS-like labeling scheme may result in an excess of per-flow information to be handled by the optical nodes. In order to avoid scalability problems, here we assume that each OVC represents a top-level explicitly routed path formed by an aggregation of lower-level connections including several traffic flows such as what proposed in [68]. Following this approach, the number of OVCs managed by a single optical core router is not supposed to be too high and to affect the correct label processing.

Figure 7.2 shows an example of the connection-oriented OPS network. An OVC forwarding table is setup in the node 1 which indicates that packets with label 25 coming from port 0 with a pink wavelength should be forwarded to the output port 2 with a blue wavelength and a new label 12.

While the information on the output fibre is given by the routing protocols (no subject of study here), the choice of the wavelength may be taken locally by each node taking into account the availability of time resources. This problem can be solved by following different strategies:

- **Static**. The OVC is assigned to a given wavelength at OVC setup and this assignment is hold over the OVC life. Therefore packets belonging to the same

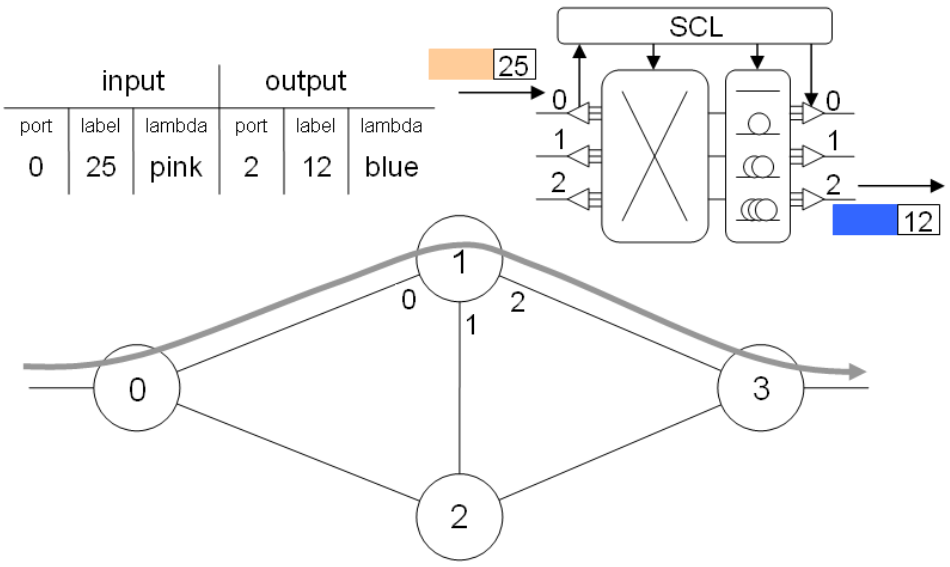

Figure 7.2: Connection-oriented OPS network

OVC are always switched to the same wavelength and the contentions can be only solved in time;

- **Dynamic**. The OVC is assigned to a wavelength at OVC setup but it can be changed during OVC life. When heavy congestion arises on the assigned wavelength (i.e., when the time domain cannot solve a contention), the OVC is temporary switched to another wavelength. When congestion disappears, the OVC is switched back to the original wavelength.

The static wavelength selection requires minimum control complexity since processing is performed only at OVC setup. At the same time, it preserves the correct order of packets belonging to the same OVC since new arrivals cannot overtake older packets. However it does not optimize the resources obtaining high PLR figures. On the other hand, when a dynamic algorithm is executed, the OVC is switched to an alternative wavelength that is not (or is less) congested and new incoming packets on that OVC will experience in general less queuing time than older packets and will very likely overtake them along the network path. At the same time, the amount of execution of the algorithm affects the processing load on the SCL ranging from no efforts if static approach is used to fairly demanding efforts if a new wavelength search is executed per each incoming packet (e.g., [99] [15]).

7.3 Problems addressed in this thesis

In this thesis we focus on a connection-oriented OPS network scenario taking into account both static and dynamic approach. We address two problems, namely the problem of setting up of the OVC, properly configuring the forwarding table at the nodes, and the problem of providing QoS.

Static

When a packet belonging to OVC l arrives:

1. Look-up the forwarding table to determine $n(l)^{out}$ and $\lambda^{out}(l)$ assigned to l;
2. Determine the minimum delay D_j available for $\lambda^{out}(l)$;
   (a) If $D_j > D_M$, the packet is lost;
   (b) If $D_j \leq D_M$, select the FDL j to send the packet to.

Dynamic

When a packet belonging to OVC l arrives:

1. Look-up the forwarding table to determine $n^{out}(l)$ and $\lambda^{out}(l)$ assigned to l;
2. Determine the minimum delay D_j available for $\lambda^{out}(c)$;
3. If $D_j > D_M$ (i.e., $\lambda^{out}(l)$ is congested), search for the set of wavelength $\Lambda \in n^{out}(l)$ not busy;
   (a) If $\Lambda = \emptyset$, the packet is lost;
   (b) If $\Lambda \neq \emptyset$, select the new wavelength $\lambda_{new} \in \Lambda$ to transmit the packet on and update the forwarding table $\lambda^{out}(l) = \lambda_{new}$;
4. Determine the delay D_j and select the FDL j to send the packet to.

Figure 7.3: Contention resolution techniques in connection-oriented OPS networks.

Concerning the former problem, at the OVC setup, each node must assign both the output port and the output wavelength to the OVC in such a way that the packets belonging to that OVC are always switched to the same output. This double setup problem is different with respect to the *classical* RWA problem in circuit-switched network because here the wavelengths are shared among several OVCs (in a packet-switched basis). In this study we do not deal with the problem of selecting the output port which depends on the routing protocol but we are interested in the election of the wavelength which may be set locally by each node using a *OVC-to-wavelength setup assignment* (OWSA) algorithm. In particular we show that intelligent OWSA procedures can considerably improve the performance of the switches. The intelligence relies on grouping the flows coming from the same input wavelength which allows to obtain the conflict-free situations and hence reduce the contention probability.

Concerning the latter problem, existing solutions to provide QoS in OPS networks are based on the following strategy: 1) design a contention resolution algorithm which minimizes the Packet Loss Rate (PLR), thus 2) apply a QoS mechanism (some form of resources reservation on top of the contention resolution algorithm) able to differentiate the PLR among two or more classes. Given that we are dealing with a connection-oriented model, here we suggest a new method based on the well known ATM scheme of defining different service categories which consists of defining different OPS service categories, each one based on a different contention resolution algorithm specifically designed to cope with the requirements of that category. With this technique, besides the PLR, also the preserving the correct packet sequence and the computational complexity can be considered as important metrics for the QoS provisioning problem.

Let us spent some words on the out-of-order packet delivery problem. It is a serious problem since causes expensive reordering operations to be performed at the edges of the optical network and makes mandatory the use of very large memories due to the high speed of optical links. It is demonstrated in [71] [65] that even a small percentage of out-of-ordered packets seriously affects end-to-end protocols behavior, causing a considerable throughput degradation at the application level.

When considering TCP-based traffic it is well known that these phenomena influence the typical congestion control mechanisms adopted by the protocol and may result in a reduction of the transmission window size and consequently in bandwidth under-utilization. In particular the TCP congestion control is very affected by the loss or the out-of-order delivery of bursts of segments. This is exactly what may happen in the OPS network where traffic is typically groomed and several IP datagrams (and therefore TCP segments) are multiplexed in an optical packet, because optical packets must satisfy a minimum length requirement to guarantee a reasonable switching efficiency. Therefore out-of-order or delayed delivery of just one optical packet may result in out-of-order or delayed delivery of several TCP segments triggering (multiple duplicate ACKS and/or timeouts that expire) congestion control mechanisms and causing unnecessary reduction of the window size.

Another example of how out-of-sequence packets may affect application performance is the case of delay-sensitive UDP-based traffic, such as real-time traffic. In fact unordered packets may arrive too late and/or the delay required to reorder sev-

eral out-of-sequence packets may be too high with respect to the timing requirements of the application.

These brief and simple examples make evident the need to limit the number of unordered packets. In general out-of-order delivery is caused by the fact that packets belonging to the same flow of information can take different paths through the network and then can experience different delays [5]. In traditional connection-oriented networks, packet reordering is not an issues since packets belonging to the same connection are supposed to follow the same virtual network path and therefore are delivered in the correct sequence, unless packet loss occurs.

In an OPS network, using the wavelength domain for contention resolution (i.e. using dynamic policies), this may not be the case. Packets traveling along the same network path may use different wavelengths according to the choices of the algorithm. Therefore it may happen that packets of the same flow are delivered out of sequence, even though still following the same network path. Intuitively the reason is that the OVC is switched to an alternative wavelength that is less congested and new incoming packets on that OVC will experience less queuing time than older packets.

A possible solution could be to assume that this problem is solved at the egress edge-nodes that should take care of re-sequencing the various packet flows. This assumption in our view is not very realistic. It can be feasible for some flow of high value traffic, but is unlikely that will happen for all the flows of best effort traffic, because of the amount of memory and processing effort that would be necessary. Therefore we argue that it is important and necessary to control out-of-order delivery of packets directly in the OPS network nodes.

As explained above, this may cause an undesirable reduction in throughput as well as costly reordering operations with consequent unacceptable delays.

7.4 Simulation scenario

The performances of the proposed mechanisms are evaluated in order to assess their merits. The simulation results presented in the following section have been obtained by means of an ad-hoc event-driven simulator of the optical packet switch. The parameters of the switch are:

- N indicates the number of input and output fibers;
- W indicates the number of wavelengths per fiber;
- $\mathbf{Q}_B$ indicates the set of possible delays of B FDLs;
- D indicates the delay granularity;
- L indicates the average number of OVCs per input wavelength.
- ρ indicates the offered load.

- $\mathbf{M}$ indicates the fibre-to-fibre traffic matrix, whose generic element $\mathbf{M}_{i,j}$ is a real number ranging between 0 and 1 representing the percentage of traffic coming from input fibre i and going to output fibre j with respect to ρ. Three different traffic matrix are defined named: *uniform* $\mathbf{M}^U$, *power-of-two* $\mathbf{M}^P$, and *unbalanced* $\mathbf{M}^B$. For the case of $N = 4$, the matrices are as follows:

$$M^U = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad M^P = \frac{1}{15} \begin{bmatrix} 1 & 2 & 4 & 8 \\ 2 & 4 & 8 & 1 \\ 4 & 8 & 1 & 2 \\ 8 & 1 & 2 & 4 \end{bmatrix}$$

$$M^B = \frac{1}{30} \begin{bmatrix} 15 & 0 & 0 & 0 \\ 15 & 3 & 10 & 2 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 10 & 0 \end{bmatrix}$$

  These matrices represent a good sample of all possible traffic patterns: the classical uniform matric to evaluate a fair situation, the power-of-two matrix which demonstrates performance degradations when applied to the switches, and the unbalanced matrix to consider not balanced situations.

The distribution of the OVC requests follow an exponential model: both the interarrival time and connection duration of the OVCs are exponential distributed. The mean value of the interarrival times, connection duration, and required bandwidth are selected accordingly to generate the required offered load ρ.

The interarrival time of the packets is exponential distributed with a mean that depends on the OVC bandwidth. The packets are an exponential distributed size with average and minimum lengths of 500 and 40 bytes respectively.

We define the following measures to evaluate the performance of the switch:

- *Average Packet Loss Rate* (PLR). It is the usual performance measure for packet switches and also indicates the capability of an algorithm to reduce the congestion situation.

- *Out-of-sequence packets* (OS). The out-of-sequence packets delivery causes considerable throughput degradations and delay increases [71] due to the expensive reordering operations to be performed at the edges of the optical network. This measure indicates the percentage of out-of-sequence packets belonging to the same OVC.

- *Forwarding opacity* (FO). It is measured as the percentage of packets that are forwarded searching a new wavelength over the total number of simulated packets. The resulting value estimates the overload on the switch control function. The higher the percentage, the higher the overload.

In the following performance evaluation sections, we will show only the most significant measures according to the purposes of the study.

Chapter 8

The OVC setup in connection-oriented OPS networks

8.1 Problem description

In an OPS connection-oriented scenario, the configuration of the OVC forwarding table has a significant role to improve the network performance. In this context, a basic observation, as stated in [21], is that packets following OVCs incoming on the same input wavelength cannot overlap, because of the serial nature of the wavelength as a transmission line. Therefore such packets contend for output resources only with packets incoming on different wavelengths. As a consequence, if OVCs incoming on the same input wavelength are the only ones forwarded to the same output wavelength, contention will never arise (*conflict-free* configuration).

Figure 8.1 shows an example of a switch with $N = 2$ ports and $W = 3$ wavelengths per port. On wavelength λ_2^{in} of port n_1^{in} three OVCs are active: two (l_1 and l_2) are switched to λ_1^{out} of n_1^{out}, and the other (l_3) to λ_3^{out} of n_1^{out}. On port n_2^{in} there are three OVCs (l_4, l_5 and l_6) coming from different wavelengths. l_4 is switched to λ_3^{out} of n_1^{out} while l_5 and l_6 are switched to λ_2^{out} of n_2^{out}. By observing the figure it is trivial to understand that packets from l_1 and l_2 will never overlap (there are subject to an conflict-free allocation), whilst packets from l_3 and l_5 may overlap with packets from l_4 and l_6, respectively.

It is possible to quantify the influence of the overlapping to the performance of the switch. For the evaluation we use the simulation environment described in Appendix I where we also explain the meaning of the parameters. The following results have been obtained using a switch with $N = 4$, $W = 16$ and a degenerate buffer of length $B = 6$. The granularity of the FDL has been set to $D = 0.4$ because it is the optimal value for the static approach [21]. Each input and output wavelength is supposed to carry $L = 10$ different OVCs for a total of 640 incoming OVCs. We put our attention to a particular output wavelength. At different values of ρ, we carried out a set of simulations changing the relative load of the OVCs δ which is the percentage of the OVCs coming from the maximum loaded input wavelength. Formally, if the OVC i has a load of ρ_i, δ is calculated as

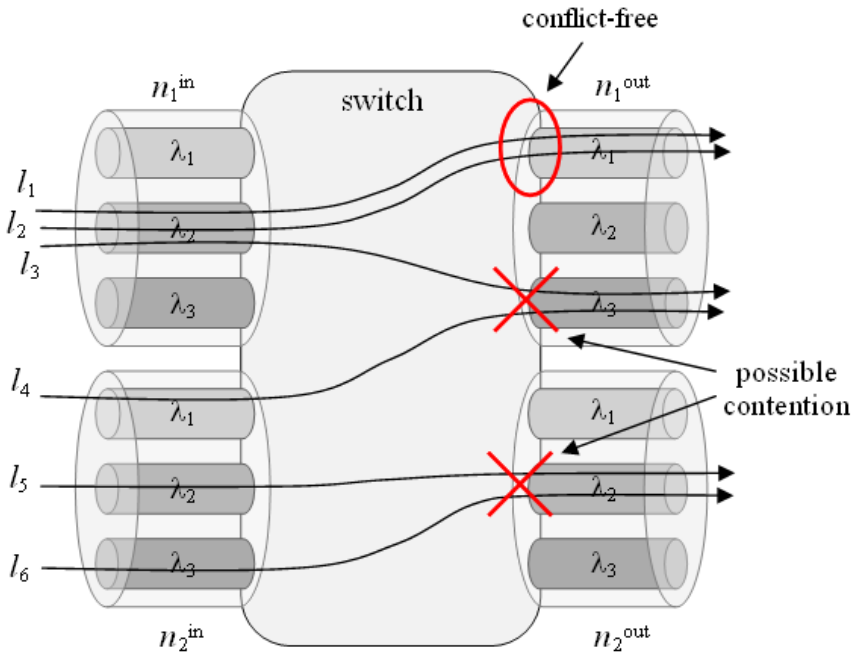

Figure 8.1: Example of OVC forwarding table configurations able to avoid and produce contentions.

$$\delta = \frac{\max_{i \in L} \rho_i}{\sum_{i=1}^{L} \rho_i}$$

For example, if the overall load is $\rho = 0.8$ and the relative load is $\delta = 0.5$, this means that there is an OVC which contributes with a 50% of the 0.8 load while the remaining load is uniformly provided by the other 9 OVCs.

Figure 8.2 shows the results changing the relative load from $\delta = 0.1$ to $\delta = 1$ and overall load from $\rho = 0.6$ to $\rho = 1$. It is clear that when $\delta = 1$, no contentions are possible (i.e., conflict-free configuration) and therefore the PLR is 0. Decreasing δ, the PLR increases. The increase strongly depends on the overall load. At $\rho = 1$, the curve is practically flatten with an elbow close to $\delta = 0.95$. At $\rho = 0.7$ and $\delta = 0.3$, the PLR is less than 10^{-7}.

It has to be underlined that the same behavior can be observed for whatever number of OVCs L contending for the same output wavelength. We carried out different simulations varying L from 2 to 20 and the obtained results present negligible differences.

This observation outlines that the configuration of the forwarding table has a strong impact on the switch performance. Previous works do not consider this issue and always assume an average situation where the OVCs are already established and fixed in the simulations [21] [23] [22]. But we just noticed that the OVC setup

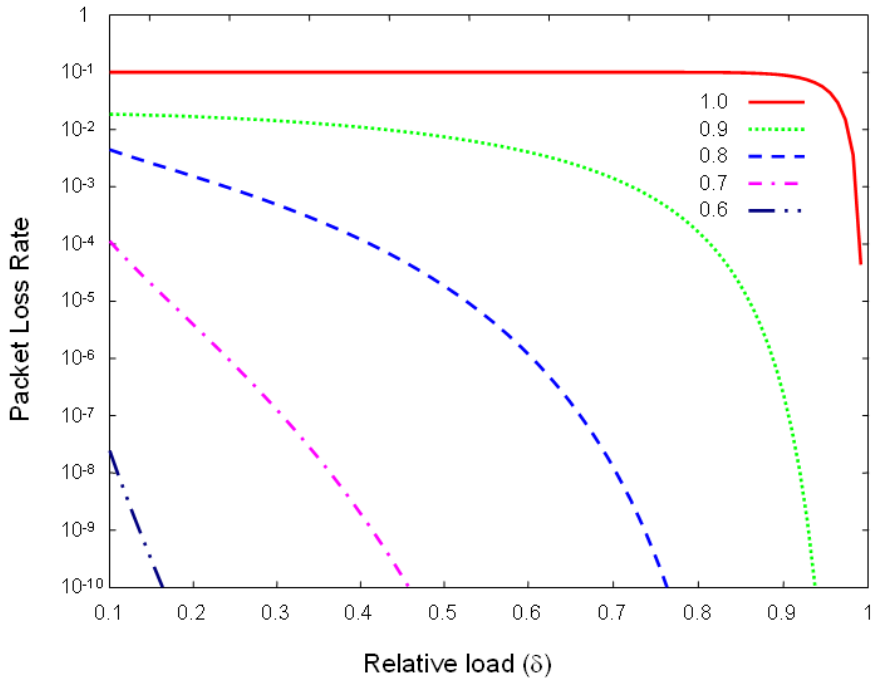

Figure 8.2: Packet loss rate as a function of the relative load at different overall load.

procedure is an interesting problem to be addressed. When a request to setup a new OVC arrives to a switch, this procedure is in charge of determining both the output port n^{out} and the output wavelength λ^{out}. While the former depends on the routing protocol, the latter may be set locally by each node using a *OVC-to-wavelength setup assignment* (OWSA) algorithm.

In the following section, we concentrate only on the OSWA problem proposing some heuristic OWSA procedures. These procedures are evaluated considering a static wavelength selection policy. However, the performance using dynamic policies is also affected since a non optimal setup continuously requires updates of the forwarding tables, moving the OVCs from the original wavelength to another. This clearly overloads the control functions, and increases the probability of breaking the correct packet sequence.

8.2 OWSA algorithms

Four different algorithms for the OWSA problem, namely **Random**, **Round-Robin**, **Balance**, and **Grouping** are suggested:

- **Random** (RND). When a request to setup a new OVC arrives, the SCL recognizes the output port n^{out} to reach the next hop and selects a random wavelength λ^{out} of n^{out}. Then, the new entry is added to the forwarding table.
- **Round-Robin** (RR). In this case, the SCL maintains a set of pointers ptr; each one pointing to the last selected wavelength of each output port. When

a request to setup a new OVC arrives, the SCL recognizes the output port n^{out} to reach the next hop, increases by 1 the corresponding pointer $ptr_{n^{out}}$ and selects the pointed wavelength λ^{out} of n^{out}. Then, the new entry is added to the forwarding table.

- **Balance** (BLC). In this case, we assume that the setup request contains also an information on the average load of the OVC. The SCL uses this information to maintain a matrix $\mathbf{V}$, where each entry $\mathbf{V}_{i,j}$ indicates the overall load of the output wavelength i of output port j. At OVC setup request, the SCL determines the output port n^{out} and selects the wavelength λ^{out} with the minimum load, i.e., $\min_{\lambda^{out} \in W} \{V_{\lambda^{out}, n^{out}}\}$.

- **Grouping** (GRP). This algorithm tries to take benefit from the conflict-free configuration. At OVC setup request, the SCL recognizes the $(\lambda^{in}, n^{in}, n^{out})$ tuple of this OVC. Therefore, it searches in the forwarding table if there is another OVC l' with the same tuple. If l' exists, it selects the same λ^{out} assigned to l'. If not, it searches a wavelength λ^{out} with no assignments. If all wavelengths are already in use, SCL applies the BLC algorithm.

It has to be underlined that the idea of exploiting the conflict-free configuration has firstly been adopted in [21] to develop a dynamic contention resolution algorithm for connection-oriented OPS networks. Here we use the same concept to intelligent setup the OVC forwarding table at the nodes.

Figure 8.3 shows the steps of these algorithms. FT stands for Forwarding Table.

8.3 Performance evaluation

We carried out several simulations in order to evaluate the performance of the previous described OWSA algorithms. We set up the simulator (described in Appendix I) considering $N = 4$, $W = 16$, $C = 10$ Gbps, $L = 10$, a degenerate buffer $\mathbf{Q}_6$, and $D = 0.4$ (optimal value for static approach [21]). In this case we only evaluate the PLR. Under static approach, there is no possibility to break the packet sequence (OS is always 0%) neither to change the assigned wavelength (FO is always 0%).

Figure 8.4, Figure 8.5, and Figure 8.6 shows the PLR as a function of the offered load comparing the RND, RR, BLC, and GRP algorithms using $\mathbf{M}^U$, $\mathbf{M}^P$, and $\mathbf{M}^B$ traffic matrix, respectively.

We can see that the BLC and GRP algorithms outperform the other strategies. Confirming our expectations, the performance improves even more using the GRP algorithm.

Further simulations (not presented here), showed that either varying L from 2 to 20 or changing the traffic distribution, GRP always presents the best PLR.

Input: l, n^{in}, λ^{in}, n^{out}, FT, ptr, V.
Output: λ^{out}

λ^{out} = algorithm(...)
Add entry FT(l) = $(\lambda^{in}, n^{in}, \lambda^{out}, n^{out})$

RND

1. λ^{out} = random$(1, \ldots, W)$;

RR

1. $ptr_{n^{out}} = |ptr_{n^{out}} + 1|_W$;
2. $\lambda^{out} = ptr_n$.

BLC

1. Search $\min_{i \in W} \{V_{i,n^{out}}\}$;
2. $\lambda^{out} = i$.

GRP

1. Search l' in FT with $(\lambda^{in}(l'), n^{in}(l'), n^{out}(l')) = (\lambda^{in}(l), n^{in}(l), n^{out}(l))$;
2. If l' exists, $\lambda^{out} = \lambda^{out}(l')$;
3. If not, search in FT a λ^{out} not used;
4. If all wavelengths are used, apply BLC algorithm.

Figure 8.3: OWSA algorithms.

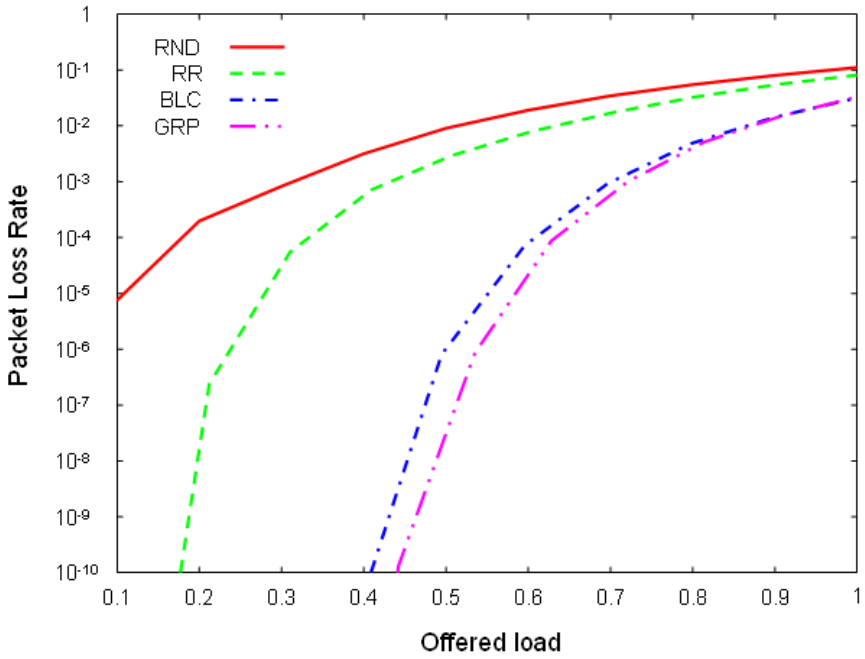

Figure 8.4: Packet loss rate as a function of the offered load comparing the RND, RR, BLC, and GRP algorithms under uniform traffic matrix.

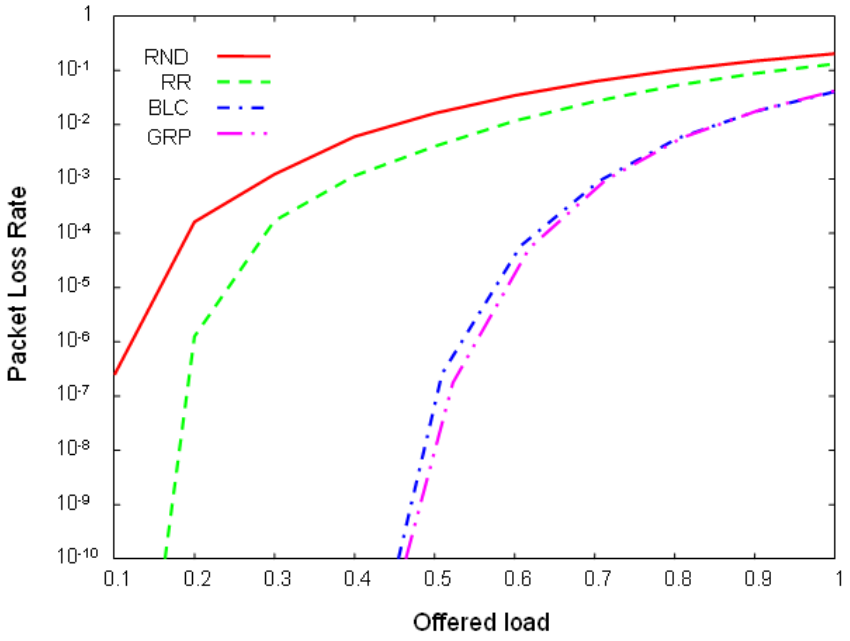

Figure 8.5: Packet loss rate as a function of the offered load comparing the RND, RR, BLC, and GRP algorithms under power-of-two traffic matrix.

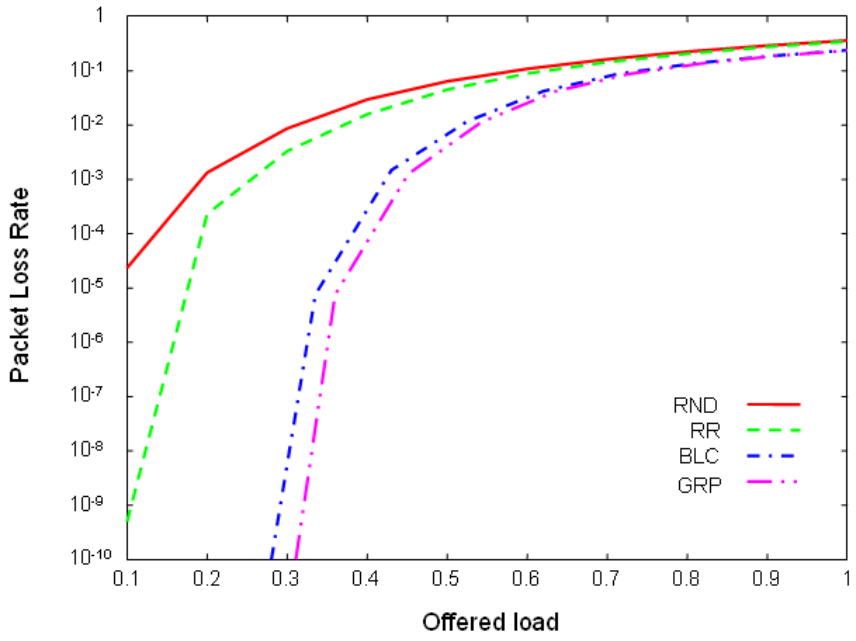

Figure 8.6: Packet loss rate as a function of the offered load comparing the RND, RR, BLC, and GRP algorithms under unbalanced traffic matrix.

8.4 Quality differentiation at the OVC setup

The previous results suggest that it is possible to obtain a clear performance improvements in terms of PLR simply applying intelligent OWSA algorithm. Basically, grouping the conflict-free flows and balancing the load yield the better results.

We can further exploit these facts to setup OVC with quality differentiation. For example let us consider that two quality OVC are available in the network, namely **High Quality** (HQ) and **Low Quality** (LQ). If a request to setup an OVC arrives to a node, we hence have two alternatives:

- If the request regards the establishment of an HQ OVC, the SCL applies the GRP algorithm previously explained since it performs the lowest PLR.
- For the LP OVC, at first, the SCL applies the first two steps of the GRP algorithm. These steps do not affect the HQ OVCs since no contentions are possible among conflict-free OVCs. If it does not find any l' with the same $(\lambda^{in}, n^{in}, n^{out})$ tuple of the new OVC, the SCL applies the BLC algorithm only between those wavelengths not used to transport HQ OVCs. If all wavelengths already transport at least one HQ OVC, the SCL applies the BLC algorithm as it is.

Figure 8.7 describes this quality differentiation setup procedure.

Eventually, it is possible to reject a request in order to not reduce the performance of the HQ OVCs already established. This possibility is not considered here and it is let to be investigated in further works.

When a request to setup an OVC l arrives

1. If the OVC is HQ, apply the GRP algorithm
2. If the OVC is LQ:
   (a) Search l' with $(\lambda^{in}(l'), n^{in}(l'), n^{out}(l')) = (\lambda^{in}(l), n^{in}(l), n^{out}(l))$;
   (b) If l' exists, $\lambda^{out}(l) = \lambda^{out}(l')$;
   (c) If not, search for the set of wavelengths $\Lambda \in n^{out}(l)$ without HP OVCs assigned;
   (d) If $\Lambda = \emptyset$, apply the BLC algorithm

Figure 8.7: Procedure for HQ and LQ OVC.

To evaluate the performance of this procedure, we set up the simulator considering $N = 4$, $W = 16$, $L = 10$, a degenerate buffer $\mathbf{Q}_6$, and $D = 0.4$ (optimal value for static approach). In this case we only evaluate the PLR. Under static approach, there is no possibility to break the packet sequence (OS is always 0) and to change the assigned wavelength (FT is always 100).

Figure 8.8 shows the PLR as a function of the offered load under uniform traffic matrix considering an HQ load of 5%, 25%, and 50% of the overall load, respectively.

We can see that there is a clear QoS differentiation between LQ and HQ. As expected, under low offered load it is possible to achieve good results. Increasing the HQ relative load, the performance gets worse, reaching 10^{-2} LQ packets losses at 0.5 load.

The same conclusions can be drawn considering Figure 8.9 which shows the PLR as a function of the offered load under power-of-two and unbalanced traffic matrix. In this case, the HP load is 25% of the overall load. In particular, we can observe that the unbalanced traffic matrix causes high performance degradations. We can suppose that this behavior is mainly due to the asymmetries of the traffic matrix which match very bad with the switch symmetries.

8.5 Summary

In this part of the thesis, we have addressed the problem of setting up the OVCs at the nodes assigning a proper wavelength to the connection.

An original policy called GRP based on grouping the conflict-free flows (i.e., flows coming from the same input wavelength) has been proposed and compared to common approach such as random, round-robin and load balancing techniques. Results have been demonstrated that considerable switch performance improvements can be obtained by this policy. For example, in the scenario studied in this work, the GRP algorithm yields a PLR one order of magnitude lower than load balancing when the switch is lightly loaded. This concept has been efficiently inferred to provide quality differentiation between two type of OVCs.

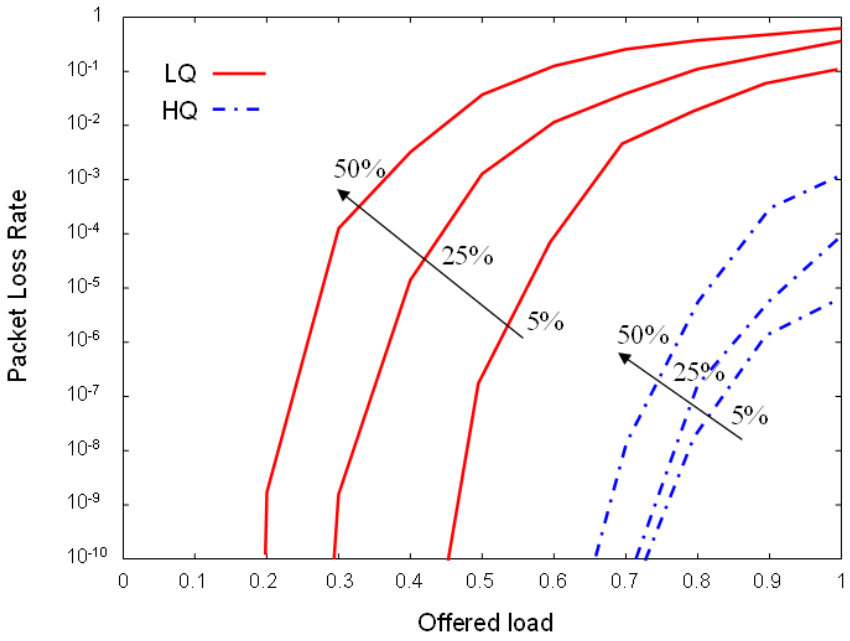

Figure 8.8: Packet loss rate as a function of the offered load. HQ load increases from 5% to 50% with respect to the overall load under uniform traffic matrix.

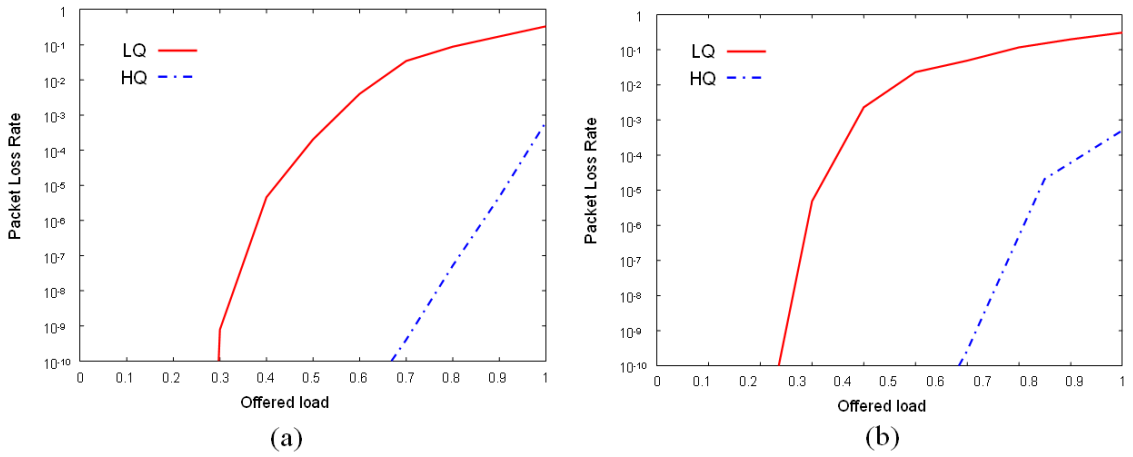

Figure 8.9: Packet loss rate as a function of the offered load with HQ load is 25% under (a) power-of-two traffic matrix and (b) unbalanced traffic matrix.

Chapter 9

QoS management in connection-oriented OPS networks

9.1 State-of-the-art

Recent studies have proposed enhanced contention resolution policies in order to provide quality-of-service (QoS) differentiation according to the DiffServ model [90]. The very limited queuing space and the impossibility of pre-emptying packets already buffered makes impossible the implementation of conventional fair queuing scheduling algorithms commonly used in electrical switches. At the same time, QoS management schemes must be kept very simple to be effective in OPS where each node must be able to schedule tens of Tbit/s of traffic.

Existing QoS schemes use basically the same method: 1) design a contention resolution algorithm which minimizes the Packet Loss Rate (PLR), thus 2) apply a QoS mechanism (some form of resources reservation on top of the contention resolution algorithm) able to differentiate the PLR among two or more classes. We can classify them as following:

- *Queueing priority* [67]. All incoming packets are stored into the buffer starting with the highest priority packets. If there are not free slots, remaining packets are dropped. This approach can be implemented only for synchronous networks where packets arrive at the same moment.

- *Threshold dropping*. Some resources (either wavelengths or delays) are reserved for higher priority class, while the rest is available for any class. From threshold dropping techniques are derived the following policies:

  - *Threshold-based priority policy* [19] [22]. A threshold value (which indicates that a possible congestion is imminent) is predetermined. If the network is not congested (queue occupancy above the threshold value) every packet is allowed to enter the queue.(i.e. no packet discard). When the queue length is longer than the threshold, only prior packets are allowed to enter the queue while low priority packets are discarded.

  - *Threshold-based with RED priority policy* [76]. The threshold mechanism can also be associated with a Random Early Detection (RED) strategy. When the buffer occupancy reaches the first threshold, the discard probability for lower priority packets begin to be mode than zero and its value is increased to 1 when the buffer occupancy reaches the second threshold.
  - *Wavelength-based priority policy* [18] [22] [12]. The lower priority packets can access only a subset of the wavelength resources and in any case they share it with higher priority packets.

- *Look-ahead* [41]. When a packet arrives, the header is extracted and the payload is delayed by a fixed amount of time at the input switch by means of an additional pool of optical buffers. This allows to the SCL to take scheduling decisions knowing a certain amount of packets arrivals. This solution to be effective needs additional hardware which means increase the cost. Also, the packet delay is affected since, to be effective, this scheme needs long FDLs as the results in [41] show.

- *Offset time-based* [106] [107] [41]. In this scheme, the edge node firstly sends a control packet and, after a given time called offset, sends the packet. The control packet is processed at each hop to determine the path of the packet. In order to provide QoS, different offset times are imposed at the edge node.

- *Use of electrical buffers* [12] [76]. In this case, to reduce the optical buffer requirements and make possible the use of random access, the optical packets can be converted in the electrical domain and stored electronically.

- *Queuing priority with overwriting* [56]. It is possible to build a complex buffer structure with two optical switches and a pool of FDLs in order to allow the overwriting of low-priority packets.

The problem of resource reservation schemes is that a fixed amount of resources is always reserved independently of the traffic profile. This implies a switch over-dimensioning with the relative cost increase. Moreover, such scheme does not present good enough results, for instance in [22] the PLR for low priority class is 10^{-2} with a load of 0.8. To achieve acceptable levels of PLR with this method, the scheduling requires very high computational complexity or very large optical memories.

The offset time method shows good results when applied to optical burst switching [107] where bursts comprise several packets. However, it seems not effective in OPS where the overhead of the control packets introduces considerable bandwidth wastage.

The hybrid E/O buffer method is not scalable with the bitrate since electronic devices cannot keep up with the speed of optical links and the E/O bottleneck is maintained.

The queueing priority with overwriting is costly and requires a computational-demanding scheduling. Indeed, no further investigations have been performed.

9.2 The Service Category-to-Algorithm Wavelength Selection technique

In this thesis, we propose a novel strategy able to improve the switch performance and provide the required QoS. It consists of defining different OPS service categories, like it was done in ATM networks, and the strategy is based on the fact that, in a QoS environment, it is not practical to provide the best handling to a traffic class that does not really require it. Therefore, if a set of K categories of service is available in the network, each category should be handled according to its requirements. For this reason we suggest to implement a set of K different handlings (i.e., algorithms) in the switches. When a packet belonging to an OVC with category i arrives to a switch, the SCL will execute the corresponding algorithm i to forward the packet which guarantees only the required service. We refer to this technique as *Service Category-to-Algorithm Wavelength Selection* (SCAWS).

9.2.1 Example of defining three different OPS service categories

For this study, we consider a system with the following three categories of service:

- **Best Effort** (BE) with no specific QoS requirements.
- **Loss Sensitive** (LS) for multimedia broadcasting applications which requires bounded losses;
- **Real Time** (RT) for interactive applications which requires strict performance (very low PLR and very short delay);

Others could be defined, but the point here is the definition of the different service categories, not the categories themselves.

We hence design three algorithms to be implemented in the SCL. The algorithms are the following:

- **Two-State Wavelength Selection** (TSWS) for supporting the BE category of service;
- **Losses Bounding Wavelength Selection** (LBWS) for supporting the LS category of service. It can also support the BE category when there are low LS connection demands;

  Sequence Keeping Wavelength Selection (SKWS) for supporting the RT category of service. It can also support the BE and LS categories when there are low RT connection demands.

The aim of the TSWS algorithm is to reduce the control overload (low FO) while maintains an acceptable level of the PLR. This algorithm tries to improve the performance of the static approach assigning two wavelengths to the OVC during the

When a packet belonging to OVC l arrives:

1. Look-up the forwarding table to determine $n^{out}(l)$ and wavelengths $\lambda_0^{out}(l)$ and $\lambda_1^{out}(l)$ assigned to l;
2. Determine the minimum delay D_j available for $\lambda_0^{out}(l)$ and $\lambda_1^{out}(l)$;
   (a) If $D_j > D_M$ (i.e., both wavelength are busy), the packet is lost;
   (b) If not, select the FDL j to send the packet to.

Figure 9.1: TSWS algorithms.

setup procedure (i.e., the GRP algorithm is executed twice). This assignment is kept constant all over the OVC life and single packets are always forwarded to the less congested wavelengths. This means that the wavelength searching step of the contention resolution algorithm is never needed (FO is always 0%).

Figure 9.1 shows the steps to follow when running this algorithm. Note that D_M is the maximum available delay provided by the FDLs (see Section 2.4.2 for more details).

The aim of the LBWS algorithm is to achieve a bounded PLR. Each OVC is assigned to a wavelength at setup using the GRP algorithm. This assignment may change if the OVC experiences a PLR above of a predetermined value R (*required PLR*). For this scope, a window T is defined. Every T the algorithm computes the PLR of each OVC. These PLRs are then ordered in descending way; starting from the higher value, the algorithm compares the PLRs with R, if it is higher, a new GRP algorithm is executed to reassign the OVC to another wavelength. Clearly, the value of T affects the switch performance: high values may not guarantee the required PLR; contrarily, low values can increase the control overload with an extreme situation of executing a new GRP algorithm per each incoming packet. It is important to notice that the value of R can be different from one OVC to another since their requirements can be distinct. For sake of simplicity, in this work we assume the same value for all OVCs that use this algorithm.

Figure 9.2 shows the steps to follow when running this algorithm.

The aim of the SKWS algorithm is to achieve excellent level of PLR maximizing the resource utilization and throughput. This is achieved taking per-packet decisions. At the same time, SKWS needs to control the delay preserving the correct packet sequence belonging to the same OVC. Indeed, since very short optical buffers are available in OPS networks, the delay is only due to the propagation delay and to rebuild the original information at the edge of the optical network. The latter may introduce considerable delays if extensive reordering operations are needed due to the out-of-order delivery of the packets [71]. For this reason, the design of the SKWS algorithm also considers the maintenance of the correct sequence of the packets belonging to the same OVC.

When a packet belonging to OVC l arrives:

1. Look-up the forwarding table to determine $n^{out}(l)$ and $\lambda^{out}(l)$ assigned to l;
2. Determine the minimum delay D_j available for $\lambda^{out}(l)$;
   (a) If $D_j > D_M$, the packet is lost;
   (b) If $D_j \leq D_M$, select the FDL j to send the packet to.

Every window T:

1. Determine the PLR of the OVCs: PLR^T;
2. Order (descending) the PLR^T;
3. From $l = 1$ to $L \times W \times N$:
   (a) If $PLR_l^T > R$, execute a new GRP algorithm

Figure 9.2: LBWS algorithms.

For the purpose of this work, given a stream of ordered packets at the switch input, we define the packet n to be out-of-order when the first bit of packet n leaves the switch before the last bit of packet $n-1$. This may happen in general in a DWDM environment that is equivalent to a bunch of links in parallel. Strictly speaking the packet sequence is also maintained as long as the first bit of packet n does not leave the switch before the first bit of packet $n-1$, meaning that the head of packet n is allowed to partially overlap with the tail of packet $n-1$. Unfortunately such less restrictive case is more difficult to control, especially when considering a cascade of switches. In fact, taking into account that, in general, the optical packets can aggregate more than one IP packet, the relative position of subsequent IP packets included in two subsequent optical packets cannot be controlled if overlapping is permitted. Therefore, a strict sequence keeping (i.e. avoiding packet overlapping) represents the unique procedure that assure the maintenance of sequence both at the optical packet level and at the IP packet level. Consequently in this work we have adopted this procedure.

To keep the correct packet order, the SCL stores the time-stamps t^{out} (one per each OVC) at which the last bit of the last packet is scheduled to leave the switch. This time is calculated as the sum of the packet arrival time, its duration and the delay assigned in the buffer. When a packet belonging to the OVC l arrives, the SCL recalls the time $t^{out}(l)$ and determine if the new packet needs additional delay to keep the order. This delay must be equal at least as long as the residual transmission time of the previous packet belonging to the same OVC l. Due to the discrete number of delays provided by the optical buffer, the additional delay is calculated as the integer multiple of D greater than $t^{out}(l)$.

Figure 9.3 shows the steps to follow when running the SKWS algorithm.

When a packet with duration d belonging to OVC l arrives at time t:

1. Look-up the forwarding table to determine $n^{out}(l)$, $\lambda^{out}(l)$ and $t^{out}(l)$ assigned to l;
2. If $t^{out}(l) \geq t$, the previous packet has already left the node:
   (a) If $\lambda^{out}(l)$ is busy, search for the set of wavelength $\Lambda \in n^{out}(l)$ not busy:
   (b) If $\Lambda = \emptyset$, the packet is lost;
   (c) If $\Lambda \neq \emptyset$, determine the wavelength w with the minimum delay D_j and select the FDL j to send the packet on;
3. If $t^{out}(l) < t$, the previous packet is still in the node:
   (a) Calculate the minimum delay D_{min} to add to the packet to preserve the order $D_{min} = \left\lceil \frac{t^{out}(l)-t}{D} \right\rceil D$;
   (b) If $\lambda^{out}(l)$ cannot provide D_{min}, search the set of FDL F able to provide a delay $D_j \geq D_{min}$
   (c) If $F = \emptyset$, the packet is lost;
   (d) If $F \neq \emptyset$, select the FDL with minimum D_j, if more than one is available, select the wavelength w that introduces the minimum gap between two subsequent packets;
4. Update the forwarding table $t^{out}(l) = t + d + D_j$.

Figure 9.3: SKWS algorithms.

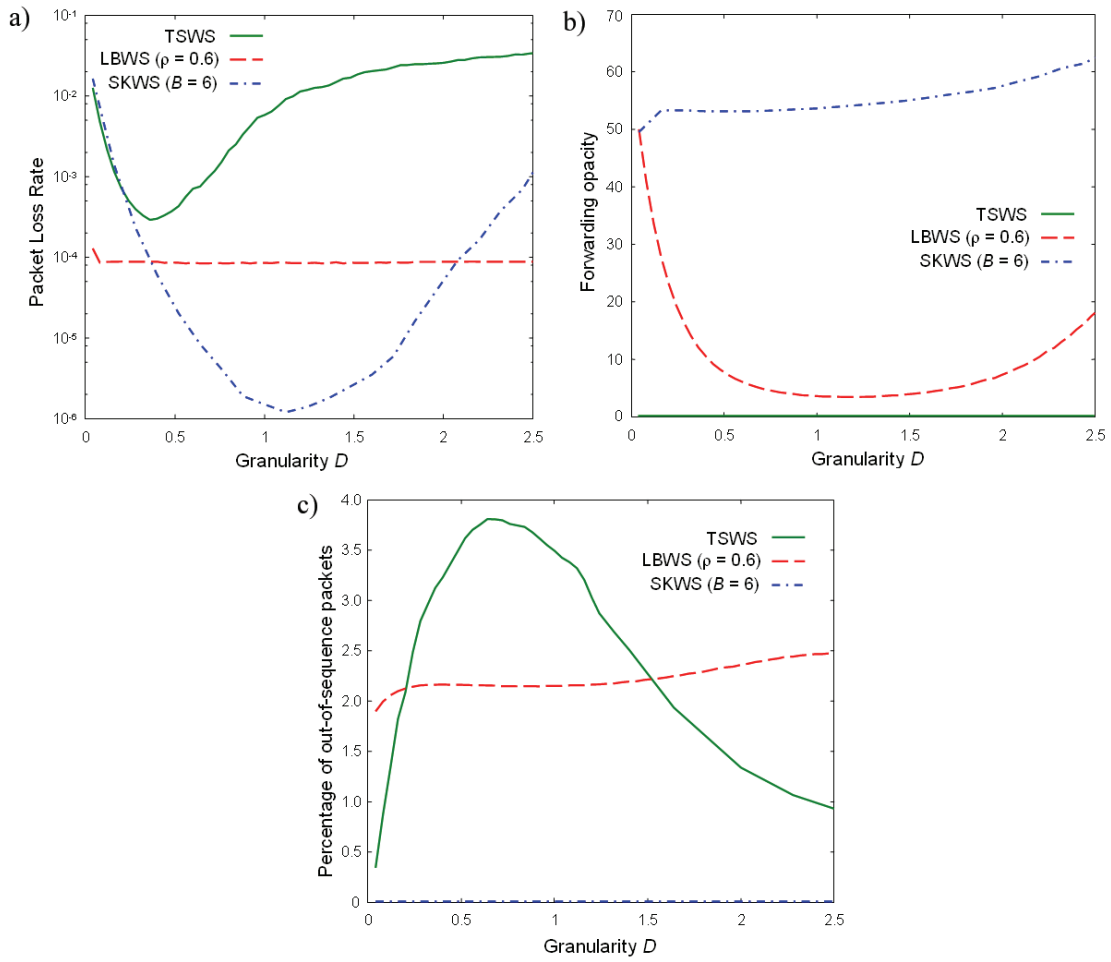

Figure 9.4: (a) Packet loss rate, (b) Forwarding opacity, and (c) Out-of-sequence packets as a function of D normalized to the average packet duration, comparing TSWS, LBWS and SKWS

9.2.2 Performance evaluation

In the this section, the algorithms are evaluated separately in order to find their specific characteristics. Afterwards, we integrate them in the same switch and evaluate the SCAWS technique.

In the following figures, we set up the simulator (described in Chapter 7) with $N = 4$, $W = 16$, $C = 10$ Gbps, and $L = 3$. The buffer configuration is a degenerate buffer $\mathbf{Q}_8$ (i.e., the length is $B = 8$) except for SKWS which uses a shorter buffer $\mathbf{Q}_6$. The offered load is $\rho = 0.8$, except for LBWS where it is $\rho = 0.6$ because it is not possible to bounding the PLR of high amount of traffic maintaining an acceptable control complexity. R is set to 10^{-4} and T to 20 D which are reasonable values offering a good trade-off between complexity and PLR.

Figure 9.4 shows the simulation results when independently evaluating the three algorithms.

Figure 9.4(a) plots the PLR as a function of D normalized to the average packet duration, comparing the TSWS, LBWS and SKWS algorithms. In this figure we can see that SKWS achieves the better PLR of 10^{-6} with $D = 1.2$. Contrarily to the usual concave behavior shown by other algorithms, LBWS exhibits constant values less than 10^{-4} which is the value set as required. TSWS presents the worst PLR but it is important to remark that its aim is to have low control complexity.

Figure 9.4(b) plots the FO measure comparing the TSWS, LBWS, and SKWS. It is clear that SKWS imposes the higher overload on the switch control; while LBWS shows low computational requirements reaching values close to 4%. The LBWS curve indicates that keeping bounded PLR require less computations for value of D ranging between $D = 1$ and $D = 1.4$, with a minimum in $D = 1.2$. Finally, TSWA does not need to reconfigure its OVC-to-wavelength assignment; therefore FO is always 0%.

Figure 9.4(c) shows the percentage of out-of-sequence packets comparing the TSWS, LBWS, and SKWS algorithms. As expected, SKWS maintains the correct sequence delivering. LBWS presents values around $2 \div 2.5\%$ while TSWS exhibits a concave behavior with a maximum of 3.7% in $D = 0.7$.

9.2.3 Optical buffer architecture to integrate different SCAWS

The results previously obtained assess the goodness of the proposed algorithms indicating that their aims have been fully accomplished: TSWS imposes low control overload and reaches acceptable PLR; LBWS requires low control overload and is able to guarantee a bounded PLR; finally, SKWS requires high control overload but achieves very good PLR maintaining the correct order of the packet sequence. The next step is hence the integration of these algorithms in the same SCL and the verification of the mutual impacts on the performance measures.

The integration is not trivial because the previous results also indicate that the algorithms achieve the better performance with different values of the fiber granularity D, the optimum D for LBWS and SKWS is 1.2 while it is 0.4 for TSWS (see Figure 9.4(a) and Figure 9.4(b).

Note that the rate between these two values ($D = 1.2$ and $D = 0.4$) is exactly 3. Exhaustive simulations (not presented here for lack of space) show that this peculiar factor of 3 is valid for whatever traffic matrix. It only depends on the traffic characteristics like average packet size. Based on this factor, the integration of the different contention resolution algorithms can be done using the following buffer architecture. Firstly, we fix $D = 0.4$ and set up two degenerate buffers: $\mathbf{Q}'$ with $D_j = jD$ delays and length B' for the BE packets and $\mathbf{Q}''$ with $D_j = 3jD$ delays and length B'' for RT and LS packets. Then, these buffers are merged in a non-degenerate buffer $\mathbf{Q} = \mathbf{Q}' \cup \mathbf{Q}''$ in such a way that the delays that are common in $\mathbf{Q}'$ and $\mathbf{Q}''$ are available for any category. Figure 9.5 shows an example with $B' = B'' = 4$, and a resulting length $B = 6$ of buffer $\mathbf{Q}$.

For the evaluation under multi-category, we set $N = 4$, $W = 16$, $C = 10$ Gbps, $\rho = 0.8$, $L = 3$, and, finally, the required PLR and measure window for LS packets to $R = 10^{-5}$ and $T = 20\ D$, respectively. Regarding the distribution of traffic, in Figure 9.6, Table 9.1 and Figure 9.7 we assume that 50% of the OVCs transport BE

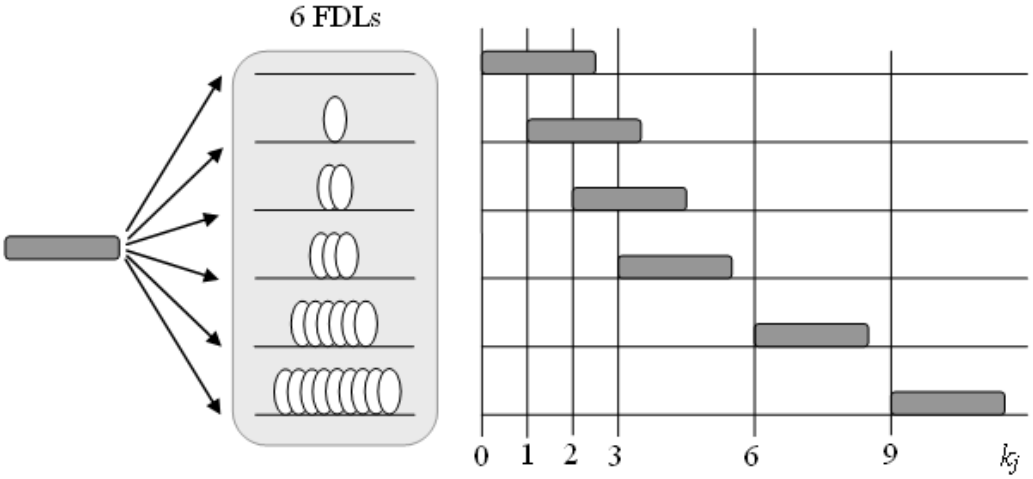

Figure 9.5: Non-degenerate buffer configuration with 6 FDLs. BE packets can use delays $\{0, D, 2D, 3D\}$, while the RT and LS packets can use delays $\{0, 3D, 6D, 9D\}$

packets, 30% transport RT packets, and the rest LS packets. In Figure 9.8 we analyze the PLR changing this distribution.

Figure 9.6 plots the PLR for the entire system as a function of D normalized to the average packet duration. In the figure, we include secondary x-axis which indicates the granularity perceived by SKWS and LBWS algorithms (exactly 3 times D). As expected, any categories of service achieves the optimal PLR in correspondence of $D = 0.4$. Hence, we use this value to obtain the following results.

In Table 9.1, we compare the SCAWS technique with the *Empty Queue Wavelength Selection* (EQWS) algorithm [21] - the best performed dynamic algorithm - and the *Minimum Gap* (MINGAP) algorithm [15] - the best performed connectionless algorithm. Both EQWS and MINGAP use the buffer threshold approach [22] to provide QoS (the values of D and of thresholds are those providing the lowest PLRs).

The results show that the SCAWS technique provides the lowest PLR for both LS and BE traffic. Moreover, as expected, the higher control complexity is required to forward the RT traffic (FO is 66.14%) while LS and BE impose low overload (5.93% and 0% respectively). In contrast, MINGAP imposes the same (very high) FO for any category, while EQWS requires higher FO for BE traffic which is an evident nonsense. At the same time, the packet sequence of RT traffic is preserved using the SCAWS technique, while it reaches 2% and 5% using EQWS and MINGAP, respectively. Previous studies [5] [65], confirm that even a small percentage of out-of-sequence (like that caused by EQWS algorithm) may impact harmfully on the network performance. We must also consider that this percentage is counted at the output of a single switch; by assuming n switch in series along a path this percentage increases accordingly.

Figure 9.7 plots the PLR as a function of the buffer depth B for any category. The results indicate that a significant improvement of the performance can be obtained with a small increase of the number of FDLs B of buffer $\mathbf{Q}$.

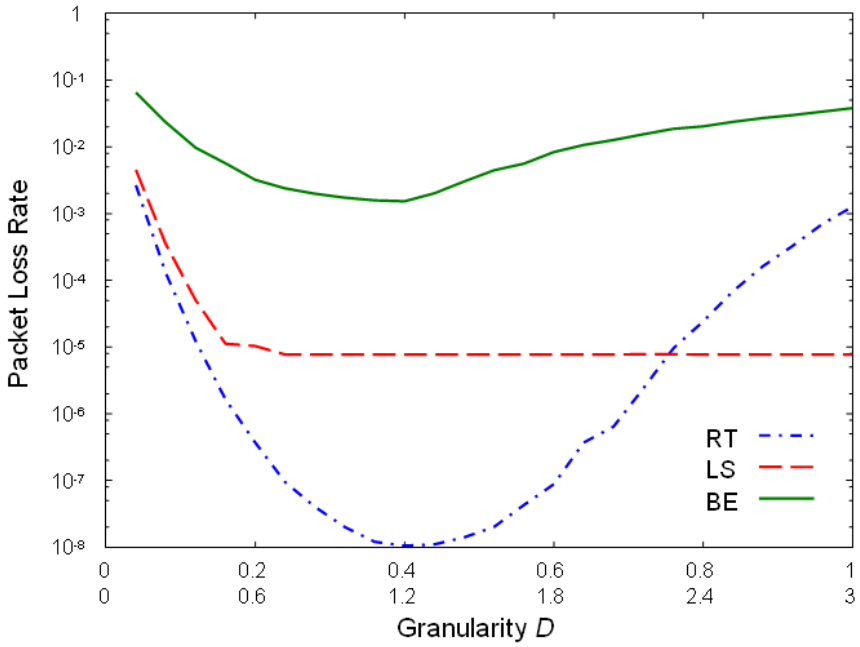

Figure 9.6: Packet loss rate as a function of D normalized to the average packet duration.

Table 9.1: PLR, FO and OS comparing SCAWS technique with EQWS and MINGAP both adopting a buffer threshold technique

	SCAWS		
Category	**PLR**	**FO**	**OS**
RT	$1.08\ 10^{-8}$	66.14%	0%
LS	$7.68\ 10^{-6}$	5.93%	1.76%
BE	$1.55\ 10^{-3}$	0%	3.29%
	EQWS		
RT	$3.00\ 10^{-8}$	16.20%	2.02%
LS	$2.75\ 10^{-4}$	30.82%	2.33%
BE	$5.24\ 10^{-2}$	52.51%	3.41%
	MINGAP		
RT	0	81.33%	5.39%
LS	$9.78\ 10^{-4}$	81.05%	5.03%
BE	$3.96\ 10^{-3}$	80.92%	4.62%

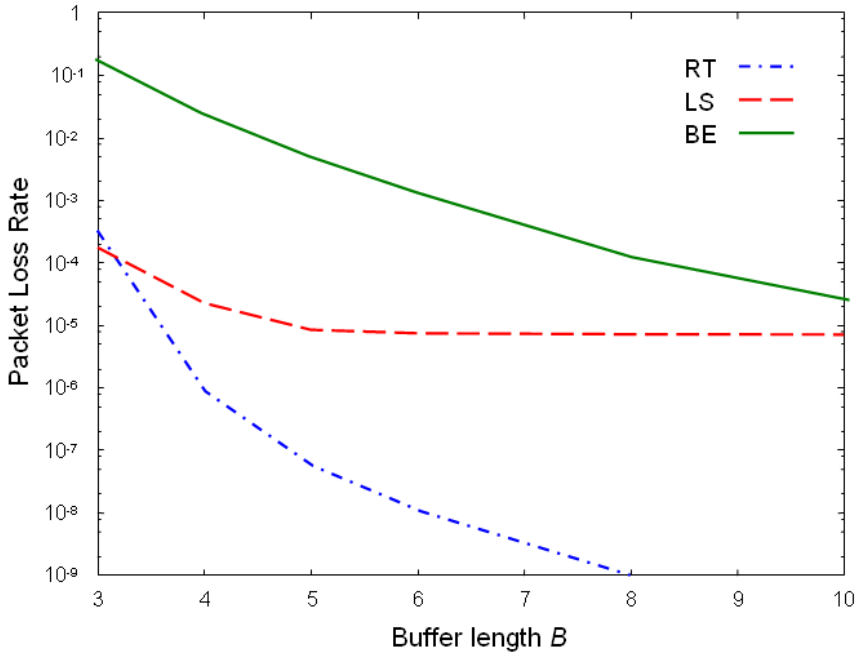

Figure 9.7: Packet loss rate as function of the buffer length B.

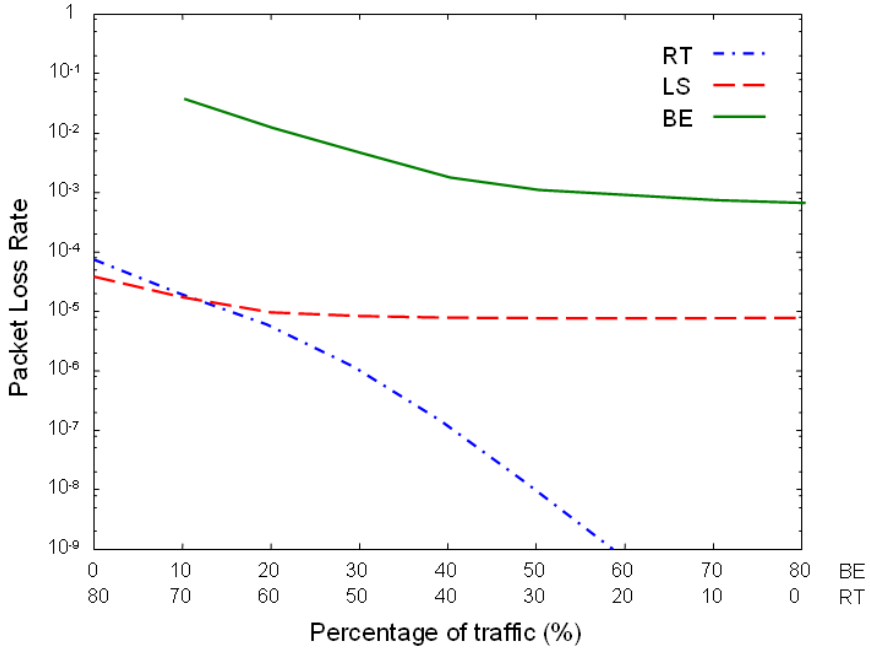

Figure 9.8: Packet loss rate as function of traffic relative load percentage.

Finally, Figure 9.8 shows the PLR changing the percentage of the relative load between the RT and BE traffic while maintaining fixed to 20% the relative load of LS traffic. This means that for instance, when the percentage of RT is 20%, the percentage of BE is 60%. We can see that the PLR of LS cannot be guaranteed if there is a high percentage of RT traffic (i.e., more than 60%). On the other side, if RT is not present, BE traffic is not able to fully exploit the switch capacity and the PLR remains relatively high. A way to improve the performance of the BE traffic when RT and LS present low loads is to apply the SKWS algorithm also to some BE OVCs. In this case, a smart algorithm should be developed in order to decide when, which, and how many BE OVCs can be forwarded according to the SKWS algorithm. This study is not developed here and is let for future investigations.

Future works will deal with the integration of the quality differentiation method with the SCAWS technique in order to obtain a more flexible environment. At the same time, SCAWS opens up future interesting developments on the routing problem for a whole network scenario.

PART IV

Conclusions and future works

Chapter 10

Conclusions and future works

This work has been focused on the optical packet switching paradigm, a long-term and more flexible alternative to the circuit-switched optical network currently being deployed by operators to support the extraordinary data traffic growth in the public domain. This innovative paradigm aims at optimizing the utilization of the WDM channels by means of fast and highly dynamic resource allocation based on a statistical multiplexing scheme, overcoming the typical inefficiency of the circuit transfer modes.

One of the open problem in such an environment is providing QoS schemes able to cope with the QoS requirements of the client networks. In this context, the dissertation specifically addresses such a problem using OPS technique for both wide area and metropolitan environments.

For the metropolitan environment and according to what developed in the electrical metro network, we have identified that 3 different classes of service must be supported, namely guaranteed service, priority service and best effort (like in standardized electrical metropolitan networks such as FDDI). Two network architectures based on composite topologies have been considered, namely the multi-PON and multi-ring networks. Both networks have been proposed within the European Project IST DAVID. Our contributions on this topic have concentrated into three different tasks per each network architecture:

1. extensive performance evaluation to identify the weakness of the architecture and the MAC protocol;

2. propose optimized mechanism to improve the performance;

3. design a mechanism to provide the missing service; the priority service for the multi-PON network and the guaranteed service for the multi-ring network.

Simulation results have demonstrated the effectiveness of the proposed mechanisms for both networks.

For the multi-PON architecture, the performance results have been obtained using a real scale simulator including self-similar traffic model and different traffic patterns between interconnected PONs. Two main weaknesses have been identified and

overtaken by optimized mechanisms. The validity of the proposals have been demonstrated, in fact both throughput and maximum end-to-end delay measures achieve better results than the original proposal. A QoS strategy for two classes of traffic has been also suggested. A good class differentiation has been achieved in a very robust way, which must be credited to the QoS strategy. In every scenario studied in this work, the achieved throughput for HP traffic perfectly matches its relative load percentage, at expense of the BE traffic. Nonetheless, the overall throughput is always close to 95%.

For the multi-ring architecture, two main drawbacks have been identified and improved by optimized mechanisms. The first one regarding the exploitation of the spatial reuse capability while the second one ability of overtaking congestion situations. The validity of the proposals have been demonstrated by numerical results. Finally we discussed the problem of allocating resources to provide guaranteed and best-effort services in different configurations of the multi-ring metro network. If the Hub performs ring-to-ring permutations, the Hybrid solution is preferable when the GS traffic is less than 50% of total resources. Whilst TD and FD solutions can be adopted when the bandwidth on network links is not a bottleneck. If the Hub performs wavelength-to-wavelength permutations, better guarantees can be offered to both best-effort and guaranteed services at the expense of increasing the complexity of the scheduling problem.

For the wide area environment, we have considered a connection-oriented OPS network. In such a scenario, we have identified two different problems: the first one regards the setup of the OVCs (OVC-to-wavelength setup assignment problem) and the second one concerns the packet contention problem under QoS requirements.

For the former, original policies have been proposed. Results have been demonstrated that considerable switch performance improvements can be obtained by grouping the conflict-free flows (i.e., flows coming from the same input wavelength). For example, in the scenario studied in this thesis, the GRP algorithm yields a PLR one order of magnitude lower than balancing the OVC load when the switch is lightly loaded. This concept has been efficiently inferred to provide quality differentiation between two type of OVCs, namely high-quality and low-quality OVCs.

For the latter, the novel SCAWS (Service Category-to-Algorithm Wavelength Selection) technique has been designed to provide QoS. In particular, we have defined a system with three different OPS service categories based on three different contention resolution algorithms. An ad-hoc buffer architecture has been designed to coordinate and optimize the behavior of the system. The obtained results highlight its goodness compared to other approaches (i.e., the EQWS and MINGAP algorithms using buffer threshold technique). For example in the studied scenario, the loss-sensitive service achieves 7.68×10^{-6} packet loss rate and a forwarding opacity (estimation of the complexity) of 5.93% which present better results than those obtained with MINGAP and buffer threshold approach: 9.74×10^{-4} and 81.05%, respectively.

The results obtained in this dissertation provide good inputs for further investigations. On the metro side, composite topologies prove that they are very suitable for coping with very large amount and heterogenous traffic. Nonetheless, recovery mechanisms to provide survivability issues are more complex. In both architecture

studied in this thesis, the Hub results a critical point. Duplication of the Hub may result in a excessive and costly solution and therefore coordination strategies among nodes and Hub must be developed.

The same issue is also valid on the wide area side. Despite of its importance, very few (no) works have been focused on the survivability in OPS networks. In this context, the study of routing problems in a global optical network scenario is of prime importance.

From the experience obtained in this thesis we recently start to focus on the Optical Burst Switching (OBS) paradigm which is currently one of the more interesting technology. The idea of transmitting the control information prior to the burst allowing the intermediate nodes sufficient time to reconfigure the switching matrix opens up several possibilities to design effective QoS schemes.

Appendix A

Acronyms

AP	Absolute Priority
ASON	Automatic Switched Optical Network
ATM	Asynchronous Transfer Mode
AWG	Array Waveguide Grating
BE	Best Effort
CAC	Call Admission Control
CAPEX	Capital Expenditure
DAVID	Data And Voice Integration over DWDM
DWDM	Dense Wavelength Division Multiplexing
FDL	Fiber Delay Line
FEC	Forward Equivalent Class
FDDI	Fiber Distributed Data Interface
FLP-AO	Fixed Length Packets and Asynchronous Operation
FLP-SO	Fixed Length Packets and Synchronous Operation
FO	Forwarding Opacity
FTTH	Fibre To The Home
GMPLS	Generalized Multiprotocol Label Switching
GS	Guaranteed Service
HP	High Priority
HQ	High Quality
IETF	Internet Engineering Task Force
IP	Internet Protocol
ISP	Internet Service Provider
LA	Limited Attempts
LAN	Local Area Networks
LBWS	Losses Bounding Wavelength Selection
LOBS	Labeled Optical Burst Switching
LQ	Low Quality
LS	Loss Sensitive
LSP	Label Switched Path
MAC	Medium Access Control
MAN	Metropolitan Area Networks

MPLS	Multiprotocol Label Switching
OBS	Optical Burst Switching
OPEX	Operational Expenditure
OPS	Optical Packet Switching
OS	Out-of-Sequence packet
OVC	Optical Virtual Circuit
OWSA	OVC-to-Wavelength Setup Assignment
PLR	Packet Loss Rate
PON	Passive Optical Network
PSC	Passive Star Coupler
PWRN	Passive Wavelength Routing Node
QOWSA	QoS OVC-to-Wavelength Setup Assignment
QoS	Quality of Service
RAM	Random Access Memory
RED	Random Early Detection
RT	Real Time
RPR	Resilience Packet Ring
SCL	Switch Control Logic
SDH	Synchronous Digital Hierarchy
SKWS	Sequence Keeping Wavelength Selection
SOA	Semiconductor Optical Amplifier
TE	Traffic Engineering
TSWS	Two-State Wavelength Selection
TWC	Tunable Wavelength Converters
VLP-AO	Variable Length Packets and Asynchronous Operation
VLP-SO	Variable Length Packets and Synchronous Operation
WAN	Wide Area Networks
WDM	Wavelength Division Multiplexing
WDS	Wavelength and Delay Selection

Appendix B

Related publications

B.1 Papers

1. **D. Careglio**, J. Solé-Pareta, S. Spadaro, "Novel contention resolution technique for QoS support in connection-oriented optical packet switching", in *Proceedings of IEEE International Conference on Communications* (ICC 2005), Seoul, Korea, May 2005.

2. C. Develder, A. Stavdas, A. Bianco, **D. Careglio**, H. Lonsethagen, J. Fernandez-Palacios, R. Van Caenegem, S. Sygletos, F. Neri, J. Solé-Pareta, M. Pickavet, N. Le Sauze, P. Demeester, "Benchmarking and viability assessment of optical packet switching for metro networks", *IEEE/OSA Journal of Lightwave Technologies*, vol. 22, no. 11, Nov. 2004, pp. 2435–2451.

3. A. Bianco, **D. Careglio**, J. Finochietto, E. Leonardi, G. Galante, F. Neri, J. Solé-Pareta, S. Spadaro, "Multi-class scheduling algorithms for DAVID metro network", *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 8, Oct. 2004, pp. 1483–1496.

4. F. Callegati, **D. Careglio**, W. Cerroni, C. Raffaelli, J. Solé-Pareta, P. Zaffoni, "Keeping the packet sequence in optical packet-switched networks", in *Proceedings of 9th European Conference on Networks and Optical Communications* (NOC 2004), Eindhoven, The Netherlands, Jun. 2004.

5. A. M. Hill, **D. Careglio**, J. Solé-Pareta, A. Rafel, "Relative costs of WDM rings and PONs for metro optical packet networks", in *Proceedings of Photonic in Switching Conference 2003* (PS2003), Paris, France, Sep. 2003.

6. **D. Careglio**, J. Solé-Pareta, S. Spadaro, "Heuristics for providing guaranteed service in DAVID metro network", in *Proceedings of 29th European Conference on Optical Communications* (ECOC2003), Rimini, Italy, Sep. 2003.

7. S. Bjornstad, C. M. Gauger, M. Nord, E. Baert, F. Callegati, **D. Care-glio** et al., "Optical burst switching and optical packet switching", *Chapter 4 of Book Advanced Infrastructure for Photonic Networks - Extended Final Report of*

COST Action 266, pp. 115-154. R. Inkret et al., Editorial Faculty of Electrical, Engineering and Computing, University of Zagreb, Sep. 2003.

8. J. Solé-Pareta, X. Masip-Bruin, S. Sánchez-López, S. Spadaro, **D. Care-glio**, "Some open issues in the optical networks control plane", in *Proceedings of 5th IEEE International Conference on Transparent Optical Networks* (ICTON2003), Warsow, Poland, Jun. 2003.

9. **D. Careglio**, A. Rafel, J. Solé-Pareta, S. Spadaro, A.M. Hill, G. Junyent, "Quality of Service strategy in an optical packet network with a multi-class frame-based scheduling", in *Proceedings of 2003 International Workshop on High Performance Switching and Routing* (HPSR 2003), Torino, Italy, Jun. 2003.

10. S. Bjornstad, M. Nord, D. R. Hjelme, N. Stol, F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, C. M. Gauger, C. Develder, J. Cheyns, E. Van Breusegem, E. Baert, D. Colle, M. Pickavet, P. Demeester, M. Lackovic, **D. Careglio**, G. Junyent, M. Klinkowsi, M. Marciniak, M. Kowalewski, "Optical burst and packet switching: node and network design, contention resolution and quality of service", in *Proceedings of 7th International Conference on Telecommunications* (ConTEL2003), Zagreb, Croatia, Jun. 2003.

11. **D. Careglio**, J. Solé-Pareta, S. Spadaro, "Optical slot size dimensioning in IP/MPLS over OPS networks", in *Proceedings of 7th International Conference on Telecommunications* (ConTEL2003), Zagreb, Croatia, Jun. 2003.

12. F. Callegati, **D. Careglio**, W. Cerroni, J. Solé-Pareta, C. Raffaelli, P. Zaffoni, "Time-wavelength exploitation in MPLS over OPS networks", *MPLS worshop*, Girona, Spain, Mar. 2003.

13. **D. Careglio**, J. Solé-Pareta, S. Spadaro, G. Junyent, "Performance evaluation of interconnected WDM PONs metro networks with QoS provisioning", in *Proceedings of 7th IFIP Working Conference on Optical Network Design and Modelling* (ONDM2003), Budapest, Hungary, Feb. 2003.

14. **D. Careglio**, G. Giner, J. Solé-Pareta, S. Spadaro, G. Junyent, "Evaluacin de la red ptica metropolitana multi-anillo del proyecto DAVID" (in spanish), *XII Jornadas Telecom I+D*, Madrid/Barcelona/Valencia, Nov. 2002.

15. T. Cinkler, S. Bjornstad, **D. Careglio**, D. Colle, C. Gauger, M. Karasek, A. Kuchar, S. de Maesschalck, F. Matera, C. Mauz, M. Settembre, "On the evolution of the optical infrastructure - COST 266 views", in *Proceedings of 4th IEEE International Conference on Transparent Optical Networks* (ICTON2002), Warsow, Poland, Apr. 2002.

16. F. Callegati, **D. Careglio**, W. Cerroni, J. Solé-Pareta, "Assessment of packet loss for an optical feedback buffer node using slotted variable length packet and

heavy tailed traffic", in *Proceedings of 4th IEEE International Conference on Transparent Optical Networks* (ICTON2002), Warsow, Poland, Apr. 2002.

17. A. Kuchar, T. Cinkler, S. Bjornstad, **D. Careglio**, D. Colle, C. Gauger, M. Karasek, S. de Maesschalck, F. Matera, C. Mauz, M. Settembre, "COST 266 views on the development of advanced infrastructure for photonic network", *IST Optimist International Workshop on Trends of Technologies for Photonic Networks*, Torino, Italy, Feb. 2002.

18. **D. Careglio**, J. Solé-Pareta, S. Spadaro, "Performance evaluation of metro optical networks based on multiple WDM PONs interconnected by a PWRN", in *Proceedings of 2nd International Workshop on All-Optical Networks* (WAON2001), Zagreb, Croatia, Jun. 2001.

19. J. Solé-Pareta, **D. Careglio**, S. Spadaro, J. Masip, J. Noguera, G. Junyent, "Modelling and performance evaluation of a national scale switchless based network", *Lecture Notes in Computer Science*, vol. 1938, pp. 337–347, Oct. 2000.

B.2 Papers in revision

1. F. Callegati, **D. Careglio**, W. Cerroni, G. Muretto, C. Raffaelli, J. Solé-Pareta, P. Zaffoni, "Keeping the packet sequence in optical packet-switched networks", submitted to *Optical Switching and Networking Journal*.

2. **D. Careglio**, J. Solé-Pareta, S. Spadaro, "Service category definition for supporting QoS in connection-oriented optical packet switching", submitted to *IEEE/OSA Journal of Lightwave Technology*.

B.3 Project deliverables

1. Deliverable D4, "Requirements for burst/packet networks in core and metro supporting high quality broadband services over IP", *FP6-506760 NOBEL Project*, Jun. 2004.

2. Deliverable D101, "Network model validation and benchmarking", *IST-1999-11742 DAVID Project*, Jun. 2003.

3. S. Bjornstad et al, "Optical packet and burst switching", *Intermediate report of COST 266 Action*, Jul. 2002.

4. Deliverable D131, "Specification of management of multi-layer optical packet networks", *IST-1999-11742 DAVID Project*, Jun. 2002.

5. Deliverable D122, "Optimisation and traffic performance of optical router and MAC protocol", *IST-1999-11742 DAVID Project*, Jun. 2002.

6. Deliverable D121, "Traffic models for optical packet networks with quality of service differentiation", *IST-1999-11742 DAVID Project*, Jun. 2001.

7. Deliverable D6, "Network scenarios and requirements", *IST-1999-11387 LION Project*, Oct. 2000.

B.4 Other publications

1. M. Klinkowski, F. Herrero, **D. Careglio**, J. Solé Pareta, "Adaptive routing algorithms for optical packet switching network", in *Proceeding of 9th IFIP Working Conference on Optical Network Design and Modelling* (ONDM2005), Milan, Italy, Feb. 2005.

2. F. Herrero, **D. Careglio**, J. Solé Pareta, M. Klinkowski, "Algoritmos de enrutamiento para conmutación de paquetes ópticos", (in spanish), *XIV Jornadas Telecom I+D*, Madrid, Spain, Nov. 2004.

3. M. Klinkowski, **D. Careglio**, X. Masip-Bruin, S. Spadaro, S. Sánchez-López, J. Solé-Pareta, "A simulation study of combined routing and contention resolution algorithms in connection-oriented OPS network scenario", in *Proceedings of 6th IEEE International Conference on Transparent Optical Networks* (ICTON2004), Wroclaw, Poland, Jul. 2004.

4. S. Spadaro, J. Solé-Pareta, **D. Careglio**, K. Wajda, A. Szymanski, "Positioning of the RPR standard in contemporary operator environments", *IEEE Network*, vol. 18, no. 2, Mar./Apr. 2004, pp. 35–40.

5. S. Spadaro, M. Quagliotti, J. Solé-Pareta, **D. Careglio**, A. Manzalini, F. Saluta, R. Stankiewicz, A. Lason, J. Rzasa, "Teletraffic engineering method for intelligent optical networks", in *Proceeding of 8th IFIP Working Conference on Optical Network Design and Modelling* (ONDM2004), Ghent, Belgium, Feb. 2004.

6. M. Klinkowski, **D. Careglio**, M. Marciniak, J. Solé-Pareta, "Performance analysis of the simple prioritized buffering algorithm in optical packet switch for DiffServ Assured Forwarding", in *Proceedings of 5th IEEE International Conference on Transparent Optical Networks* (ICTON2003), Warsow, Poland, Jun. 2003.

7. S. Spadaro, J. Solé-Pareta, **D. Careglio**, K. Wajda, A. Szymanski, "Assessment of resilience features for DPT rings", in *Proceedings of Eurescom Summit 2002*, Heidelberg, Germany, Sep. 2002.

Bibliography

[1] M. Ajmone Marsan, A. Bianco, E. Leonardi, A. Morabito, F. Neri, "All-optical WDM multi-rings with differentiated QoS", *IEEE Communications Magazine*, vol. 37, no. 2, Feb. 1999, pp. 58–66.

[2] J. D. Angelopoulos, N. Leligou, H. Linardakis, A. Stavdas, "A QoS-sensitive MAC for slotted WDM metropolitan rings", in *Bianco, A., Neri, F. (eds.): Next Generation Optical Network Design and Modelling*, IFIP TC6 / WG6.10 Sixth Working Conference on Optical Network Design and Modeling (ONDM 2002), Torino, Italy Feb. 2002, pp. 3–16.

[3] A. Banarjee, J. Drake, J.P. Lang, B. Turner, K. Kompella, Y. Rekhter, "Generalized multiprotocol label switching: an overview of routing and management enhancements", *IEEE Communications Magazine*, vol. 39, no. 1, Jan. 2001, pp. 144–150.

[4] A. Banarjee, J. Drake, J.P. Lang, B. Turner, K. Kompella, Y. Rekhter, "Generalized multiprotocol label switching: an overview of signaling enhancements and recovery techniques", *IEEE Communications Magazine*, vol. 39, no. 7, Jul. 2001, pp. 144–151.

[5] J.C.R. Bennett, C. Patridge, "Packet reordering is not a pathological network behavior", *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, Dec. 1999, pp. 789–798.

[6] A. Bianco, E. Leonardi, M. Mellia, F. Neri, "Network controller design for SONATA - a large-scale all-optical passive network", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, Oct. 2000, pp. 2017–2028.

[7] A. Bianco, M. Bonsignori, E. Leonardi, F. Neri, "Variable-size packets in slotted WDM ring networks", in *Bianco, A., Neri, F. (eds.): Next Generation Optical Network Design and Modelling*, IFIP TC6 / WG6.10 Sixth Working Conference on Optical Network Design and Modeling (ONDM 2002), Torino, Italy Feb. 2002, pp. 167–182.

[8] A. Bianco et al., "Frame-based matching algorithms for input-queued switches", in *Proceedings of 2002 International Workshop on High Performance Switching and Routing* (HPSR2002), Kobe, Japan, May 2002.

[9] A. Bianco, G. Galante, E. Leonardi, F. Neri, "Measurement based resource allocation for interconnected WDM rings", *Photonic Network Communications*, vol. 5, no. 1, Jan. 2003, pp. 5–22.

[10] A. Bianco, D. Careglio, J. Finochietto, E. Leonardi, G. Galante, F. Neri, J. Solé-Pareta, S. Spadaro, "Multi-class scheduling algorithm for the DAVID metro network", *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 8, Oct. 2004, pp. 1483–1496.

[11] D.J. Blumenthal et al., "All-optical label swapping networks and technologies", *IEEE/OSA Journal of Lightwave Technology*, vol. 18. no. 12, Dec. 2000, pp. 2058–2075.

[12] S. Bjornstad et al., "Optical burst and packet switching: node and network design, contention resolution and quality of service", in *Proceedings of 7th International Conference on Telecommunications* (ConTEL2003), Zagreb, Croatia, Jun. 2003.

[13] F. Callegati, "Which packet length for a transparent optical network?", in *Proc. SPIE Symposium Broadband Networking Technology*, Dallas, TX, Nov. 1997.

[14] F. Callegati, "Optical buffers for variable length packtes", *IEEE Communications Letters*, vol. 4, no. 9, Sep. 2000, pp. 292–294.

[15] F. Callegati, W. Cerroni, G. Corazza, "Optimization of wavelength allocation in WDM optical buffers", *Optical Network Magazine*, vol. 2, no. 6, Nov. 2001, pp. 66–72.

[16] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "Dynamic DWDM exploitation in connection-oriented optical packet switches", in *Bianco, A., Neri, F. (eds.): Next Generation Optical Network Design and Modelling*, IFIP TC6 / WG6.10 Sixth Working Conference on Optical Network Design and Modeling (ONDM 2002), Torino, Italy Feb. 2002, pp. 151–166.

[17] F. Callegati, D. Careglio, W. Cerroni, J. Solé-Pareta, "Assessment of packet loss for an optical feedback buffer node using slotted variable length packet and heavy tailed traffic", in *Proceedigns of 4th IEEE International Conference on Transparent Optical Networks* (ICTON2002), Warsow, Poland, Apr. 2002.

[18] F. Callegati, G. Corazza, C. Raffaelli, "Exploitation of DWDM for optical packet switching with QoS guarantees", *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 1, Jan. 2002, pp. 190–201.

[19] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "DWDM for QoS management in optical packet switches", in *Proceedings of 2nd International Workshop on Quality of Service in IP networks* (QoS-IP2003), Milano, Italy, Feb. 2003.

[20] F. Callegati, D. Careglio, W. Cerroni, J. Solé-Pareta, C. Raffaelli, P. Zaffoni, "Time-wavelength exploitation in MPLS over OPS networks", *MPLS worshop*, Girona, Spain, Mar. 2003.

[21] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "Dynamic wavelength assignment in MPLS optical packet switches", *Optical Network Magazine*, vol. 5, no. 5, Sep./Oct. 2003, pp. 41–51.

[22] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "Wavelength and time domain exploitation for QoS management in optical packet switches", *Computer Networks*, vol. 44, no. 4, Mar. 2004, pp. 569–582.

[23] F. Callegati, D. Careglio, W. Cerroni, J. Solé-Pareta, C. Raffaelli, P. Zaffoni, "Keeping the packet sequence in optical packet-switched networks", in *Proceedings of 9th European Conference on Networks and Optical Communications* (NOC 2004), Eindhoven, The Netherlands, Jun. 2004.

[24] N. Caponio, A. M. Hill, F. Neri, R. Sabella, "Single layer optical platform based on WDM/TDM multiple access for large scale 'switchless' networks", *European Transactions in Telecommunications*, vol. 11, no. 1, Jan/Feb. 2000, pp. 72–82.

[25] D. Careglio, J. Solé-Pareta, S. Spadaro, "Performance evaluation of metro optical networks based on multiple WDM PONs interconnected by a PWRN", in *Proceedings of 2nd International Workshop on All-Optical Networks* (WAON2001), Zagreb, Croatia, Jun. 2001.

[26] D. Careglio, G. Giner, J. Solé-Pareta, S. Spadaro, G. Junyent, "Evaluacin de la red ptica metropolitana multi-anillo del proyecto DAVID" (in spanish), *XII Jornadas Telecom I+D*, Madrid/Barcelona/Valencia, Nov. 2002.

[27] D. Careglio, J. Solé-Pareta, S. Spadaro, "Performance evaluation of an optical metro network based on a multi-tree topology", *Research Report UPC-DAC-2002-40*, Sep. 2002.

[28] D. Careglio, G. Junyent, J. Solé-Pareta, S. Spadaro, A. Rafel, "Studies on advanced optical metro networks ", *Research Report UPC-DAC-2002-43*, Sep. 2002.

[29] D. Careglio, J. Solé-Pareta, S. Spadaro, G. Junyent, "Performance evaluation of interconnected WDM PONs metro networks with QoS provisioning", in *Proceedings of 7th IFIP Working Conference on Optical Network Design and Modelling* (ONDM2003), Budapest, Hungary, Feb. 2003.

[30] D. Careglio, J. Solé-Pareta, S. Spadaro, "Optical slot size dimensioning in IP/MPLS over OPS networks", in *Proceedings of 7th International Conference on Telecommunications* (ConTEL2003), Zagreb, Croatia, Jun. 2003.

[31] D. Careglio, A. Rafel, J. Solé-Pareta, S. Spadaro, A.M. Hill, G. Junyent, "Quality of Service strategy in an optical packet network with a multi-class frame-based

scheduling", in *Proceedings of 2003 International Workshop on High Performance Switching and Routing* (HPSR 2003), Torino, Italy, Jun. 2003.

[32] D. Careglio, J. Solé-Pareta, S. Spadaro, "Heuristics for providing guaranteed service in DAVID metro network", in *Proceedings of 29th European Conference on Optical Communications* (ECOC2003), Rimini, Italy, September 2003.

[33] G.-K. Chang, K.-I. Sato, D.K. Hunter, *IEEE/OSA Journal of Lightwave Technology, Special Issue on Optical networks*, vol. 18, no. 12, Dec. 2000.

[34] T.-K. Chang et al., "Implementation of STARNET: a WDM computer communications network", *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, Jun. 1996, pp. 824–839.

[35] D. Chiaroni et al., "First demonstration of an asynchronous optical packet switching matrix prototype for multiterabit-class routers/switches", in *Proceedings of 27th European Conference on Optical Communications* (ECOC 2001), Amsterdam, The Netherlands, Oct. 2001.

[36] D. Chiaroni, "Packet switching matrix: a key element for the backbone and the metro", *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 7, Sep. 2003, 1018–1025

[37] T. Cinkler et al., "On the evolution of the optical infrastructure - COST 266 views", in *Proceedings of 4th IEEE International Conference on Transparent Optical Networks* (ICTON2002), Warsow, Poland, Apr. 2002.

[38] M. Crovella, A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes", *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, Dec. 1997, pp. 835–846.

[39] R. Davey, A. Lord, D. Payne, "Optical networks: a pragmatic European operator's view", in *Proceedings of Optical Fiber Communication* (OFC2002), invited paper, Anaheim, CA, Mar. 2002.

[40] http://david.com.dtu.dk/

[41] C. Develder, J. Cheyns, M. Pickavet, P. Demeester, "Service differentiation mechanisms for variable length packets in an optical switch with recirculating FDL buffer", in *Proceedings of Photonic in Switching* (PS 2003), Versailles, France, Sep. 2003.

[42] C. Develder, A. Stavdas, A. Bianco, D. Careglio, H. Lonsethagen, J. Fernandez-Palacios, R. Van Caenegem, S. Sygletos, F. Neri, J. Solé-Pareta, M. Pickavet, N. Le Sauze, P. Demeester, "Benchmarking and viability assessment of optical packet switching for metro networks", *IEEE/OSA Journal of Lightwave Technologies*, vol. 22, no. 11, Nov. 2004, pp. 2435–2451.

[43] L. Dittman et al., "The IST project DAVID: a viable approach towards optical packet switching", *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 9, Sep. 2003, pp. 1026–1040.

[44] E. Mannie et al., "Generalized multi-protocol label switching architecture", *draft-ietf-ccamp-gmpls-architecture-07.txt*, May 2003.

[45] T.S. El-Bawab, J.-D. Shin, "Optical packet switching in core networks: between vision and reality", *IEEE Communications Magazine*, vol. 40, no. 9, Sep. 2002, pp. 60–65.

[46] C. Fan, M. Maier, M. Reisslein, "The AWG‖PSC network: a performance enhanced single-hop WDM network with heterogeneous protection", in *Proceedings of IEEE Infocom 2003*, San Francisco, CA, Apr. 2003.

[47] S. Floyd, V. Jacobson, "Random Early Detection gateways for congestion avoidance", *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, Aug. 1993, pp. 397–413.

[48] A. Fumagalli et al., "CORD: contention resolution by delay lines", *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 6, Jun. 1996, pp. 1014–1029

[49] A. Fumagalli, J. Cai, I. Chlamtac, "A token based protocol for integrated packet and circuit switching in WDM rings", in *Proceedings of IEEE Globecom 1998*, Sydeny, Australia, Nov. 1998, pp. 2339–2344.

[50] P. Gambini et al., "Transparent optical packet switching: network architecture and demostrators in the KEOPS project", *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 7, Sep. 1998, pp. 1245–1259.

[51] R. Gaudino, et al., "RINGO: a WDM ring optical packet network demonstrator", in *Proceedings of European Conference on Optical Communications* (ECOC2001), Amsterdam, The Netherlands, Oct. 2001.

[52] N. Ghani, S. Dixit, T.-S. Wang, "On IP-over-WDM integration", *IEEE Communications Magazine*, vol. 38, no. 3, Mar. 2000, pp. 72–84.

[53] N. Ghani, S. Dixit, T.-S. Wang, "On IP-WDM integration: a retrospective", *IEEE Communications Magazine*, vol. 41, no. 9, Sep. 2003, pp. 42–45.

[54] J. J. Gordon, "Long range correlation in multiplexed pareto traffic", in *Proceedings of International IFIP-IEEE Conference on Broadband Communications*, Montreal, Canada, Apr. 1996, pp. 28–39.

[55] P.B. Hansen, S.L. Danielsen, K.E. Stubkjear, "Optical packet switching without packet alignment", in *Proceedings of 24th European Conference on Optical Communications* (ECOC1998), Madrid, Spain, Sep. 1998.

[56] H. Harai, M. Murata, "Prioritized buffer management in photonic packet switches for DiffServ assured forwarding" in *Bianco, A., Neri, F. (eds.): Next Generation Optical Network Design and Modelling*, IFIP TC6 / WG6.10 Sixth Working Conference on Optical Network Design and Modeling (ONDM 2002), Torino, Italy Feb. 2002, pp. 231–245.

[57] A.M. Hill, F. Neri, *IEEE Communications Magazine, Special Issue on Optical switching networks: from circuits to packets*, vol. 39, no. 3, Mar. 2001.

[58] A.M. Hill, D. Careglio, J. Solé-Pareta, A. Rafel, "Relative costs of WDM rings and PONs for metro optical packet networks", in *Proceedings of Photonic in Switching Conference 2003* (PS2003), Paris, France, Sep. 2003.

[59] D.K. Hunter, W.D. Cornwell, T.H. Gilfedder, A. Franzen, I. Andonovic, "SLOB: a switch with large optical buffers for packet switching", *IEEE/OSA Journal of Lightwave Technology*, vol. 16, no. 10, Oct. 1998, pp. 1725–1736.

[60] D.K. Hunter, M.C. Chia, I. Andonovic, "Buffering in optical packet switches", *IEEE/OSA Journal of Lightwave Technology*, vol. 16, no. 12, Dec. 1998, pp. 2081–2094.

[61] D.K. Hunter et al., "WASPNET: a wavelength switched packet network", *IEEE Communications Magazine*, vol. 37, no. 3, Mar. 1999, pp. 120–129.

[62] D.K. Hunter, I. Andonovic, "Approaches to optical Internet packet switching", *IEEE Commununications Magazine*, vol. 38, no. 9, Sep. 2000, pp. 116–122.

[63] T. Inukai, "An efficient SS/TDMA time slot assignment algorithm", *IEEE Transactions in Communications*, vol. 27, no. 10, Oct. 1979, pp. 1449–1455.

[64] R. Inkret et al., *Advanced infrastructure for photonic networks - Extended final report of COST Action 266*, Editorial Faculty of Electrical, Engineering and Computing, University of Zagreb, Sep. 2003.

[65] S. Jaiswal, G. Iannacone, C. Diot, J. Kurose, D. Towsley, "Measurement and classificatiion of out-of-sequence packets in a tier-1 IP backbone", in *Proceedings of IEEE Infocom 2003*, San Francisco, CA, vol. 2, Mar. 2003, pp. 1199–1209.

[66] J.-P. Jue, M.S. Borella, B. Mukherjee, "Performance analysis of the Rainbow WDM optical network prototype", *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, Jun. 1996, pp. 945–951.

[67] M. Klinkowski, D. Careglio, M. Marciniak, J. Solé-Pareta, "Performance analysis of the simple prioritized buffering algorithm in optical packet switch for DiffServ Assured Forwarding", in *Proceedings of 5th IEEE International Conference on Transparent Optical Networks* (ICTON2003), Warsow, Poland, Jun. 2003.

[68] K. Kompella, Y. Rekhter, "LSP Hierarchy with Generalized MPLS TE", *draft-ietf-mpls-lsp-hierarchy-08.txt*, IETF draft, Sep. 2002.

[69] T. Koonen, G. Morthier, J. Jennen, H. de Waardt, P. Demeester, "Optical packet routing in IP-over-WDM networks deploying two-level optical labeling", in *Proceedings of European Conference on Optical Communications*, Amsterdam, The Netherlands, Oct. 2001.

[70] A. Kuchar et al., "COST 266 views on the development of advanced infrastructure for photonic network", *IST Optimist International Workshop on Trends of Technologies for Photonic Networks*, Torino, Italy, Feb. 2002.

[71] M. Laor, L. Gendel, "The effect of packet reordering i a backbone link on application throughput", *IEEE Network*, vol. 16, no. 5, Sep. 2002, pp. 28–36.

[72] T. T. Lee, L. Soung-Yue, "Parallel routing algorithm in Benes-Closs networks", in *Proceedings of IEEE Infocom 1996*, vol. 1, San Francisco, CA, Mar. 1996, pp. 279–286.

[73] N. Le Sauze, et al., "A novel, low cost optical packet metropolitan ring architecture", in *Proceedings of 27th European Conference on Optical Communications* (ECOC2001), Amsterdam, The Netherlands, Oct. 2001.

[74] M. Maier, M. Reisslein, A. Wolisz, "Towards efficient packet switching metro WDM networks", *Optical Networks Magazine*, vol. 3, no. 6, Nov. 2002, pp. 44–62.

[75] M.A. Marsan, A. Bianco, E. Leonardi, A. Morabito, F. Neri, "All-optical WDM multi-rings with differentiated QoS", *IEEE Communications Magazine*, vol. 37, no. 2, Feb. 1999, pp. 58–66.

[76] J. Masip, J. Solé-Pareta, S. Borgione, B. Bostica, M. Burzio, "Providing differentiated service in optical packet networks", in *Proceedings of 16th International Teletraffic Congress* (ITC16), Edinburgh, UK, Jun. 1999.

[77] M. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, "Achieving 100% throughput in an input-queued switched", *IEEE/ACM Transactions in Communications*, vol. 47, n. 8, Aug. 1999, pp. 1260–1267.

[78] E. Modiano, "WDM-based packet networks", *IEEE Communications Magazine*, vol. 37, no. 3, Mar. 1999, pp. 130–135.

[79] B. Mukherjee, "WDM-based local lightwave networks - Part I: single-hop systems", *IEEE Network*, vol. 6, no. 3, May 1992, pp. 12–27.

[80] B. Mukherjee, "WDM-based local lightwave networks - Part II: multihop systems", *IEEE Network*, vol. 6, no. 4, Jul 1992, pp. 20–32.

[81] M.J. O'Mahony, D. Simeonidou, D.K. Hunter, A. Tzanakaki, "The application of optical packet switching in future communication networks", *IEEE Communications Magazine*, vol. 39, no. 3, Mar. 2001, pp. 128–135.

[82] A. Okada, T. Sakamoto, Y. Sakai, K. Noguchi, M. Matsuoka, "All-optical packet routing by an out-of-band optical label and wavelength conversion in a full-mesh network based on a cyclic frequency AWG", in *Proceedings of Optical Fiber Communication Conference* (OFC2001), vol. 4, Anaheim, CA, Mar. 2001, pp. ThG5-1-ThG5-3.

[83] H. Papadimitriou, K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, New York, Dover, 1998.

[84] J. Solé-Pareta, D. Careglio, S. Spadaro, J. Masip, J. Noguera, G. Junyent, "Modelling and performance evaluation of a national scale switchless based network", *Lecture Notes in Computer Science*, vol. 1938, pp. 337–347.

[85] J. Solé-Pareta, X. Masip-Bruin, S. Sánchez-López, S. Spadaro, D. Careglio, "Some open issues to define routing functions for control plane in optical core networks", in *Proceedigns of 5th IEEE International Conference on Transparent Optical Networks* (ICTON2003), Warsow, Poland, Jun. 2003.

[86] C. Qiao, M. Yoo, "Optical burst switching (OBS) - a new paradigm for an optical Internet", *Journal on High Speed Networks*, vol. 8, no. 1, Jan. 1999, pp. 69–84.

[87] R. Ramaswani, K.N. Sivarajan, *Optical networks: a practical perspective*, Morgan Kaufmann Publishers, San Francisco, 2nd edition, 2002.

[88] R. Braden et al., "Integrated services in the Internet architecture: an overview", *IETF RFC 1633*, Jun. 1994.

[89] E. Crawley et al., "A framework for QoS-based routing in the Internet", *IETF RFC 2386*, Aug. 1998.

[90] S. Blake et al., "An architecture for differentiated services", *IETF RFC 2475*, Dec. 1998.

[91] E. Rosen et al., "Multiprotocol label switching architecture", *IETF RFC 3031*, Jan. 2001.

[92] D. Awduche et al., "Overview and principles of Internet traffic engineering", *IETF RFC 3272*, May 2002.

[93] IEEE 802.17, *Resilient Packet Ring Working Group*, http://grouper.ieee.org/groups/802/17/

[94] K.V. Shrikhande et al., "HORNET: a packet-over-WDM multiple access metropolitan are ring network", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, Oct. 2000, pp. 2004–2016.

[95] S. Spadaro, J. Solé-Pareta, D. Careglio, K. Wajda, A. Szymanski, "Assessment of resilience features for DPT rings", in *Proceedings of Eurescom Summit 2002*, Heidelberg, Germany, Sep. 2002.

[96] S. Spadaro, J. Solé-Pareta, D. Careglio, K. Wajda, A. Szymanski, "Positioning of RPR standard in contemporary operators' environment", *IEEE Network*, vol. 18, no. 2, Mar/Apr. 2004, pp. 35–40.

[97] W. Stallings, *Local and Metropolitan Area Networks*, Prentice Hall, 2000.

[98] A. Stavdas, S. Sygletos, M. O'Mahoney, H.L. Lee, C. Matrakidis, A. Dupas, "IST-DAVID: concept presentation and physical layer modeling of the metropolitan area network", *IEEE/OSA Journal of Lightwave Technology*, vol. 21, no. 2, Feb. 2003, pp. 372-383.

[99] L. Tančevski, S. Yegnanarayanan, G. Castañon, L. Tamil, F. Masetti, T. McDermott, "Optical routing of asynchrnous, variable length packets", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, Oct. 2000, pp. 2084–2093.

[100] R.E. Tarjan, *Data Structures and Network Algorithms*, Society for Industrial and Applied Mathematics, Pennsylvania, Nov. 1983.

[101] K. Thompson, G.J. Miller, R. Wilder, "Wide-area internet traffic patterns and characteristics", *IEEE Network Magazine*, vol. 11, no. 6, Nov./Dec. 1997, pp. 10–23.

[102] H. R. van As, "Media access techniques: the evolution towards terabit/s LANs and MANs", *Computer Networks and ISDN Systems*, vol. 26, no. 6-8, Mar. 1994, pp. 603-656.

[103] W. Willinger, M.S. Taqqu, R. Sherman, D.V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level", *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, Feb. 1997, pp. 71–86.

[104] L. Xu, H.G. Perros, G. Rouskas, "Techniques for optical packet switching and optical burst switching", *IEEE Communications Magazine*, vol. 39, no. 1, Jan. 2001, pp. 136–142.

[105] S. Yao, B. Mukherjee, S. Dixit, "Advances in photonic packet switching: an overview", *IEEE Communications Magazine*, vol. 38, no. 2, Feb. 2000, pp. 84–94.

[106] M. Yoo, C. Qiao, "Supporting multiple classes of service in IP over WDM networks", in *Proceedings of IEEE Globecom 1999*, Rio de Janeiro, Brazil, Dec. 1999, pp. 1023–1027.

[107] M. Yoo, C. Qiao, S. Dixit, "QoS performance of optical burst switching in IP-over-WDM networks", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, Oct. 2000, pp. 2062–2071.