Node design in optical packet switched networks

Nord, Martin; Dittmann, Lars

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

NODE DESIGN IN OPTICAL PACKET SWITCHED NETWORKS

Martin Nord

April 2005

Research Center COM
Technical University of Denmark

ABSTRACT

This thesis reports the results from the Ph.D. project “Node Design in Optical Packet Switched Networks”, carried out at Research Center COM, Technical University of Denmark. The study covers motivation, realisation and performance of the Optical Packet Switching (OPS) network paradigm, for use in a future telecommunication network layer.

The introduction discusses the rationale for introducing OPS in the optical layer. All building blocks needed to realise optical packet switches have been demonstrated, but component integration is needed to decrease cost, and make OPS a serious contender for the optical layer.

The two next chapters provide an overview of optical technology and analysis of OPS designs. Use of the wavelength domain for contention resolution in asynchronous operation is studied. Emphasis is put on Shared-Per-Node (SPN) contention resolution pools with Tuneable Wavelength Converters (TWCs) to combine low Packet Loss Rate (PLR) and reduced TWC count. A parallel design, passively separating and recombining the switching planes, is proposed to overcome scalability constraints, and to enable hybrid networks and migration scenarios. Furthermore, the proposed *SoftRSV* algorithm increases the efficiency of shared Fibre Delay Lines (FDL) buffer scheduling, thereby increasing flexibility in choice between FDLs and TWCs in the pool.

Input processing schemes for hybrids networks, supporting OPS and Optical Circuit Switching (OCS), are proposed and demonstrated. The design combines Class of Service (CoS) segregation, header erasure and the first switching stage using only an automatic polarisation controller and a TWC as active components.

Applying Quality of Service (QoS) differentiation to tailor network performance to the needs of different applications facilitates efficient dimensioning. This thesis compares QoS differentiation approaches and proposes efficient Access Restriction (AR) method for the SPN TWC design. Moreover, a number of AR QoS differentiation schemes, with different combinations of PLR and jitter differentiation, are proposed to take advantage of an optimum FDL and TWC mix in the SPN contention resolution pool, without making jitter sensitive applications suffer.

Optical Packet Switched Ring Network designs (OPSRN) for the Metropolitan Area Network (MAN) are proposed and experimentally verified. The distributed MAC protocol, *Asynchronous Insertion Priority Scheduling with Insertion Threshold* (*AIPSwIT*) enables Variable Length Packets (VLP), high throughput and a high degree of fairness, both for balanced and unbalanced traffic matrices.

RESUMÈ (IN DANISH)

Denne rapport beskriver resultaterne fra Ph.D. projektet ”Node Design in Optical Packet Switched Networks”, som er udført ved Forskningscenter COM, Danmarks Tekniske Universitet (DTU). Rapporten omfatter en motivation og beskriver realisering og ydelse af det optiske pakkekoblede netværksparadigme, til anvendelse i et fremtidigt lag i telekommunikationsnetværk.

Introduktionen indeholder en diskussion af rationalet bag introduktionen af pakkekobling i det optiske netværkslag. Alle nødvendige byggesten for at realisere optiske pakkeswitche er allerede demonstreret, men øget integration af komponenter er nødvendigt for at reducere pris og gøre optisk pakkekobling til et seriøst alternativ i det optiske netværkslag.

De to efterfølgende kapitler giver et overblik over optisk teknologi og analyse af designs af optiske pakkeswitchede net. Brugen af bølgelængdedomænet til at løse blokeringsproblemet undersøges. Der lægges særlig vægt på såkaldte ”Shared-Per-Node” (SPN) puljer af tunbare bølgelængdekonvertere til afhjælpning af blokeringsproblemet, hvilket kombinerer en lav pakketabssandsynlighed med et reduceret antal bølgelængdekonvertere. Et parallelt design, der passivt separerer og kombinerer switchplanerne, foreslås for at overkomme begrænsninger i skalérbarhed, og for at muliggøre en udvikling mod hybride netværk. Ydermere vises, at den foreslåede algoritme *SoftRSV* øger effektiviteten af hukommelses-schedulering med fælles hukommelse baseret på fiber-forsinkelseslinjer, der derigennem giver øget fleksibilitet i valget mellem forsinkelseslinjer og bølgelængekonvertere i puljen.

Metoder til processering i indgangene til hybride netværk, der understøtter både optisk pakkekobling og kredsløbskobling, foreslås og demonstreres. Designet realiserer opsplitning i serviceklasser, sletning af pakke-header samt det første trin i switchen, udelukkende ved hjælp af en automatisk polarisationskontrol og en tunbar bølgelængdekonverter som de aktive komponenter.

Ved at anvende differentiering af kvaliteten af service (QoS) som et middel til at skræddersy ydelsen til forskellige applikationer, kan et netværk dimensioneres effektivt. Denne rapport sammenligner metoder til differentiering af QoS og foreslår en effektiv ”Access Restriction” (AR) metode for SPN designet med tunbare bølgelængdekonvertere. Derudover foreslås et antal metoder til differentiering af QoS baseret på AR, med forskellige kombinationer af differentiering af pakketabssandsynlighed og jitter, med henblik på at udnytte en optimal fordeling mellem forsinkelseslinjer og tunbare bølgelængdekonvertere i SPN puljen, uden at det går ud over jitter-følsomme applikationer.

Optiske pakkekoblede netværk med ringtopologi (OPSRN) til Metro netværket (MAN) foreslås og verificeres eksperimentelt. Den distribuerede MAC protokol, *Asyncronous Insertion Priority Scheduling with Insertion Threshold (AIPSwIT)*, åbner mulighed for variabel pakkelængde, høj kapacitet og en høj grad af retfærdighed, både for balancerede og ubalancerede trafikmatricer.

ACKNOWLEDGMENENTS

I would like to start by expressing my gratitude towards Professor Lars Dittmann for supervising me. Thanks to Professor Villy B. Iversen for collaboration and discussions on teletraffic theory. Furthermore, I would like to thank the other members of the Networks Competence Area for a warm welcome and for organising a number of social activities. Thanks to Henning Christiansen, Ole Asmus and Anders Fosgerau for help on IT related issues. Thanks to Henrik Christiansen for introducing me to OPNET, and to José Soler for collaboration on the first OPNET projects. Also thanks to other Ph.D. students at COM, for scientific discussions in the corridors and social activities on conferences. Special thanks to Mads L. Nielsen and Martin N. Petersen, for collaboration in the lab, and for making my years in Copenhagen a wonderful experience- it has been a great pleasure for me to show you around in Copenhagen.

I am grateful towards the Research Council of Norway and Telenor R&D, Norway, for funding this project. From Telenor R&D, I would especially like to thank the head of the Network Infrastructure division, Nils Flaarønning, for his warm support during the preparation- and course of this project. Thanks to IP Network colleagues for their encouragements and support during my studies, in particular the group leader, Harald Pettersen, and those who I have worked with on optical networking issues: Steinar Bjørnstad, Torodd Olsen, Evi Zouganeli, Aasmund Sudbø, Bjørn Johan Slagsvold, and Asbjørn Kleivstul.

During the project I have had the pleasure of participating in the following European projects: Cost 266, IST DAVID, IST STOLAS, e-Photon/ONe and Cost 279. I would like to thank all project members for collaboration and discussions.

Alcatel Research and Innovation are acknowledged for supplying the state-of-the art devices for experiments on wavelength converters, and Beatrice Dagens for collaboration on resulting articles.

Special thanks to Steinar Bjørnstad and Harald Øverby for frequent discussions and collaborations on articles, in particular during their stays at COM during their Ph.D. studies, and for feedback on this thesis. I also very much appreciate the feedback from Mads L. Nielsen, Evi Zouganeli, and Villy B. Iversen.

Finally, warm thanks to my family and friends, for being supportive throughout my stay in Denmark, by keeping in touch and visiting me. Also thanks to the friends I have made here in Copenhagen. You have all helped me keeping a healthy perspective on life, even during the most hectic periods of this project.

ABBREVIATIONS

AA-MZI	All-Active Mach Zehnder Interferometer
ADSL	Asymmetric Digital Subscriber Line
AIPSwIT	Asynchronous Insertion Priority Scheduling with Insertion Threshold
AN	Access Network
APC	Automatic Polarisation Controller
APDP	Adaptive Preemptive Drop Policy
AQM	Active Queue Management
AR	Access Restriction
ASON	Automatic Switched Optical Network
AsySrvTS	Asymmetric Server Traffic Scenario
ATM	Asynchronous Transfer Mode
ATMR	Asynchronous Transfer Mode Ring
AWG	Array Waveguide Grating
BD	Buffer Discard
BE	Best Effort
BERT	BER Test-set
BPF	Band Pass Filter
BTB	Back-To-Back
CAPEX	Capital Expenditures
CBR	Constant Bit Rate
CL	Connectionless
CO	Connection Oriented
CoS	Class of Service
CS	Circuit Switched
CW	Continuous Wave
DBQR	Distributed Queue Bidirectional Ring
DFB	Distributed Feedback
DiffServ	Differentiated Services
DM	Direct Mapping
DWRON	Dynamic Wavelength Routed Optical Networks
E/O	Electrical-to-Optical
ECL	External Cavity Laser
EDFA	Erbium-Doped Fibre Amplifiers
FB	Forwarding Blocking
FCFS	First-Come, First-Served
FDL	Fibre Delay Line
FI	Fairness Index
FIFO	First In, First Out
FIOWC	Fixed Input-and Output wavelength Wavelength Converter
FIWC	Fixed Input-wavelength Wavelength Converter
FLP	Fixed Length Packets
FOWC	Fixed Output-wavelength Wavelength Converter
FR	Fixed Receiver

FRF	Fixed Reflection Filter
FT	Fixed Transmitter
GMPLS	Generalised Multi-Protocol Label Switching
GS	Guaranteed Service
HLDLC	High-Level Data Link Control
HOL	Head Of Line
HP	High Priority
IB	Insertion Blocking
IF	Input Fibre
IntServ	Integrated Services
IP	Internet Protocol
IPD	Intentional Packet Dropping
ISP	Internet Service Provider
ITU	International Telecommunication Union
IW	Input Wavelength
IWL	Input Wavelength
JET	Just-Enough-Time
JF	Jitter Free
JT	Jitter Tolerant
LLUI	Link Load Uniformity Index
LNTS	Link Network Traffic Share
LP	Low Priority
LSP	Label Switched Path
LQF	Longest Queue First
m.p.d.	mean packet duration
MAN	Metropolitan Area Network
MZI	Mach Zehnder Interferometer
MIN WDS	Minimum Gap queue Wavelength and Delay Selection
MMR	Multi-Meta Ring
MOD	Modulator
MSS	Multiple Slot Sizes
MPLS	Multi-Protocol Label Switching
MTIT	Multi-Token Interarrival Time
O/E	Optical-to-Electrical
OBS	Optical Burst Switching
OCS	Optical Circuit Switching
OF	Output Fibre
OLS	Optical Label Switching
OPEX	Operational Expenditures
OPS	Optical Packet Switching
OPSRN	Optical Packet Switched Ring Network
OW	Output Wavelength
OWL	Output Wavelength
PBS	Polarisation Beam Splitter
PDP	Preemptive Drop Policy
PIC	Photonic Integrated Circuit

PJF	Partially Jitter Free
PLR	Packet Loss Rate
PM	Polarisation Maintaining
PPG	Pulse Pattern Generator
PPP	Point to Point Protocol
PRBS	Pseudo Random Bit Sequence
PS	Packet Switched
QoS	Quality of Service
RAM	Random Access Memory
RB	Receiver Blocking
REC	Receiver
RIX	Ring Interchanger
RN	Ring Node
RNB	Reconfigurably Non-Blocking
RR	Random Routing
RT	Real Time
RT	Relative Throughput
RTT	Round Trip Time
RX	Receiver
SAR	Segmentation And Reassembly
SDE	Service Differentiation Efficiency
SDH	Synchronous Digital Hierarchy
SFR	Slot Filling Ratio
SLA	Service Level Agreement
SOA	Semiconductor Optical Amplifier
SNB	Strictly Non-Blocking
SOP	State Of Polarisation
SPN	Shared Per Node
SPWP	Shared Per Waveband Plane
SRLF	Server Relative Load Factor
SRR	Synchronous Round Robin
SWRON	Static Wavelength Routed Optical Networks
SymSrvTS	Symmetric Server Traffic Scenario
TAG	Tell-And-Go
TAW	Tell-And-Wait
TBF	Tuneable Bandpass Filter
TCP	Transmission Control Protocol
TDM	Time Division Multiplexing
TE	Traffic Engineering
TR	Tuneable Receiver
TT	Tuneable Transmitter
TWC	Tuneable Wavelength Converter
UniformTS	Uniform Traffic Scenario
VLP	Variable Length Packets
VoD	Video on Demand
VoIP	Voice over IP

VOQ	Virtual Output Queue
WA	Wavelength Allocation algorithm
WAN	Wide Area Network
WC	Wavelength Conversion
WCR	Wavelength Conversion Ratio
WDM	Wavelength Division Multiplexing
WP	Waveband Plane
WR	Wavelength Routing
WRON	Wavelength Routed Optical Networks

TABLE OF CONTENT

PH.D. PUBLICATIONS

p1. S. Bjørnstad, M. Nord, and D. R. Hjelme. “Transparent optical protection switching scheme based on detection of polarisation fluctuations”, Techn. Digest OFC 2002, pp. 433-434, (Anaheim, CA, USA), 2002.

p2. M. Nord. “Optical Switching Technologies for optical line-, burst- and packet switches”, Telenor R&D Report 32/2002, (URL:www.telenor.com/rd/pub/publications/pub02/index.shtml), 2002.

p3. M. Nord. “Optical Switching Technologies for Fast Optical Packet Switching”, Proc. 4th IEEE International Conference on Transparent Optical Networks (ICTON 2002), vol. 1, pp. 241-244, (Warsaw, Poland), 2002.

p4. M. Nord, S. Bjørnstad, and C. M. Gauger. “OPS or OBS in the core network?”, Proc. 7th IFIP working conference on optical network design & modeling (ONDM 2003), pp. 179-198, (Budapest, Hungary), 2003.

p5. M. L. Nielsen, M. Nord, M. P. Nordal, B. Dagens, A. Labrousse, R. Brenot, B. Martin, S. Squedin, and M. Renaud. “40 Gb/s standard-mode wavelength conversion in all-active MZI with very fast response”, IEE Electronics Letters, 39(4), pp. 385, February 2003.

p6. M. L. Nielsen, M.P Nordal, M. Nord, and B. Dagens. “Compact All-Optical Parity Calculator Based on a Single All-Active Mach Zehnder Interferometer with All-SOA Amplified Feedback”, Techn. Digest OFC 2003, vol. 1, pp. 274-275, (Atlanta, GA, USA), 2003.

p7. M. Nord. “Node Design in Optical Packet- and Optical Burst Switching”, (invited paper) Proc. 5th IEEE International Conference on Transparent Optical Networks (ICTON 2003), pp. 136-143, (Warsaw, Poland), 2003.

p8. M. Nord. In: COST 266, Work Group 2 report: Optical Packet and Burst Switching (ed. S. Bjørnstad). ”Chapter 4 – Analysis of required switching times and identification of suitable switching technologies in optical packet switching”, (URL:www.ure.cas.cz/dpt240/cost266), 2003.

p9. M. Nord, and S. Bjørnstad. “DWDM in Metropolitan Area Networks” (ed. F. Matera), Chapter 2.3 in Extended Final Report of Cost Action 266: “Advanced Infrastructure for Photonic Networks”, (ed. R. Inkret, A. Kuchar, and B. Mikac), Published by Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, ISBN 953-184-064-4, (URL:www.ure.cas.cz/dpt240/cost266/index.html), 2003.

p10. M. Nord et al. “Optical Packet and Burst Switching” (ed. S. Bjørnstad, C. Gauger, and M. Nord), Chapter 4.1-4.2, in Extended Final Report of Cost Action 266: “Advanced Infrastructure for Photonic Networks”, (ed. R. Inkret, A. Kuchar, and B. Mikac), Published by Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, ISBN 953-184-064-4, (URL:www.ure.cas.cz/dpt240/cost266/index.html), 2003.

p11. S. Bjørnstad, M. Nord, and D. R. Hjelme. “QoS differentiation and header/payload separation in optical packet switching using polarization multiplexing”, Proc. ECOC 2003, pp. 28-29, paper Mo 3.4.6, (Rimini, Italy), 2003.

p12. M. Nord, and M. Berger. ”Other related cost studies: power consumption”. Chapter 2.4.6 in IST DAVID Deliverable D101: Network concepts validation and benchmarking, (ed. N. Le Sauze), 2003.

p13. M. L. Nielsen, J. D. Buron, M. Nord, and M. N. Petersen. “SOA-based functional devices for future optical networks”, (invited paper) Proc. International Conference on Solid State Devices and Materials (SSDM), F-9-1, (Tokyo, Japan), 2003.

p14. M. Nord, S. Bjørnstad, M. L. Nielsen, and B. Dagens. “Novel strictly non-blocking Node Designs for asynchronous OPS MAN”, Proc. Photonics in Switching (PS’2003), PD paper 1, published by La Société de l'Electricité, de l'Electronique, et des Technologies de l'Information et de la Communication, (Versailles, France), 2003.

p15. M. Nord. “Multi-plane waveband based optical packet switch design with partially shared wavelength conversion”, Proc. 8th IFIP working conference on Optical Network Design & Modelling (ONDM 2004), (Ghent, Belgium), 2004.

p16. M. Nord, S. Bjørnstad, M. L. Nielsen, and B. Dagens. “Demonstration of optical packet switching scheme for header-payload separation and class-based forwarding”, Techn. Digest OFC 2004, vol. 1, pp. 563-565, paper TuQ2, (Los Angeles, CA, USA), 2004.

p17. S. Bjørnstad, D. R. Hjelme, N. Stol, and M. Nord. “A packet switched hybrid optical network with service guarantees”, Proc. 16th Norwegian Electro-Optics Meeting 2004, (Tønsberg, Norway), 2004.

p18. M. Nord. “Optical packet switch design with relaxed maximum hardware parameters and high service-class granularity for flexible switch node dimensioning”, Proc. 11th International Telecommunication Network Strategy and Planning Symposium (networks 2004), pp. 453-458, (Vienna, Austria), 2004.

p19. M. Nord. “An efficient Quality of Service differentiation algorithm for buffer-less optical packet switches with partially shared wavelength converters”, Proc. 9th IEICE OptoElectronics and Communications Conference/ 3rd Conference on Optical Internet (OECC/COIN 2004), pp. 586-587, (Yokohama, Japan), 2004.

p20. M. Nord. “Performance analysis of a low-complexity and efficient QoS differentiation algorithm for bufferless optical packet switches with shared wavelength converters in asynchronous operation”, Proc. IEEE 1st International Conference on Broadband Networks (BroadNets 2004), pp. 334-336, (San Jose, CA, USA), 2004.

p21. M. Nord. “Replacing Shared-Per-Node Wavelength Converters by Fibre Delay Lines in an Asynchronous Optical Packet Switch”, Proc. ECOC 2004, vol. 3, pp. 758-759, We4.P.145, (Stockholm, Sweden), 2004

p22. M. Nord, and H. Øverby. ”Packet Loss Rate and Jitter Differentiating Quality-of-Service Schemes for Asynchronous Optical Packet Switches”, OSA Journal on Optical Networking, 3(12), pp. 866-881, (URL:www.osa-jon.org/abstract.cfm?URI=JON-3-12-866), November 2004.

p23. M. Nord, S. Bjørnstad, and M. L. Nielsen. “Distributed MAC Protocol for Optical Packet Switched Ring Network Supporting Variable Length Packets”, OSA Journal on Optical Networking, 4(4), pp. 213-225, 2005.

p24. M. L. Nielsen, M. Nord, and B. Dagens. ”2 x 20 Gbit/s WDM to 40 Gbit/s OTDM translation using MZI wavelength converter”, (to be) submitted to OSA Optics Express, 2005.

p25. H. Øverby, M. Nord, and N. Stol. “Evaluation of QoS Differentiation Mechanisms in Asynchronous Bufferless Optical Packet Switched Networks”, accepted for publication in IEEE Communications Magazine, 2004.

p26. M. Nord. ”Fairness Support in Flexible Asynchronous Optical Packet Switched Ring Networks”, submitted to Elsevier Journal of Optical Switching and Networking (OSN), 2005.

p27. S. Bjørnstad, M. Nord, Torodd Olsen, D. R. Hjelme, and N. Stol. ”Burst, packet and hybrid switching in the optical core network”, accepted for publication in Telektronikk, ISSN 0085-7130, (URL:www.telektronikk.com), 2006.

1. Introduction

1.1. Historical Background

From early times, optics has played an important role in communication technologies, due to the speed of light and the relatively long distances enabled, when line of sight requirements are respected. As early as 1184 BC, the Greeks used a system of torches to send a message of victory from Troy to the city of Argos, ~600 km away [1]. This was a far more convenient (and less fatal) way, than using runners to convey messages, as after the Battle of Marathon. The optical telegraph was put in place for military applications in France, in the wake of the French revolution [2]. Decades later this was made obsolete by the invention of the electrical telegraph, telephone and radio, which enabled higher bandwidth communication over transoceanic distances.

The invention of the laser in 1958 [3], marked the beginning of the telecommunication industry's move towards fibre optical communication. By the 1970s and early 1980s, transmission systems combined multi-mode silica based low-loss optical fibre with semiconductor Fabry-Perot lasers [4]. These were directly modulated at 32-140 Mbit/s in the 1.3 µm spectral region, and transmission distance was limited to ~10 km by intermodal dispersion [4]. By using single-mode fibre, the electrical regenerator distance increased to ~40 km and the bitrate increased to a few hundred Mbit/s. Using the low-loss region around 1.55 µm further increased the regenerator spacing, and use of narrow spectrum Distributed Feedback (DFB) lasers overcame chromatic dispersion, thus enabling bitrates above 1 Gbit/s. Around 1990, the availability of high power semiconductor pump lasers enabled realising the Erbium-Doped Fibre Amplifier (EDFA) [4]. The EDFA can simultaneously amplify several signals within its spectral operation range (1530-1565 nm). Whilst earlier systems had one regenerator per single-channel fibre, the EDFA enabled cost-efficient Wavelength Division Multiplexing (WDM) transmission systems, mainly by replacing this array of regenerators by a single EDFA. Combining external modulators (to reduce the frequency chirping and thus signal spectrum), with increasingly sophisticated transmission techniques, WDM technology has in recent years demonstrated transmission of Tbit/s over transoceanic distances [4].

1.2. Optical Layer Evolution

Tbit/s capacity corresponds to transmitting the content of hundreds of DVDs per second. Having networks with such capacities available will greatly change our everyday life, since it enables services that are much more bandwidth demanding than today's telephone and Internet applications. However, due to bottlenecks in the telecommunication network, most users still observe a bandwidth that does not match their Mbit/s broadband connection. Consequently, e.g. downloading files from the Internet can be a tedious task, and real-time streaming of video is limited to low bandwidths. This occurs since all data pass through a complex protocol hierarchy. The throughput experienced by the user depends on the capacity of each layer, and on their interaction. Therefore, Tbit/s transmission capacity in the optical layer is an insufficient condition for Tbit/s network throughputs. Chapters 1.2.1 - 1.2.4 give the rationale for ongoing- and future changes in the design of telecommunication networks, with an emphasis on the optical layer.

1.2.1. Current Telecom Industry Trends

Most investments in the currently installed WDM technology took place during the boom in the late 1990s, fuelled by exponential growth in traffic and a beneficial economical climate. The traffic drivers stemmed from data centric applications, such as browsing on the World Wide Web, business adaptation of Internet applications, deployment of broadband access, and use of peer-to-peer software for file sharing [5].

Data traffic is fundamentally different from the static bandwidth requirements of voice traffic. E.g. when browsing on the Internet, the traffic generated by a user is very bursty: There are short periods with high traffic, when opening a new page, whilst there may be idle periods of several minutes, when reading the content of the page. Since aggregation of bursty traffic streams does not necessarily create one smooth stream, a particular challenge in packet switched networks relates to the bursty traffic patterns that may occur [6]. Moreover, many data applications tolerate larger delay and delay variations than real-time voice traffic. Finally, transport protocols, such as Transmission Control Protocol (TCP), can make delay tolerant applications more robust towards loss of packets. These differences call for a rethinking of network design, which traditionally has been optimised for voice traffic.

The quantity of network traffic is an important factor in determining demand for equipment, eventually motivating new architectures when

proven to be more cost-effective. We here concentrate on the North-American Internet traffic, which is by far the largest in the world. Several claims of annual Internet traffic growth rate factors of 8 and 16 were made during the boom. More sober analysis reveals that annual growth factors of approximately 2 (between 1.7 and 2.5) were more correct for the 1997-2002 period [7]. These numbers are in the same range as those provided by the market research and consulting firm RHK, which also reports that the annual growth factor has declined from 2 in 2001 to 1.66 in 2003, with a prediction of 1.5 for 2004 [5]. On the one hand, this decline in relative growth rate can be expected since the impact of successive growth is muted by the large base of existing traffic [5]. On the other hand, traffic growth is disruptive and depends on a number of factors [7]. Hence, whilst an annual growth factor of 1.5 may very well be representative for the coming years, an annual growth factor of 2 is also likely for the remainder of this decade [7]. If this holds, the upper estimate of data amount switched by US Internet backbones will increase from the 1997 value of ~4 Petabytes per month [7], to ~32,000 Petabytes per month in 2010. On the other hand, if the growth rate is only 1.5 from 2004 and onwards, this value is ~4,300 Petabytes per month. Hence, by 2010 this traffic calls for an average throughput in the 10-100 Terabit/s range. In addition, the dimensioning should take day-time variations into account, calling for even higher maximum network throughput.

From a qualitative viewpoint, one can expect that an increased number of Internet users with improved access capacity will ensure a continuing strong growth, by adopting bandwidth demanding services such as video on demand, online gaming and videoconferencing. This is evidenced by a rapid increase in the number of worldwide Asymmetric Digital Subscriber Lines (ADSL) connections, which has increased from 35.9 millions at the end of 2002, to 100 millions per March 2005, with a worldwide increase of 58 % last year [8]. Whilst ADSL provides access bandwidth in the 1-6 Mbit/s range, newer standards, such as ADSL2, Very-high-data-rate DSL (VDSL) and VDSL2 are being introduced, which enable ~10-100 Mbit/s. Finally, several recent government initiatives and industry programs aim at bringing fibre to homes and businesses, mainly in Asia, but also with some lag in Europe and USA [9]. Such fibre based solutions enable access bandwidths around ~0.1-1 Gbit/s. However, the users will not be able to fully benefit from their high access capacities unless other bottlenecks in the network are removed. Hence, if the adoption of bandwidth demanding applications takes off, it can trigger the start of a positive cycle, involving an upgrade of the metro- and core segment. This opens up for new networking technologies, suitable for bandwidth intensive, data centric traffic.

1.2.2. Static Optical Layer

Consider a realistic example of today, termed a *first generation* optical network. It is constituted by multiple protocol stacks residing on top of each other, as described in [4]. An example of such a protocol stack is illustrated in Fig. 1.1 a).

What defines a first-generation optical network is its use of manually configured point-to-point WDM channels as the physical layer. Over this layer, main services have traditionally been voice and private lines, which interface with an Asynchronous Transfer Mode (ATM) and/or Synchronous Digital Hierarchy (SDH) layer. The ATM layer is a connection oriented layer, combining packet switching with virtual circuits, thereby enabling guaranteed Quality of Service (QoS). SDH is a Time Division Multiplexing (TDM) circuit switched layer, thereby enabling guaranteed bandwidth and latency. SDH is optimised for voice traffic, and is very suitable for multiplexing and accessing low TDM bit rate streams. In addition, SDH provides monitoring capability and is very resilient. However, SDH suffers from a number of drawbacks for dynamic data traffic [4]:

- SDH is static, since it requires manual intervention to set-up circuits.
- SDH equipment is designed to have lower tributary interface speeds than the line speeds. Hence, when data traffic enters at increasingly high speeds from router ports, the SDH equipment needs to operate at even higher speed, e.g. a 10 Gbit/s SDH switch for 2.5 Gbit/s IP router ports. Further increases in IP port line speeds, may thus not be supported by SDH switches.
- The coarse bandwidth granularity of SDH leads to over provisioning. As an example, transporting a 100 Mbit/s Ethernet signal calls for a 155 Mbit/s connection.
- The high resilience is achieved on the expense of using ring topologies which is a sub-optimum match with the meshed traffic. This drawback is accentuated by the low protection granularity, which prevents omitting protection for traffic that does not need it.

First generation IP clients include web browsing, file transfers and low-bandwidth streaming services. The IP network layer gives a survivable global connectivity, through connection-less packet routing and Best-Effort (BE) forwarding in IP routers. These packets are aggregated and encapsulated using the Point to Point Protocol (PPP), then framed by the High-Level Data Link Control (HDLC) for transmission of the synchronous SDH link. Alternatively, they can be sent on ATM defined virtual circuits. However, since ATM switches cells of fixed size (53 B),

whilst a large fraction of IP traffic is ~576 B or 1500 B, this solution results in a large overhead.

A main drawback of this first generation optical network is that it only uses optical technology to provide static bandwidth between neighbouring network nodes. All switching and processing is handled by electronics, be it SDH switches, ATM switches, or IP routers. These operations are becoming increasingly difficult to carry out as the aggregate capacity and/or bitrate of single channels increase. E.g., IP routers constitute a future bottleneck, since their capacity is ultimately limited by power- and space considerations [10]. Other hurdles for practical realisation of Tbit/s routers are robustness of high speed interconnections, power dissipation of integrated circuits and the need for multicard switching fabrics [11]. Some of these challenges can be alleviated by introducing optics inside the router. E.g. using high-speed optical interconnects may be better than a large number of Gbit/s electrical interconnects [11]. Moreover, the switch-fabric may also be optical, or a hybrid electro-optical solution [11]. However, since the remaining functionalities are electronic, several arrays of transponders are needed to perform O/E/O conversion. The most demanding conversions are those at the router port interfaces, since they have to work at the optical transmission bitrate. This is a severe drawback, as transponders are costly, and since their complexity, power and space consumption increase with the bitrate [4]. An important, practical issue is that the increasing power consumption increases the heat dissipation. In turn, this puts high requirements on the ventilation of the chassis, and on the cooling of the facilities to ensure correct operating temperature.

1.2.3. Circuit Switched Optical Layer

The recent availability of commercial Optical Add/Drop Multiplexers (OADM) and Optical Cross Connects (OXC) paves the way for implementation of a *second generation* optical networks. This network type employs Optical Circuit Switching (OCS) in an optical layer above the WDM layer, as depicted in Fig. 1.1 b). As opposed to electrical switches, optical switches do not process every bit. OADMs and OXCs thus enable a higher total switching capacity, occupy a smaller footprint, and have a lower cost per port than electrical switches [4]. Hence, switching and routing high-capacity connections is much more economical in the optical layer than in the electrical layer [4]. Using control plane solutions under development [12], such as the Automatic Switched Optical Network (ASON) or Generalised Multi-Protocol Label

Switching (GMPLS) to set-up lightpaths on demand, the optical layer creates a virtual topology for its client layer. This topology relaxation is illustrated in the example of Fig. 1.1 c). The IP routers use this virtual topology to bypass unnecessary electrical switching and processing. Hence, the second generation of optical networks alleviates the electrical switching capacity bottleneck.

Furthermore, new deployments are likely to benefit from a protocol stack that is more adapted to the increased importance of data-centric applications. It is a design goal to minimise overlapping of functionalities in the different layers, and tailor the service level to the application. The IP layer is becoming the most popular service network interface, and adding Multi-Protocol Label Switching (MPLS) capability to IP routers, enables a connection oriented service, through the concept of Label Switched Paths (LSPs). This increases both QoS and Traffic Engineering (TE) capabilities. The objective of TE is to "put the data traffic where the network bandwidth is available" in an efficient and effective way [13], whilst maintaining the specified QoS. Hence, TE contributes to an improved bandwidth utilisation. In IP/MPLS over OCS, one can reduce congestion by balancing the load, by using explicit routing to send MPLS encapsulated IP packets over non-shortest paths. MPLS also defines fast protection and restoration mechanisms [13]. The flexible label stack in MPLS enables high granularity. E.g. in case of a cut on a fibre used by several LSPs, a pre-defined back-up LSP can be used to exclusively protect flows that need protection, thus not wasting bandwidth to protect other flows.

These beneficial properties of MPLS may make ATM redundant as an IP server layer. A further change in the protocol stack is likely to be the disappearance of SDH as a ubiquitous link layer protocol in the core of the network, due to its inadequateness for data traffic in particular. However, IP/MPLS packets cannot be sent directly over lightpaths; a protocol is needed for management, monitoring and BER measurements. This can either be undertaken by direct SDH framing in the IP router ports (thus without separate SDH equipment), or by using a Digital Wrapper protocol around the optical channel [13]. This type of protocols is very flexible, and can include Forward Error Correction, the incorporation of a Data Communication Network (used for control plane communication), as well as encapsulating signals using Fiber Channel and Gigabit Ethernet protocols [4]. Increased flexibility makes it possible to tailor restoration properties to individual lightpaths. The latter contributes e.g. to offer low cost channels, when leasing capacity to Internet Service Providers (ISP) that only offers a BE service, by omitting protection of these channels.

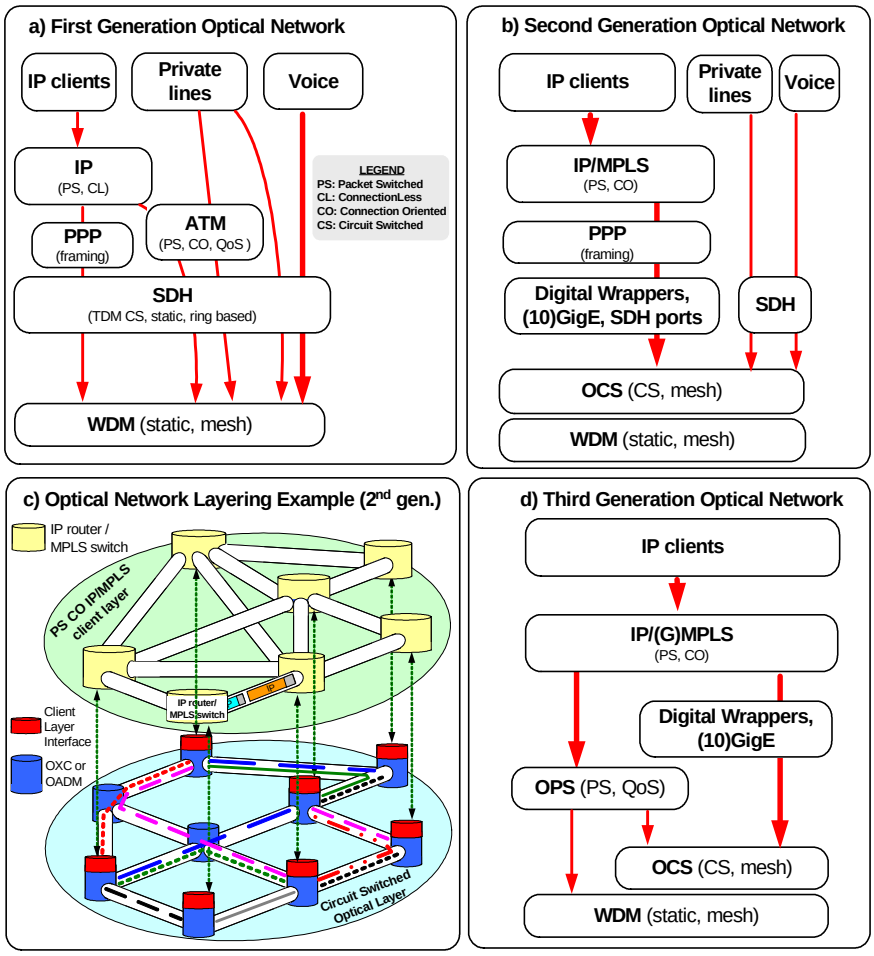

Fig. 1.1. Possible evolution scenario for core telecommunication protocol stacks: a) 1st generation, b) 2nd generation, c) illustration of how an OCS provides a virtual topology for an IP/MPLS layer, d) 3rd generation optical network.

1.2.4. Statistically Multiplexed Optical Layer

To maintain a reasonable cost per user it is important to efficiently share the infrastructure cost. The classical telephone network uses Time Division Multiplexed (TDM) based circuit-switching. Such *statical multiplexed* telephone conversations thus occupy a fixed bandwidth between two end-points, regardless of whether anything is being said or not. The advent of computer networks introduced *statistical multiplexing*, where a user only occupies network bandwidth when actually sending data. Hence, packet/burst switching networks require less total capacity, to obtain the same throughput, which is referred to as statistical multiplexing gain. The drawback is that these networks do not offer absolute transfer guarantees, since data may be lost in the network. Examples of early electrical packet networks include ARPANET, the predecessor of Internet, and Ethernet, for the local area network, implemented in 1969 and 1973, respectively [14].

Third-generation optical networks combine the high capacity of fibre-optical transmission and -switching with the efficiency of statistical multiplexing. Optical Packet Switching (OPS) and Optical Burst Switching (OBS) represent such switching paradigms, further discussed in Chapter 3.2 - 3.3. The OPS or OBS layer will operate on top of either a static WDM network, or a dynamic OCS layer, as depicted in Fig. 1.1 d).

An OCS network has a high minimum lightpath capacity and a coarse capacity granularity, typically 2.5, 10 and 40 Gbit/s. Many clients may not be able to fill up a reasonable fraction of this capacity. In this case, extensive grooming in the electrical domain, possibly involving O/E/O conversions in intermediate nodes, is needed to avoid waste of bandwidth. In addition, the OCS layer may not keep up with the dynamics of the traffic. The idle periods during connection set-up and tear-down, initiated as response to changes in desired connectivity or connection capacity, represent in fact wasted bandwidth. These periods will as a minimum equal the millisecond range return propagation delay, and may be orders of magnitude larger, depending on the control plane and settling time of the switches. In contrast, the sub-wavelength granularity of OPS and OBS avoids these idle periods. This contributes to increased bandwidth utilisation, provided that the packet/burst overhead is reasonably low, as discussed in Chapter 3.2.5.

Quantifying the gains of statistical multiplexing depends heavily on traffic matrix, dynamics, network topology, and assumptions made for the competing OCS and OPS/OBS networks. In the EU project IST

STOLAS (Switching Technologies for Optically Labeled Signals), a mesh network case study found that an OBS network only needed 64 % of the wavelength resources required by an OCS network (with wavelength conversion capability) [15]. The authors conclude that the OBS multiplexing gain achievable in realistic networks may be higher, since e.g. restoration was not taken into account and since the network size was rather small. A later study concludes that gains above 2 are achievable when the traffic sources have high peak bitrates (several tens of Mbit/s) and low mean-to-peak traffic ratio (<0.05) [16]. Moreover, decreasing the degree of meshness increases the gains, since there is little statistical multiplexing gain when many nodes are directly connected.

Benefits from TE, through e.g. load balancing, comes in addition to the statistical multiplexing gain. The benefits of MPLS for TE were discussed in Chapter 1.2.3. Generalised MPLS (GMPLS) is being developed to generalise the MPLS concept, and labels can now be e.g. TDM time slots, wavelengths, wavelength bands and optical packets. Hence, one can expect that OPS will bring similar TE benefits, as for MPLS. As discussed above, the coarse optical channel granularity makes it hard to fill up lightpaths with traffic of the same type. This may result in data belonging to different traffic types sharing a lightpath. If only a small fraction of this traffic needs protection, then considerable bandwidth is wasted when protecting or restoring this lightpath. In contrast, the packet granularity of OPS will reduce the bandwidth needed to restore traffic, by only protecting packets belonging to a Class of Service (CoS) that has this feature specified. How to extend the GMPLS framework to OPS, is e.g. being studied in the “Optical Label Switching” concept [17].

The sub-wavelength granularity of OPS and OBS requires reconfiguring the optical switch between each packet/burst. These reconfigurations will be controlled electrically, based on electronic scheduling. Note that this does not alter the transparency of the optical switch towards the payload.

To sum up, the attractiveness of OPS and OBS increases with:

- Coarser lightpath capacity granularity.
- Increasing traffic dynamics, both for:
  - Increasing magnitude of capacity variations between source-destination pairs.
  - Decreasing time-scale of variations relative to OCS set-up delay.
- Increased need for restoration.
- Increased network size and lower degree of meshness.
- Increased role of TE with sub-lightpath granularity.

One can argue that WDM made optical bandwidth a commodity, and that future capacity demands can thus be met by "throwing bandwidth at the problem". In this case there is little need for bandwidth efficient schemes, such as OPS and OBS. It is true that many carriers over-estimated their bandwidth needs during the boom in the late 1990s, which resulted in the current surplus of potential capacity in unlit fibres. However, the cost of bandwidth includes much more than just the fibre. In addition to system equipment, such as lasers, amplifiers and receivers, the cost of the switching equipment, as well as the operational cost of the network (including control and management), must be taken into account.

Since "economics will always demand that the network resources be used efficiently" [18], trying to handle future capacity increases by simply equipping more lightly-loaded fibres is not attractive in the long run. A recent study examining the potential for OPS supports this view, and deems OPS to be a better match than OCS for IP traffic [19]. In addition, the following main factors to promote OPS deployment were identified:

- Continued growth of the Internet.
- Telecommunication market recovery.
- Establishment of rational network migration scenarios.
- Overcoming technological barriers.

The first requirement seems to be fulfilled, as discussed in Chapter 1.2.1.

The second topic is out-of-the scope of this thesis, but since markets tend to find a balance between offer and demand, one can expect a return to profits for both vendors and carriers, in turn enabling new investments.

Proposals for technological solutions for migration scenarios are made in Chapter 3 and Chapter 4, but this issue needs to be further addressed by the research community, including economical aspects in the scenarios.

To overcome technological barriers, it is important to bear in mind that optical technology beyond transmission is very immature. This is in stark contrast to electronic technology, which has been heavily invested into during more than five decades. By exponentially increasing the number of low-cost components per integrated circuit, as predicted by G. E. Moore [20] (later known as "Moore's law"), electronic processing has demonstrated tremendous technology advances, whilst cutting manufacturing cost. Current optical switch designs mainly combine discrete components. To achieve similar cost decreases as those of electronics, the Photonic Integrated Circuit (PIC) is the "holy grail" of optical networking. The first encouraging results are reported from the labs, and also from the industry. This is illustrated by e.g. Intel's and Infinera's recent announcements of silicon-based laser and modulator [21], and a suite of PICs that combine dozens of active and passive

devices for managing light in an optical transport system [22], respectively. However, more research and development is needed find efficient manufacturing methods for technology used in optical packet and burst switching.

1.3. Optical Packet Switching Research

The rationale for OPS was established in the 1990s, and triggered a huge worldwide research effort on the topic. In Europe, some of the more prominent examples include the ACTS project Keys to Optical Packet Switching (KEOPS) [23, 24], which demonstrated main OPS building blocks. The IST project Data And Voice Integration over DWDM (DAVID) [25, 26] refined the networking concept, and has demonstrated an OPS Metropolitan Area Network (MAN) with an O/E/O interface to the Wide Area Network (WAN) [27]. IST STOLAS focuses on both OBS performance and realisation of OBS building blocks [15]. Finally, the U.K. based WASPNET project [28], and its follow-up, OPSnet, has demonstrated OPS switches using wavelength routers in asynchronous operation at bitrates of 40 Gbit/s [29, 30].

Japan has been involved in the development of optical networking solutions for a number of years, as exemplified by an OPS prototype exhibited at the OFC in 2003 [31, 32] by the National Institute of Information and Communications Technology (NICT).

Numerous groups in the U.S. have also been very active, both on OPS and OBS research [17, 33, 34]. Future high-capacity demonstrations are to be expected, as e.g. the U.S. Defence Advanced Research Projects Agency (DARPA) started in 2004 to fund the $15.8 million budget Label Switched Optical Router (LASOR) project, with a goal of demonstrating an optical packet router with a throughput of 100 Tbit/s [35].

One can group OPS research into the following categories:

- Large-scale prototypes demonstrate OPS feasibility, in order to prove the viability of this switching paradigm.
- Techno-economic studies evaluate the attractiveness of OPS networks, compared to traditional solutions.
- Performance studies help evaluating performance of statistical multiplexed optical networks.
- New OPS features and schemes, together with optical processing methods, enable offering either more advanced functionality, or less complex OPS node and network designs.

The first category demands an enormous amount of resources, and is mostly suitable for large projects that include partners from industry.

Accurate techno-economic studies are hindered by lack of realistic cost assumptions, since most OPS components are not commercially available. The cost of such components will much depend on the level of mass-production, which is difficult to assess at present.

For performance studies, analysis is an attractive method in many cases. However, some cases may not be feasible for analysis without over-simplifying network behaviour, which limits the applicability of the results. Event-driven simulations then constitute a better option. Still, simulations of network performance demand very high computational power, so many studies limit the evaluation to a single node. However, some aspects must be studied in a network context, such as fairness between users in the network.

Proposals for innovative designs are important to make OPS networks a more attractive optical layer candidate. E.g., whilst the first OPS designs resembled optical implementations of electrical switches, focus has increased on adapting the designs to the specifics of optical technology. E.g. since optical random access memory is unavailable, OPS networks, unlike IP networks, cannot be store-and-forward networks. Instead, the wavelength domain, which is specific to optical networks, enables an attractive contention resolution alternative.

Whilst the OPS research community pursues all these axes, this project focuses on the two latter categories, as detailed in Chapter 1.4.

1.4. Ph.D. Project Synopsis

1.4.1. Background and Scope

This project follows up on the substantial work on optical signal processing and optical packet switching that has taken place at the Research Center COM during the last decade, as evidenced by a number of Ph.D. projects [36-41]. A close collaboration with Alcatel Research & Innovation has enabled using state-of-the-art optical SOA based signal processing devices for these studies.

The project has involved collaboration on a number of topics with S. Bjørnstad and H. Øverby during their Ph.D. studies at the Norwegian University of Technology and Science. Hence, this thesis is complementary to theirs [42-43]. Whilst the former focuses on scalability of OPS switches with electronic buffering, hybrid circuit- and packet switching, and use of polarisation multiplexing for networking purposes,

the latter focuses on analysis of networking performance, in particular QoS differentiation.

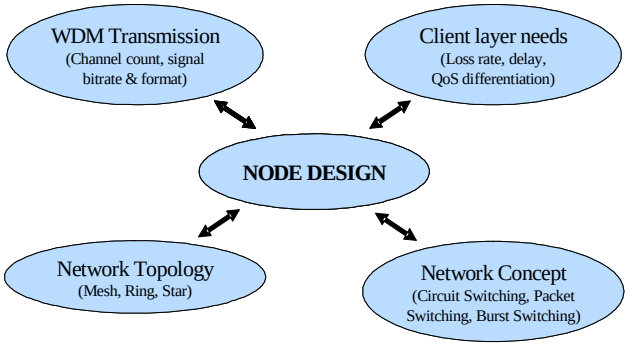

Fig. 1.2. Node design results from the interplay of a number of network parameters (non-exhaustive).

This project has a taken a wide perspective on node design, spanning from single component operation to network level performance aspects. A good node design is the result of the interplay of a number of parameters, as depicted in Fig. 1.2, which calls for a holistic network view. The work has been put in a context reflecting current trends in OPS network research:

- Asynchronous operation with Variable Length Packets (VLP), to avoid data alignment in synchronous networks, and to be compatible with Internet transfer mode and packet length distribution.
- QoS differentiation, to efficiently support different applications, network clients and/or Classes of Service (CoS).
- Hybrid networks, to combine the guaranteed QoS of OCS and the efficiency of OPS, as well as enabling migration scenarios.
- OPS in the MAN, since metro networks are expected to carry more dynamic and bursty traffic than the WAN, and since its limited scope makes it a likely first-implementation of OPS [26].
- Fairness between end-users. Increasing the performance evaluation detail level from typical average performance parameters, such as overall network throughput, increases performance insight and the value for the network designer.

1.4.2. Thesis Organisation

This thesis consists of 7 main *Chapters*, devoted to the following topics:

- Chapter 1 motivates the Ph.D. project, puts it in a proper context and describes the organisation of this thesis.
- Chapter 2 reviews the status of optical technology in order to identify suitable components and justify network assumptions.
- Chapter 3 discusses main network- and node architecture design options. This includes OPS and OBS concepts, packet handling schemes, main node building blocks with complexity analysis, and node designs for efficient contention resolution that simultaneously overcome scalability constraints.
- Chapter 4 reports a demonstration of a hybrid network design, combining OPS/OCS data segregation with OPS header-payload separation and forwarding.
- Chapter 5 discusses means of service differentiation. It compares bufferless QoS methods and evaluates a number of proposed QoS differentiation algorithms for different Shared Per Node (SPN) contention resolution pool designs.
- Chapter 6 focuses on Optical Packet Switched Ring Networks (OPSRNs) for the MAN. A novel node design is demonstrated, and a MAC protocol that supports VLP and fairness for a high network throughput is proposed and evaluated. The performance is compared to that of a Static Wavelength Routed Optical Network (SWRON).
- Chapter 7 draws the conclusions and identifies future research topics.

Different forms are chosen for these Chapters, as detailed in their introduction, and as described in the following list:

- Chapter 1 and 2 are quite general. They are thus written from scratch, referring mainly to the work of others, or to PhD publications not discussed in detail in this thesis.
- Chapter 3 discusses node design issues in the form of overview/discussion. This chapter is then written as a summary of the content from a number of Ph.D. papers, characterised as *included* papers in this thesis.
- From Chapter 4 and onwards, the content is given in the form of *incorporated* papers (incorporated in their entirety into this thesis). This is chosen since these papers address a distinguished issue each, with little overlap.

The *included* and *incorporated* papers are selected among those in the Ph.D. publication list on pp. xiv-xv, in order to give a representative view of the work in this project, whilst minimising overlap, and avoid an excessive thesis length.

Regarding the incorporated papers, they occupy a sub-chapter each (e.g. "Chapter 4.2"), and each paper is divided into several *Sections*, whose section number starts with a letter (e.g. "Section B.2"). The format of the articles has been adapted to match the format of this thesis. These incorporated papers are detailed in Table 1.1.

The references in this thesis refer to either the Ph.D. publication list (denoted by "p", e.g. [p2]) on pp. xiv-xv, or to the reference list on pp. 187-193.

The numbering of figures, tables and equations reflects the structure of the main chapters. E.g., the third figure in Chapter 4, is referred to as "Fig. 4.3".

For a complete view of this Ph.D. project, it is recommended to proceed chapter by chapter. However, if the reader is interested mainly in one topic, one can focus on the corresponding chapter. Likewise, the incorporated papers are self-explanatory, and can be read without reading other parts of the thesis.

Fig. 1.3 details the topic and method of all Ph.D. publications, and it also indicates network segment, author, whether the publication has been incorporated, included or not included (only briefly referred to) in this thesis, the status of the publication (published, or in peer review). The figure highlights that discussion and logical performance evaluation of node design with advanced features such as QoS differentiation have been the focus of this project, but also that a number of physical demonstrations have been carried out.

Table. 1.1. List of papers incorporated (in entirety) in thesis.

Topic	Chapter	Ref	Contribution type	Title
Node Design	3.6	[p21]	*ECOC 2004* conference paper	Replacing Shared-Per-Node Wavelength Converters by Fibre Delay Lines in an Asynchronous Optical Packet Switch.
Hybrid Networks	4.2	[p16]	*OFC 2004* conference paper	Demonstration of optical packet switching scheme for header-payload separation and class-based forwarding.
QoS differentiation	5.2	[p25]	*IEEE Comm. Mag. 2005* journal article*	Evaluation of QoS Differentiation Mechanisms in Asynchronous Bufferless Optical Packet Switched Networks.
	5.3	[p20]	*IEEE BroadNets 2004* conference paper	Performance analysis of a low-complexity and efficient QoS differentiation algorithm for bufferless optical packet switches with shared wavelength converters in asynchronous operation.
	5.4	[p22]	*OSA JON 2004* journal article	Packet Loss Rate and Jitter Differentiating Quality-of-Service Schemes for Asynchronous Optical Packet Switches.
Metro Networks	6.2	[p14]	*Photonics in Switching 2003* conference paper	Novel strictly non-blocking Node Designs for asynchronous OPS MAN.
	6.3	[p23]	*OSA JON 2005* journal article	Distributed MAC Protocol for Optical Packet Switched Ring Network Supporting Variable Length Packets.
	6.4	[p26]	*Elsevier OSN 2005* journal submission	Fairness Support in Flexible Asynchronous Optical Packet Switched Ring Networks

*Second author

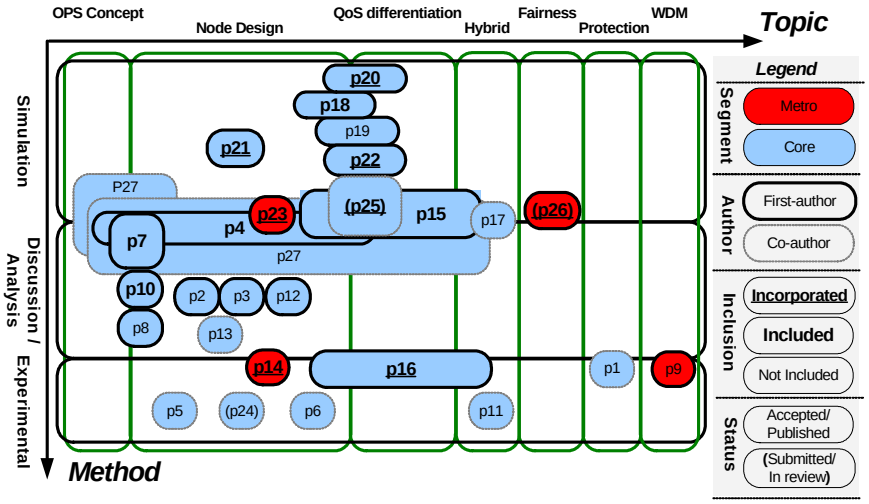

Fig. 1.3. Topic and method of Ph.D. publications.

1.4.3. Methods

The node- and network performance evaluations have been carried out by discrete, event-driven simulations. This was chosen, since most of the schemes are too complex to accurately analyse. An exception is the analytical studies presented in Chapter 5.2, carried out by H. Øverby. The choice of OPNET as simulation tool was motivated by its suitability for packet simulations (since a vast library of Kernel Procedures that facilitates the handling of objects/packets is available), its ease of analysis (using the built-in graphical interfaces), and its suitability for (future) integration with models of commercial devices.

All simulations omit the transient period when capturing data. The indicated 95 % confidence intervals are calculated using the method detailed in [44]. The results of the simulations have been compared with analytical results, where applicable, and with results from other groups, where comparable input parameters allows it. E.g. (as commented on) the PLR resulting from the SWRON in Chapter 6.4 (when ingress buffer size is set to 0) equals that found by analytical results from the Erlang-B formula. In Chapter 3.4.2, the confidence interval found was within the one found for a comparable case [45] (using parameter setting: $load$=0.7, F=8, W=64, WCR>0.7, no FDLs). As commented on in Chapter 3.4.2, the performance of the contention resolution pools has also the same qualitative behaviour as similar studies. Yet another example is the confirmation of deteriorating performance with bursty traffic, in Chapter 6.3 (as commented on). For QoS differentiation schemes, no comparative studies were found, but the trends seems reasonable. A number of small tests have been conducted to verify the good behaviour of the simulation models. These verifications and observations contribute to a high confidence level in the correctness of the results.

The demonstration of optical functionality has been carried out in the lab-facilities at Research Center COM, except [p1, p11], which took place at the Telenor R&D laboratory. The demonstrations are mainly proof-of-principle demonstrations where signal quality is evaluated by Bit Error Rate (BER) measurements.

2. Optical Technology

2.1. Introduction

Node design calls for a realistic view on the current and future status of optical technology. Main aspects of optical technology are thus discussed throughout this chapter, grouped into the following categories:

- Fibre Optical Transmission. Discusses trends in WDM transmission systems, and its effects on client equipment.
- Optical Switching. Sums up performance of optical switching technology for OPS.
- Termination. Sums up status of tuneable laser and Optical-to-Electrical (O/E) receivers suitable for asynchronous operation.
- Storage. Discusses how information can be stored in the optical domain, applicable for buffers and synchronisers.
- Signal Conditioning. Outlines how flexible networks require the ability to convert optical signal format-, wavelength-, and bitrate.
- Logical processing. Sums up main demonstrated optical logic functionalities.

2.2. Fibre Optical Transmission

There are three main options to increase the capacity of a link:

- Increase the channel bitrate.
- Decrease the channel spacing.
- Increase the amplification spectrum, beyond the conventional C-band
- In addition, one can combine these methods with polarisation multiplexing, enabling two channels to share the same wavelength, and with multi-level coding, increasing the information rate, for a given bitrate.

However, these methods tend to increase the impact of linear, non-linear and cross-talk effects. E.g., increasing the bitrate requires a larger detection bandwidth, which requires an increased number of photons per

bit to maintain the signal quality. In turn, this increases the nonlinear phase shift from Self Phase Modulation (SPM). Record-breaking experiments overcome these obstacles by employing a number of sophisticated techniques, including advanced modulation formats, distributed amplification, advanced chromatic- and polarisation mode dispersion compensation schemes, and Forward Error Correcting (FEC) codes [46]. E.g., 6 Tbit/s has been transmitted over 6000 km [47].

Regarding the bitrate, 40 Gbit/s bitrate systems are ready to be installed in commercial systems when there are demands for it. Several vendors offer 40 Gbit/s transmission systems, and carriers have completed successful field trials, as e.g. the one reported by MCI in cooperation with Ciena and Mintera in 2004 [48]. Even IP router vendors are preparing for the introduction of this technology, as illustrated by the Cisco CRS-1 router prototype with 40 Gbit/s interfaces [49].

However, most deployed systems still use 2.5 bit/s or 10 Gbit/s bitrates. A major difference between these two alternatives is that the former calls for fewer transmitters and receivers, but of higher complexity, and the latter calls for more transmitters and receivers, but of lower complexity. A number of factors have influence on whether upgrading to 40 Gbit/s systems represent a cost-advantage. In addition to the cost ratio of 40- and 10 Gbit/s equipment needed for the WDM transmission, the optimum switching granularity in the network is important and depends on the input traffic characteristics. Moreover, having more channels available increases the statistical multiplexing gains when employing the wavelength domain for contention resolution, as illustrated in Chapter 3.4. This provides a rationale for Dense WDM (DWDM), as discussed e.g. in [p9]. With the commercialisation of DWDM systems with 25 GHz channel spacing [4], 175 channels can be supported within the C-band. Using both the C-band (1530 nm – 1565 nm) and the L- band (1530 nm - 1610 nm), 320 channels have been demonstrated [50].

2.3. Optical Switching

A very extensive survey of available switching technologies has been conducted [p2], identifying main switching technology candidates for Optical Circuit Switching (OCS), Optical Burst Switching (OBS) and Optical Packet Switching (OPS). Main types of architectures are listed below. The most promising results for OPS from each of three main switch architectures were identified in [p3]:

- Space switches: Based on interconnection of 1x2 and 2x2 switches. Scalability is poor, due to limited integration potential, noise- and loss cascadability issues.
- Array Waveguide Grating Routers (AWGRs): This architecture combines AWGs with tuneable lasers, and has demonstrated 1.2 Tbit/s throughput.
- Broadcast And Select (B&S): Has demonstrated OPS with 640 Gbit/s and potential for 2.56 Tbit/s throughput, using Semiconductor Optical Amplifier (SOA) gate technology.

Studies on cost-effectiveness, footprint and power consumption are needed to make the best choice of switch fabric, given OPS network requirements such as throughput, transparency, switching time and scalability. This is currently not feasible due to immature technologies, and is out-of-the scope of this project. However, main switch parameters are compared in Chapter 3.3.3 in order to highlight important realisation issues of each architecture.

2.4. Optical Interfaces

The optical path followed by a packet starts by an E/O converter and ends with an O/E converter. The former consists of a laser and an electro-optic modulator. Both fixed and tuneable lasers can be realised, with tuning times in the ns-range [51].

For OPS and OBS networks, since the phase of the bits varies from packet to packet, or from burst to burst, the O/E converter consists of a Burst Mode Receiver (BMR) which is more complex than a synchronous receiver. To minimise packet overhead, as discussed in Chapter 3.2, the BMR must thus be able to perform Clock Recovery (CR) within a few ns. Furthermore, it must be robust towards the packet-by-packet power variations that result from packets following different signal paths. It should incorporate (electrical) 3R to make it more robust toward signal degradations. Such a device has been reported at 10 Gbit/s, together with 40 Gbit/s burst-mode BER- and packet loss measurement equipment [52].

2.5. Optical Storage

As discussed in Chapter 3, OPS and OBS would benefit from being able to store light for the following applications:

- Delay payload data, while processing the control information.
- Synchronisation / alignment of data.
- Buffering for contention resolution.

Electronic packet switched networks rely extensively on electronic First-In, First-Out (FIFO) buffers, with random access in the time domain. Furthermore, This is realised with relative ease by using transistors to control the electrons. In contrast, photons are not as easily manipulated; "stopping" light presents a great challenge. Note that recent breakthrough experiments have managed to significantly slow down light [53]. However, this technology is very immature, and demonstrations require an advanced lab set-up. Hence, optical FIFO buffers with random access in the time domain seem too impractical for use in first implementations of OPS.

A more promising path is to exploit the low noise and low loss of optical transmission. When light needs to be stored, it can simply be inserted into coils of fibre, termed Fibre Delay Lines (FDLs). Light can then be retrieved at a predetermined moment, governed by the FDL length.

In general, increasing the time resolution optimises the bandwidth utilisation. This can be achieved in a number of ways, e.g. by the following methods [34]:

- Programmable FDL: Let each delay consist of a set of FDLs with different lengths, and control which to use by using a passive splitter and then controlling optical gates at the end of each FDL. This solution was used in IST KEOPS for buffering in a switch architecture [23]. Alternatively, one can control the FDL delay by setting the wavelength of the input signal combined with demultiplexers associated with different FDL lengths [23, 28].
- Feed-forward time slot interchanger: Interconnect a number of 2x2 switches by two FDLs: one with negligible length and one with a stage-dependent length. The resolution increases with the number of stages. This solution was e.g. used for the synchronisation in IST KEOPS [24].
- Active switched recirculating FDL: Allow multiple circulations and use an optical 2x2 switch to determine at which circulation to release the packet from the FDL. This can be generalised to control FDL access by extensions to the switch matrix, as studied in Chapter 3.6.

A common drawback with optical buffering is that FDLs are bulky- a delay of e.g. 1 µs corresponds to roughly 200 m of fibre. It is therefore a design goal to limit use of FDLs to a few tens per node [54]. This design rule is respected in the studies of use of FDLs for contention resolution in Chapter 3.6 and Chapter 5.4.

2.6. Signal Conditioning

Increased flexibility in handling optical signals calls for ability to convert its wavelength, format and bitrate. Such flexibility is useful, to e.g.:

- Support multi-vendor or multi-technology optical networks, which operate with different signal wavelengths, format and bitrates.
- Resolve contention by wavelength conversion, cf. Chapter 3.4.

This project has contributed to the development of these technologies, through demonstrations of wavelength conversion at 40 Gbit/s [p5], and format-preserving bitrate conversion from 20- to 40 Gbit/s [p24]. Both demonstrations were conducted using an All Active-Mach Zehnder Interferometer (AA-MZI). A discussion of methods for wavelength conversion, regeneration and all-optical logic concluded that such SOA-based devices are promising candidates for optical signal processing [p13]. Their main advantages are compactness, potential for integration and large-scale manufacturing, and the inherent amplification in SOAs, which reduces the need for additional amplification.

Note that using optical modulation in single SOAs constitutes an alternative to interferometric WC designs [55]. To increase the detail level of WC designs, one can categorise them with respect to input- and output tunability. This is particularly important for the tunability of the probe, i.e. the laser source. In addition, in co-propagating configurations, which in general have higher bitrate potential than the counter-propagating configurations, the input tunability has an impact on the filtering used to separate new and original signal at the output. We thus denote WCs as follows:

- Tuneable Wavelength Converter (TWC): Needs a tuneable probe and tuneable Band Pass Filter (BFP) in co-propagating configurations.
- Fixed Input wavelength Wavelength Converter (FIWC): Needs a tuneable laser probe, but can replace the tuneable BFP with a fixed reflection filter and isolator/circulator, as discussed in [p16].
- Fixed Output wavelength Wavelength Converter (FOWC): Can use a fixed laser probe, and a fixed BPF.
- Fixed Input- and Output wavelength Wavelength Converter (FIOWC): Combines the advantages of FIWCs and FOWCs.

Hence, in counter-propagating configurations, the FOWCs and the FIOWC are less complex than TWCs, since they have a fixed laser source. In co-propagating configurations, FIWCs are also less complex than TWCs due to simplified filtering, but this type still requires laser tunability, as opposed to FOWCs and FIOWCs.

2.7. Optical Regeneration

Unlike electronic digital networks that reshape each signal bit during processing (switching, storage etc), optical networks are analogue in nature. The signal should transparently go from source to destination in the network. However, the pulse shape is altered by a number of causes, e.g. by Amplified Stimulated Emission (ASE) noise during amplification, by linear and nonlinear effects during transmission, and by patterning effects in SOAs. In order to maintain a sufficiently good signal quality for correct interpretation at the receiver, the signal may have to be regenerated. Full regeneration, termed "3R Regeneration", consists of:

- Signal Reamplification.
- Signal Reshaping.
- Signal Retiming.

Whether 1R, 2R or 3R regeneration is required, depends on the network. 1R is sufficient when ASE noise does not limit the system. However, when the signal is split- and reamplified several times, or subject to high fibre losses before reamplification, 2R regeneration is needed to overcome ASE induced noise limit. 3R regeneration is required to retime the signal when signal quality is impaired by jitter, which may be introduced by noise from active components, environmental fluctuations, Cross Phase Modulation (XPM) or Polarisation Mode Dispersion (PMD). 3R regeneration can be achieved using a nonlinear gate with a properly timed probe signal. As reviewed in [41], a number of techniques can be employed to perform this task, among them SOA-based Interferometers WCs (IWCs). Note that 3R regenerators include Clock Recovery (CR) for the retiming functionality.

2.8. Optical Logic Processing

In electronics, Random Access Memory (RAM) is available for logical processing. Although the lack of optical RAM imposes a limit to the complexity of functions that can be implemented using optical devices, there are still a number of feasible Boolean logic functions. As reviewed in [p13], AND, OR, XOR, and NOT have been realised using MZIs. Of these, the XOR is arguably the most interesting, since it can be used to implement simple, but useful functions, such as pattern recognition and parity checking. The former enables address-comparison, applicable to header look-up, and to simple label-swapping schemes. The latter can be used to verify the integrity of the data, without conversion to the electronic domain, as demonstrated in [p6].

3. Node Design

3.1. Introduction

This chapter discusses network- and node design issues, both for the control- and data plane. Chapter 3.2 - 3.5 sums up a number of included papers, whilst Chapter 3.6 incorporates a paper, as detailed in the following list:

- Chapter 3.2 discusses initial network design issues, the OPS and OBS concepts, packet/burst handling schemes and packet/burst formats. This is based on an invited paper at ICTON 2003 [p7], an ONDM 2003 paper [p4], and Section 4.1 – 4.2 (which was first-authored as part of this project) of the COST 266 Final Report [p10].
- Chapter 3.3 focuses on main node design options, both for the data- and the control plane, based on an invited paper at ICTON 2003 [p7].
- Chapter 3.4 details the rationale for- and the performance of a TWC based Shared Per Node (SPN) contention resolution pool design.
- Chapter 3.5 describes the proposal for a parallel switch design to overcome scalability constraints, and to support hybrid networks and QoS differentiation, based on the ONDM 2004 and Networks 2004 papers [p15, p18].
- Chapter 3.6 compares use of FDLs and TWCs in a SPN pool, by incorporating an ECOC 2004 paper [p21], which proposes the '*SoftRSV*' FDL buffer algorithm to reduce the need for TWCs in the SPN pool.

3.2. Network Design Issues

This chapter is based on findings in [p4, p7, and p10].

3.2.1. Network Context

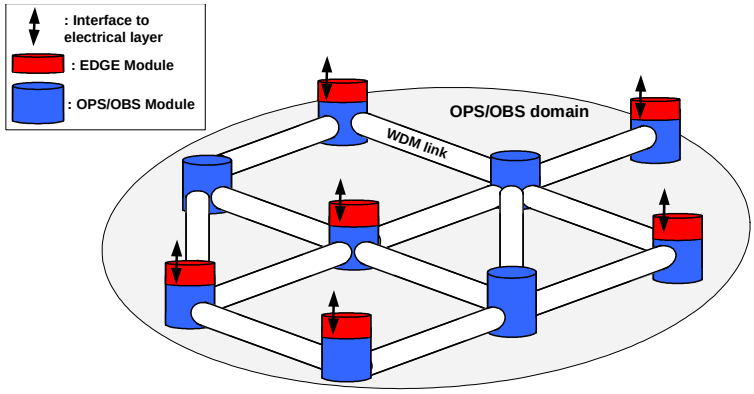

Fig. 3.1. Network consisting of nodes with edge and core routers functionalities

We consider OPS and OBS for application in a mesh-based WAN or "core network", context. We consider OPS for the MAN segment in Chapter 6. Mesh networks have lower hop count than ring networks, which give reasonable switch matrix sizes and propagation distance. Furthermore, it enables flexible load balancing and link-protection, whilst avoiding single-points of failure, as opposed to star networks. WDM systems with high channel counts are considered; the interplay between channel count and network performance is discussed in Chapter 3.4.

3.2.2. Network Transparency

Commercial transponders performing O/E/O conversions are optimised for a specific signal bitrate and transmission format. Networks based on optical switching may avoid transponders, which opens up for network *transparency*, here meaning design of a network that readily handles any signal format and bitrate. This is often cited as an attractive property of optical switching. Nevertheless, realisation of such *fully transparent* networks requires a high number of adaptive components, as discussed in [p4]. There is hence a trade-off between the flexibility and the

complexity/cost in network design. *Semi-transparent* networks, here meaning optical networks optimised for a certain signal format and bitrate, may therefore be attractive. Furthermore, such a network, due to the fixed signal format, may allow O/E/O conversion, e.g. to perform 3R regeneration, wavelength conversion and buffering; whenever a function is less costly to perform in the electronic domain. The work in this thesis assumes a common signal format and bitrate in a given network.

3.2.3. OPS and OBS Concepts

Both OPS and OBS have a clear separation of data- and control plane, since the payload of the packets/bursts stays in the optical domain during switching, whilst control information is O/E converted for electronic processing.

Note that OBS and OPS concepts are not strictly defined. Main differences were identified in [p4], as discussed in the following:

In OPS networks, the control information propagates *in-band*, i.e. on the same wavelength channel and simultaneously as the payload. In contrast, OBS uses out-of-band encoding of control information. A separate wavelength can be devoted to transmission of burst control packets (BCP) on each link. The BCPs are transmitted with a time-offset, with respect to their associated bursts. The minimum offset equals the expected BCP processing delay in the network, thereby avoiding the need for delaying bursts by FDLs during processing. The information contained in the BCPs may vary. In the *Reserve a Fixed Duration* (RFD) reservation scheme, the BCPs inform the scheduler of burst start- and end times [p4]. Combining the RFD scheme with the delayed reservation principle (as in Just Enough Time (JET) [57]), enables advanced scheduling with void minimisation, thus increased utilisation.

OPS studies typically assume a low aggregation of IP packets, and mean payload size is in the kB range. This corresponds to µs durations, at 10 Gbit/s payload bitrate. OBS assumes roughly 10-1000 times larger payloads, which calls for extensive aggregation of client packets.

The packet handling schemes may differ, as detailed in Chapter 3.2.4. Most OBS burst handling schemes are asynchronous with variable length bursts, whilst OPS are in general either slotted or asynchronous, with Variable Length Packets (VLP).

A main motivation for OBS is to use optical technology of relatively low complexity, achieved e.g. by avoiding burst alignment and having relaxed switching time requirement. However, recent studies have introduced

FDL buffers and complex reservation schemes, to improve bandwidth efficiency, which increases optical- and electronic complexity.

The higher assembly of OBS requires more buffering resources in the ingress nodes. A potential benefit is that the burst assembly may shape the ingress traffic. However, a recent study concludes that in general, burst assembly algorithms do not remove long-term dependency of traffic [58]. Hence, if the input traffic is bursty, the core nodes must be designed to handle this.

- The packet/burst handling schemes and format are detailed in Chapter 3.2.4 and 3.2.5, respectively. Further OPS/OBS control– and data plane differences are discussed in a node design perspective in Chapter 3.3.

3.2.4. Packet/Burst Handling Schemes

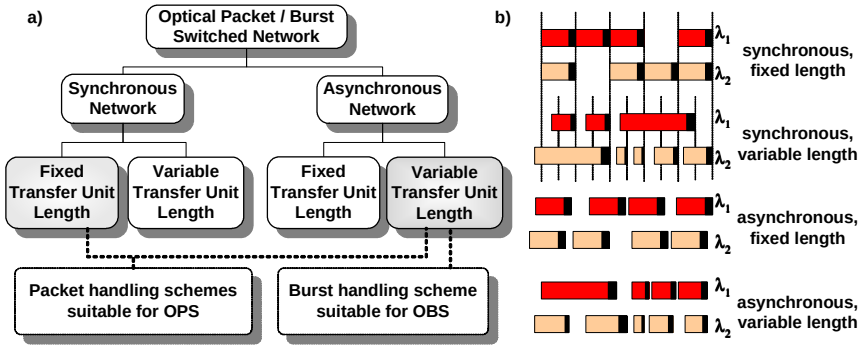

Fig. 3.2. Potential packet/burst handling schemes in OPS and OBS.

The packet handling scheme is defined by whether the data units arrive in a synchronous or asynchronous manner at the nodes, and whether they are of fixed duration or not. The basic principles for switching architectures and functionalities are independent of the packet/burst handling scheme [p4]. In Fig. 3.2 we report the four potential packet/burst handling schemes, classified according to synchronisation and size of the data units.

As indicated, asynchronous and variable length data units are considered more suitable for OBS. This is motivated by decreased complexity at the optical layer, and OBS will also benefit from a large degree of freedom in the burst assembly mechanism. This is the case most widely studied in the literature and to which we refer in the following. On the other hand, the best scheme for OPS is a matter of debate [p10]. The most studied

case is that of synchronous operation with Fixed Length Packets (FLP) [59-62], but more recently, work on asynchronous, variable length packets (VLP) have also been studied [30, 54, 63-67]. Note that synchronous operation with VLP (typically "trains" of packets) have also been studied [68]. Table 3.1 sums up main aspects for the choice between OPS in slotted operation and in asynchronous operation with VLP. The considered aspects are detailed in the following list:

- Packet alignment requires optical synchronisers at the switch interfaces to align the packets. This requires switchable FDLs and is a complex task, as discussed in Chapter 2.5.
- Segmentation/fragmentation of client packets and padding to fill the optical packet increases the packet overhead. In addition, segmentation calls for reassembly of client packets, which increases egress node complexity. E.g. with FDL buffering and/or deflection routing for contention resolution, one needs to determine whether a missing packet fragment is lost or simply delayed.
- Internal blocking occurs when the scheduler cannot utilise free capacity on an output fibre, due to blocking internally in the switch matrix. In slotted operation, a Rearrangably Non-Blocking (RNB) switch may have similar loss rates as a Strictly Non-Blocking (SNB) switch. Advantages of RNB switches compared to SNB switches include decreased component count, as for Clos-based broadcast-and-select (B&S) architectures [69]. In wavelength routers, the TWC tuning range can be reduced, as studied for slotted [70] and for asynchronous operation [71]. However, scheduling complexity increases as all the connections must be taken into account to configure the switch matrix. Since this must be accomplished within a slot period, RNB architectures tighten the control plane bottleneck.
- Contention occurs when an output fibre does not have the capacity to accommodate all packets destined for it. Similar to the performance difference of the slotted and the unslotted ALOHA protocol [44], contention in OPS is minimised for slotted operation [72].

Table 3.1. Main aspects for OPS operation mode

	Slotted (Synchronous, FLP)	**Asynchronous, VLP**
Packet alignment	Required	Not required
Segmentation, Reassembly, Padding	Required	Not required
Internal blocking avoided with:	RNB and SNB switch matrix	SNB switch matrix
Contention	Minimum	Higher

As indicated in Fig. 3.2, we consider either the synchronous, FLP mode or the asynchronous, VLP mode to be the better choice for OPS. Which is to be preferred depends on the weighting of the pros and cons of these options, taking both the technology status and network context (client layer characteristics and service requirements) into account. Since asynchronous operation provides a better match with Internet's non-uniform packet length, the interest has recently surged for this operation mode. This is the operation mode assumed in this Ph.D. project.

3.2.5. Packet and Burst Format

In a packet/burst switched network paradigm, each network layer encapsulates higher layer packets, thereby adding an *overhead*. In the OPS/OBS layer, successful processing and switching typically dictates packet fields for control information, synchronisation pattern(s) and optical guard bands (OGBs), as illustrated in the example of Fig. 3.3.

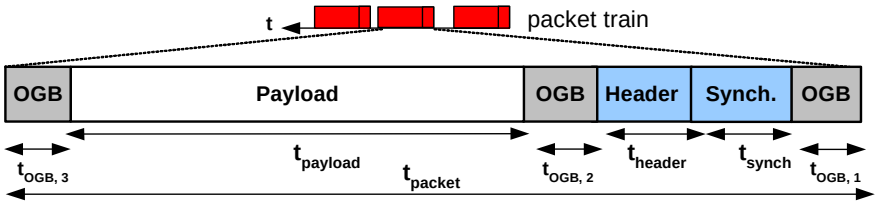

Fig. 3.3. Example of optical packet format

The overhead, defined in (3.1), describes the proportion of time the switch matrix spends settling the switch and transmitting the non-payload fields depicted in Fig. 3.3, relative to the payload duration. The durations of these fields are technology dependent. E.g. the header duration depends on the header encoding method, and the packet synchronisation/preamble field must contain a pattern long enough to allow a stable clock-recovery with unambiguous start-of-packet detection. OGBs are required to accommodate jitter in e.g. the header insertion process and between packets in a packet train. In addition comes the switching time, t_{switch}, during which the considered switch matrix path cannot be exploited.

$$Overhead_{OPS} = \frac{t_{packet} + t_{switch} - t_{payload}}{t_{payload}} = \frac{\left(t_{header} + t_{OGB} + t_{synch} + t_{switch}\right)}{t_{payload}} \quad (3.1)$$

OPS/OBS networks should be designed to handle a certain load offered by the client layer, but the overhead creates a need for the OPS/OBS

network to actually be designed for a higher "optical" load. Since contention increases with the load, the overhead should be reasonably low to efficiently utilise the resources in the optical layer. Since the penalty from the overhead counteracts the statistical multiplexing advantage of OPS compared to OCS, the packet overhead should not exceed ~10 %, to preserve a significant statistical multiplexing gain.

Since the overhead is technology dependent, thus hard to accurately quantify, it is neglected in the performance studies in this thesis. Nevertheless, the following examples illustrate the interplay of payload length, packet format and switching times, further discussed in [p8]:

- At 10 Gbit/s channels, for a serial header packet format (assuming 4 B (byte) header at 2.5 Gbit/s) with a total of 12 ns reserved for synchronisation field and OGBs:
  - The shortest IP packets, i.e. 40 B TCP acknowledgements packets cannot be transported with an overhead below ~77 %.
  - 323 B and 435 B payloads are sufficient to limit the overhead to 10 % with 1 ns and 10 ns switching time, respectively.
  - For 1500 B payloads (typical IP Maximum Transfer Unit length), 95 ns switching time can be tolerated, whilst respecting the 10 % overhead.
- For Sub-Carrier Modulation (SCM) headers (i.e. no field required for headers, as discussed in Chapter 3.3), but with the same synchronisation field and OGBs:
  - The 40 B TCP acknowledgements packets cannot be transported with an overhead below ~37.5 %.
  - 163 B and 275 B payload are sufficient to limit the overhead to 10 % with 1 ns and 10 ns switching time, respectively.
  - For 1500 B IP packets, 105 ns switching time can be tolerated, whilst respecting the 10 % overhead.

These examples show that little packet aggregation is required, with 1-10 ns switching times, if the average payload length is representative of the packet distribution found in the Internet, which is ~400 B [73]. The decreased overhead of SCM has an significant impact only for short to medium length payloads. Finally, note that the overhead increases with the payload bitrate. For the same packet format, upgrading to 40 Gbit/s requires 1240 B and 600 B payload length for serial and SCM headers, respectively, ignoring the switching time. Hence, moving to higher bitrates calls for packet aggregation, or reduction of the length of the non-payload packet format fields.

OBS has little overhead as long as the resources for conveying the BCPs are small relative to the resources used for data transmission. E.g. using

one WDM channel out of 32 per link can easily be tolerated. Similar to OPS, bursts need OGBs to accommodate finite switching times, but their relative impact decreases due to increased payload durations. Hence, higher switching times can be tolerated for OBS than for OPS.

However, for sub-ms burst durations, note that e.g. MEMS based OXC technology, with ms-range switching times [p2], gives excessive overhead. The most widely studied sub-ms switches actually have switching times below 100 ns [p2]. Consequently, it may turn out that the switching technology for OBS may be similar to the one used for OPS (e.g. using wavelength routers or SOA gates). This reduces the potential benefit from large burst durations with respect to the switching matrix.

3.3. Node Design

This chapter is based on findings in [p7].

3.3.1. Node Design in OPS and OBS

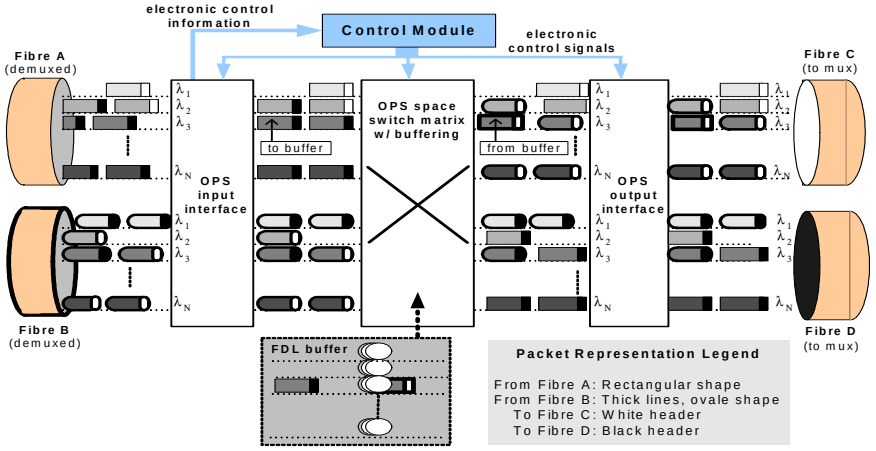

Fig. 3.4. Generic OPS node with FDL buffers in slotted operation.

Fig. 3.4 illustrates main building blocks for OPS/OBS switches, to provide the context for the remainder of Chapter 3.3.

In OPS, the *input interface* taps a fraction of the power of incoming signals, which is used to detect the preamble, marking the packet/burst arrival. A synchronisation pattern enables clock recovery (CR) of packet header, to read the control information. A fixed-length FDL delays the data whilst the scheduling takes place. In OBS, a single burst mode receiver (BMR) is sufficient per input link at the nodes to retrieve the control information. In either case, the control information is transmitted in electronic form to the control unit. The input interface also monitors incoming signals and conditions them as required, e.g. through power equalisation, regeneration and packet alignment (in slotted operation).

For each packet/burst the *control unit* makes a forwarding table lookup, and is responsible for implementing the scheduling policy by identifying a suitable switch matrix path and by resolving contention. If needed, the control unit identifies new control information to encode in the packet header or the BCP. The control unit controls the switch matrix, the FDL buffers and the interfaces to implement the scheduling.

The *switch matrix* influences the node performance by its switching time, maximum throughput, internal blocking properties and signal degradation.

The *output interface* implements control information updates and conditions the signal, if required.

Most of these functions and components are demonstrated in the labs, but are still on the research stage. Of particular importance is the analogue nature of the network transparency, which puts stringent requirements to performance monitoring, signal regenerators and amplifiers that should compatible with large power fluctuations.

3.3.2. Control Plane Design

Both OPS and OBS nodes should contain an electronic routing unit that maintains a routing table, used to generate forward lookup tables, used by the control unit to find the data units' output link. In the Internet, routing information is updated on a minute time scale [74], and the dynamics of this unit is at least on the ms time scale, since control plane communication speed is limited by propagation delay. Some OBS proposals assume that the network's relative delays of control plane and data plane are known, to correctly set BCP time offset; otherwise the routing is similar. On the other hand, OPS and OBS have different requirements to lookup in forward tables and scheduling of packets, as discussed below.

3.3.2.1. Optical packet switching control plane design

Since all-optical logic currently is very immature, electronics should be responsible for the control units more complex processing tasks, such as scheduling and identification of new headers. The sum of processing- and switch fabric reconfiguration time should not exceed the average packet duration, to avoid a data bottleneck at the switch input. Hence, the control unit should be orders of magnitude faster than the routing unit. This potential problem is termed the *control plane bottleneck*. To alleviate this bottleneck, one may need to improve the speed of electronic lookup, i.e. decrease memory access time and the number of accesses needed [74], e.g. by using efficient lookup algorithms. Furthermore, the OPS scheduling complexity must be limited, compelling simple QoS differentiation algorithms, as studied in Chapter 5.

Header processing consists of the following functionalities, which all can be performed optically:

- Header retrieval. Separate “raw” header and payload directly, or obtain a copy of the header.
- Header recovery. Convert control information to a form compatible with forward lookup.
- Forwarding table lookup. Use the control information to consult the packet forward lookup table.
- Header erasure. Erase the old optical header attached to the payload, unless already separated.
- Header reinsertion. Attach a new header to the payload.

A short description of the main methods to perform these functions follows, and is summed up in Table 3.2. Among the discussed methods, serial header, sub-carrier modulation and λ-DPSK and λ-FSK are suitable for electronic lookups, and often they have quite low bitrate header signal to enable low-cost electronics in the BMRs that convert the control information to the electronic domain.

Serial headers have Intensity Modulated (IM) headers before payload. The header is read by direct detection in electronics by a BMR. Using a lower bitrate of the header is a possibility, but this increases the overhead, as discussed in Chapter 3.2.5. Headers can be erased by fast SOA gates [23], and reinserted by couplers. An alternative is to use a two-stage WC configuration: The probe is turned on just after header exits the cross-gain modulated (XGM) WC, followed by an interferometric WC (IWC) that has a probe with the new header IM onto it, during the packet’s header field, combined with bias shift of IWC [23, 51].

Sub-carrier modulation (SCM) combines Amplitude Shift Keyed (ASK) electrical payload and header signals, modulated at baseband and Sub-Carrier Frequency (SCF), respectively. This signal drives a Mach-Zehnder Modulator (MZM), applied to a CW probe. SCM reduces overhead by transmitting header and payload simultaneously, and relaxes the timing accuracy needed for inserting a new optical header. The header is read in electronics, by High Pass Filtering (HPF) of O/E converted signal, followed by Homodyne Detection (HD) to enable direct detection [75]. Practical demodulation limits the SCF to 10-20 GHz, which again limits the payload bitrate (since the modulation bandwidth cannot overlap with that of the SCM header around the SCF). The header can be erased by use of optical Fibre Fabry-Perot (FFP) filters, and reinserted by a MZM. Alternatively, header and payload can be separated by Fibre Loop Mirrors (FLM) and reinserted by a two-stage WCs configuration [51].

λ-DPSK and *λ-FSK* are quite similar in that they see the control information as a two dimensional label: one part is the optical carrier frequency of the packet, the other part is the "orthogonally" modulated Differential Phase Shift Keyed or Frequency Shift Keyed (DPSK/FSK) label [76]. Practical realisation of DPSK is hindered by the requirement for very low laser linewidth. IM/FSK does not suffer from this, and enables a simpler direct detection system. Still, a critical issue of IM/FSK is to correctly set the extinction ratio of the IM payload, since successful FSK detection prevents the '0' level to go too low, whilst sufficient IM extinction ratio must be kept.

Considering approaches with optical forwarding, both *multi-wavelength-* and *Optical Bipolar Shift Keyed (OBSK)* headers have been demonstrated [77, 78]. The former decreases spectral efficiency in WDM systems, since it uses additional frequency channels to encode the header information. The latter is a serial header technique, but very high modulation rate of header ensures low overhead. The optical correlators are based on optical filtering in wavelength- or time domain, and the correlation signal is time-gated, thresholded and O/E converted so that an electrical pulse can open e.g. an SOA gate, in case of match. In the control plane, a copy of the packet header is sent to a set of *N* correlators in parallel, each controlling an optical gate. These designs prevent efficient decoupling of data- and control plane design. This limits control plane scalability, since an NxN switch requires N^2 correlators and since headers suffer a 1:*N* loss.

Table 3.2 sums up these header processing approaches, showing that the proposed methods typically are implemented using WCs. Note that the table uses the classification of WCs from Chapter 2.6.

Electrical header reading and lookup seems more attractive than optical solutions, since it is simpler to reconfigure forwarding tables, as well as decouple control- and data plane architectures. OPS, as intended in this thesis, is thus different from *all-optical* packet- switching approaches, where control functionalities such as forwarding are carried out in the optical domain. Due to their increased optical complexity, we consider that such concepts are further from implementation.

Table 3.2. OPS processing of different header formats (functions in bold are performed electronically).

Processing function →	**Retrieval**	**Recovery/ Reading**	**Table Lookup**	**Erasure**	**Reinsertion**
Serial I [23]	Coupler	**O/E conv.**	**Electronic**	Optical SOA gate	IM tunable laser + Coupler
Serial II [51]	Coupler	**O/E conv.**	**Electronic**	Gated XGM FIOWC	IM probe of 2nd FIWC (IWC) w/ bias shift
SCM I [75]	Coupler	**O/E conv. +HPF+HD**	**Electronic**	FFP notch SC freq.	Modulate MZM arms at SCF
SCM II [51]	Coupler	**O/E conv. +HPF**	**Electronic**	XGM FIOWC	IM at SCF FIWC (IWC) probe
SCM III [51]	FLM	**O/E conv.**	**Electronic**	(Already separated)	XGM FIOWC+ IM at SCF FIWC (IWC) probe
λ-DPSK [76]	Coupler	**O/E conv.**	**Electronic**	WC	Phase Modulation
λ-FSK [76]	Coupler	BPF+**O/E conv.**	**Electronic**	Swapping by TWC to Frequency Modulated probe	
Multi-λ [77]	Coupler	FBG correlator	Optical	Optical BPF	CW Laser + Circulator and FBGs + 1xN switch
BPSK [78]	Coupler	BPSK correlator	Optical	Swapping by optical control solitons and Cross Phase Modulation in DSF	

3.3.2.2. Optical burst switching control plane design

The out-of-band control information, increased burst size and use of offset makes the OBS control plane differ from the OPS control plane. Increasing burst duration by one to three orders of magnitude, compared to OPS packet duration, gives more time for processing, so that OBS has the potential of removing the control plane bottleneck.

It has been proposed to use variable offset in the Just Enough Time (JET) scheme for QoS differentiation [57], i.e. use an additional offset for QoS bursts, compared to low CoS bursts. Hence, high CoS bursts reserve capacity in the core nodes' burst schedulers earlier than the low CoS bursts. The reservations will therefore be made in a rather lightly loaded system, yielding lower burst loss rates. To achieve a certain isolation of the different CoS loss rates requires a certain QoS time offset, which decreases with the system's wavelength count [79]. The reservation window increases with the QoS offset, and may eventually put stringent

requirements on the control unit's memory requirements and the complexity of the scheduling algorithm [p8]. Implementing a Reserve a Fixed Duration (RFD) scheme with time offset for QoS differentiation may therefore reintroduce the control plane bottleneck in the core nodes. QoS differentiation has also an impact on the electronic memory requirements for burst assembly in ingress nodes, since the number of buffer queues is the product of the number of CoS and egress nodes. Furthermore, the time spent in the buffer, and thus the overall buffer sizes, increases linearly with the offset. The required buffer size of an ingress node that should on average distribute 320 Gbit/s overall, to 50 nodes with 5 CoS and average burst durations of 100 μs, has been estimated to 300 MB [80].

3.3.3. Data Plane Design

This section discusses how to design the data plane of a node. The *efficient throughput* of a switch is the product of the maximum throughput, i.e. switch capacity when no internal blocking, nor contention occurs, and the channel load. For this case study, a switch with 2 Tbit/s efficient throughput using 10 Gbit/s channels is considered a target, and the channel load considered is 0.8 for high resource utilisation. Hence, maximum throughput should be 2.56 Tbit/s, assuming negligible loss.

Considering the switch matrix, the number of switch ports, *N*, is the product of the number of fibres, *F*, and wavelengths per fibre, *W*. This chapter focuses on SNB architectures, suitable both for slotted and asynchronous operation. Since the considered architectures feature ns-range switching times, the switch matrices will be suitable both for OPS and OBS node designs.

3.3.3.1. Switch matrix

The scope is limited to some of the most promising switching solutions for OPS and OBS, namely Broadcast-and-Select (B&S) type switches and wavelength routers. These generic architectures are compared at the end of this section. The versions considered use multiplexers instead of couplers, when any of them would do, to minimise loss and ASE noise. Remark that most architectures are conceived to include buffering solutions, and references show how buffering can be implemented in the architectures. In general, this has a significant influence on component count and switch parameters. However, including this in the comparison is prone to inaccuracies, as logical performance analysis assumes different traffic simulation parameters and buffering technology

performance. For a plain switch matrix comparison, resources for buffering are omitted here. Still, the comparison of switch matrices gives a representative view of the main challenges of each design.

Broadcast-and-Select Type Switches

The principle of the *B&S* switch is illustrated in Fig. 3.5 a). A broadcast stage passively splits the WDM signals, and each output fibre uses a space- and wavelength selection module to select *W* single-channels. Such a module, for *F*=*W*=16, has been integrated on a single board [81], and 640 Gbit/s throughput in asynchronous operation has been experimentally verified [82]. This architecture gives multicast compatible switches, but high loss, induced by the high splitting ratios.

Fig. 3.5 b) illustrates an adaptation of B&S, called *Tune-And-Select* (TAS). The input demultiplexer, couplers and SOA gate based fibre selection is sufficient for switching, whilst the FIWCs enable contention resolution in the wavelength domain. An analysis of physical limitations of TAS (and related architectures), was recently presented [83]. Physical limitations (considering ASE noise from EDFAs and SOAs, crosstalk from demultiplexers and coherent crosstalk from SOAs) gave maximum values of *W*. For *F*=8, the maximum wavelength count, *W*, was 128, 64 and 16 wavelengths, for 2.5, 10 and 40 Gbit/s bitrates, respectively.

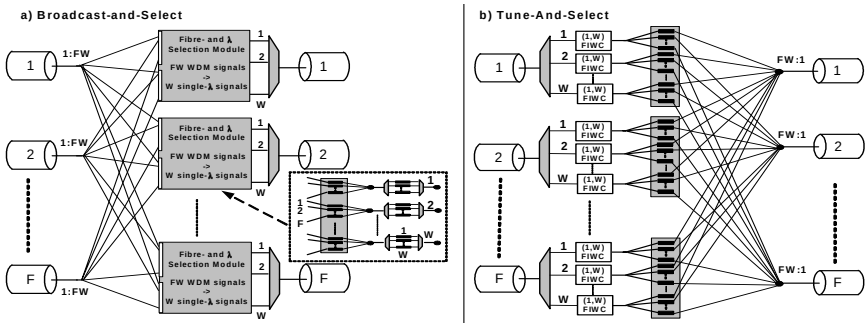

Fig. 3.5. a) Broadcast-and-Select and example of selection module. b) Tune-And-Select.

Wavelength Routers

Wavelength-routers are based on a passive fabric with preconfigured input-output paths depending on the input port and input wavelength. The Uniform-Loss Cyclic-Frequency (ULCF) planar waveguide geometry Arrayed Waveguide Grating (AWG) is well suited for the purpose. A 64x64 AWG with 50 GHz channel spacing was recently reported [84], with loss between 5.4-6.8 dB and crosstalk below –40 dB. The device is compact; a connected and packaged module, also featuring a temperature control, had dimensions of only 14x9x1.3 cm^3.

A single-stage switching matrix uses an AWG with N fixed-wavelength Input Interfaces (II_1–II_N) and N Output Interfaces (OI_1–OI_N). One equips each Input Interface with a (1,N) FIWC for the signal routing, i.e. a FIWC capable of tuning into N WDM channels, one per OI. An example of an II-OI interconnection matrix is given in Fig. 3.6, where the matrix elements indicate the wavelength to which the FIWC at II_i has to be tuned to send the signal to OI_j. As indicated by the grey background, more wavelengths than used in WDM transmission (W=2 in this example) are needed. Hence, FOWCs are required to adapt the signal at the OIs, with the consequence of allowing only single-wavelength signals at the OIs. The FOWCs brings the benefit of full contention resolution in the wavelength domain. The corresponding internal connection pattern of the AWG is illustrated in an example switch matrix, for F=W=2.

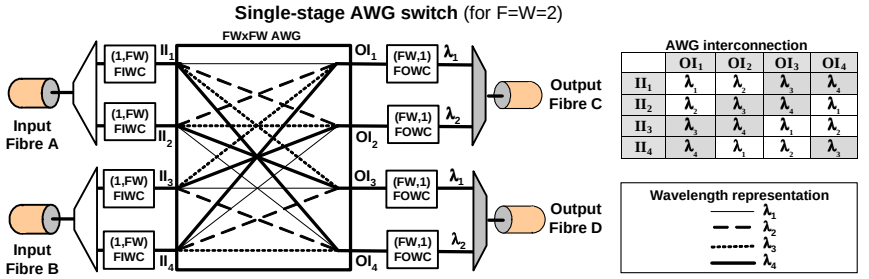

Fig. 3.6. Principle of *N*x*N* AWG and single-stage AWG switch, illustrated for *N*=4 and *F*=*W*=2.

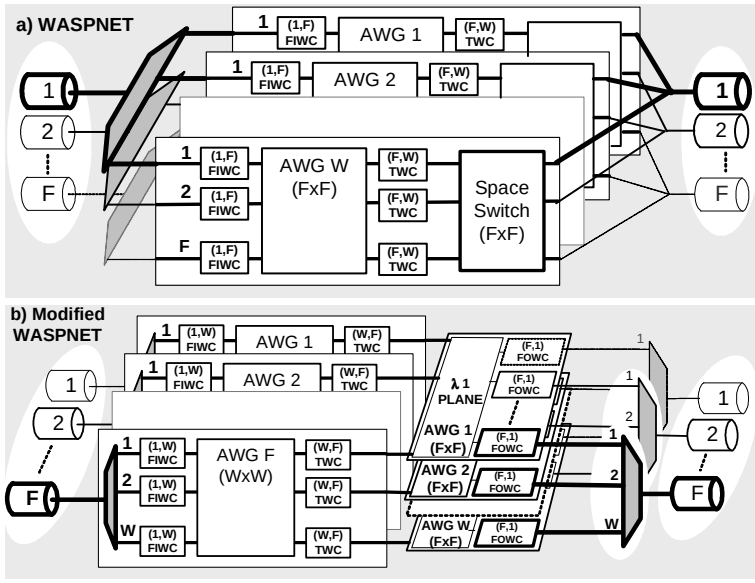

Fig. 3.7. Multi-AWG architectures: a) The WASPNET architecture uses a plane per wavelength, b) the "Modified WASPNET" architecture uses a plane per fibre.

Multi-stage AWG architectures, depicted in Fig. 3.7, have much relaxed requirements to the AWG size and WC tuning range. The *WASPNET* switch [28] is based on wavelength planes and recombination by an *F*x*F* space switch (must be capable of many-to-one switching) to a set of couplers. A variant of the WASPNET switch, "*modified WASPNET"*, was recently proposed [65]; it is based on fibre-planes and replaces the space switch by a second set of AWGs and FOWCs. Both architectures have full wavelength domain contention resolution.

3.3.3.2. Comparison

The discussed switching architectures are compared with respect to the hardware aspects. The scope of the comparison is restricted to: *i*) component count, *ii*) component complexity, and *iii*) number of interconnections. *i*) and *ii*) give an indication of total component cost, and what requirements are put on the component side, e.g. WC tuning range, AWG-, demultiplexer-, and coupler size. *iii*) indicates how many switch matrix internal interconnections that will have to be made; when performed manually, this is a tedious and costly task.

The inventory list is shown in Table 3.3, as a function of F and W. Main component challenges and scaling limitations of each architecture, as discussed below, are identified by bold format. As for the header processing, WCs are often used in the data plane. The WCs' number of allowable input channels, I_λ, and the achievable output channels O_λ, are denoted by (I_λ, O_λ), in addition to the classification of WCs into TWCs, FIWCs, FOWCs and FIOWCs.

Main parameters are graphically illustrated in Fig. 3.8, for a core mesh node scenario (F=4, W=64) that corresponds to a mesh network with a relatively high number of WDM channels per fibre, for efficient contention resolution.

Table 3.3. Scaling and complexity of components for different architectures. Note component parameters in parenthesis, and that the main challenge of each design is indicated by a bold font.

Scaling	FIWC (I_λ, O_λ)	FOWC (I_λ, O_λ)	TWC (I_λ, O_λ)	AWG (size)	SOAs	Space switch (size)	Couplers (size)	Mux (size)	Inter-connections*
TAS	**FW (1,W)**				**F^2W**		FW(F) + **F(FW)**	F(W)	**$2F^2W$ +FW**
B&S					**F^2W+W^2F**		FW(F) +**F(FW)**	2FW(W) +F(W)	**$2F^2W$ +$2FW^2$ +FW**
Single-AWG	**FW (1,FW)**	**FW (FW,1)**		**1(FWxFW)**				2F(W)	4FW
Waspnet	**FW (1,F)**		**FW (F,W)**	W(FxF)		**W(FxF)**	F(W)	F(W)	5FW
Mod. Waspnet	**FW (1,W)**	FW (F,1)	**FW (W,F)**	F(WxW) +W (FxF)				2F(W)	6FW

*Count rules (for interconnects):
1) Two components (e.g. a mux port and an SOA) are interconnected through *one* interconnect (which may require *two* splices).
2) Only switch-internal interconnects are included.
3) Since the aim is a order-of-magnitude scope, terms with F or W exclusively are ignored for simplicity.

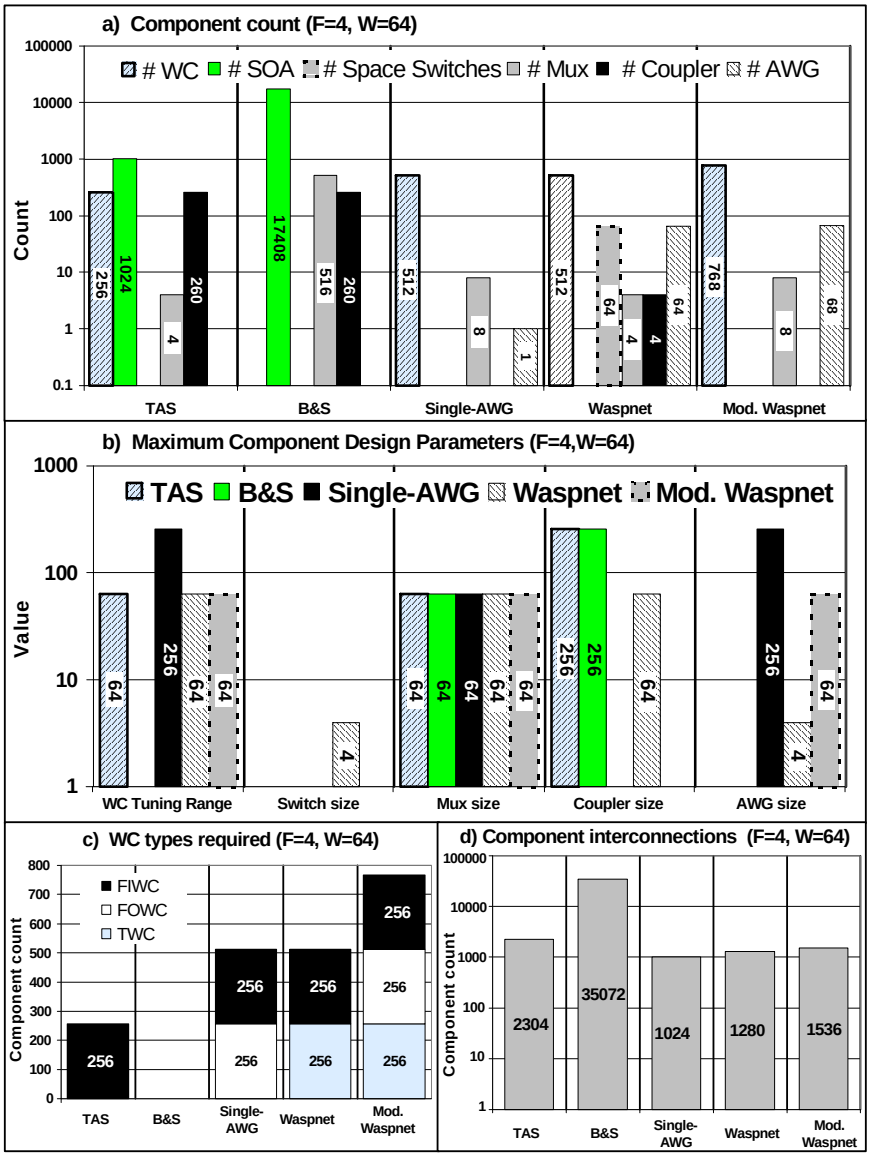

Fig. 3.8. Analysis of different architectures for (*F*=4, *W*=64). Note the log-scale.

Although signal quality issues are not discussed in detail here, note that large couplers have high intrinsic loss. Since B&S and TAS have maximum coupling ratios that scale linearly with the capacity, these architectures may have difficulties with maintaining a sufficient SNR. In addition, total signal path loss must be compensated by amplifiers, adding to CAPEX. However, this loss in TAS and B&S should be compared to

loss that occurs in the other architectures, which depends mainly on loss in AWGs, WCs, space switches and multiplexers.

For a constant $N=FW$ product, some parameters depend on the distribution of the F and W parameters. The most critical dependencies are the SOA count and the number of interconnections in TAS and B&S. Adapting the switch matrix internal (F, W) design parameters with those used in the transmission layer will add the need for N FIOWCs at the input and N FOWCs at the output, but it has the benefit of resolving contention in the wavelength domain.

It is clear that B&S, and to a certain degree TAS, suffers from a very high number of required SOAs and many interconnections. This may reduce their attractiveness, but component- and interconnection cost will depend heavily on achievable component integration level, which also decreases the space consumption. Already, modules with 32 SOA gates have been manufactured, and placed on boards containing four modules [81], but it remains to realise larger PICs, which e.g. include SOAs, couplers and multiplexers. The drawback of high SOA count can further be mitigated by economies of scale, which dictates that a high production level of a component type decreases the cost of that component. A disadvantage with B&S is that there is no inherent wavelength conversion, so that contention has to be resolved by additional resources.

Single-stage AWG switches suffer from high requirements to AWG size and tuning range. Multi-stage AWG designs overcome this limitation, at the expense of using a large number of AWGs and WCs. Compared to WASPNET, the “Modified WASPNET” switch requires more FOWCs, more large AWGs, multiplexers and interconnections. On the other hand, it avoids using couplers and the high number of complex space switches required in WASPNET switch design.

In conclusion, main data plane architectures have significant differences in main cost factors such as component count, component complexity and interconnections. Each architecture faces different scalability challenges, thus the technology progress and cost of different components will to a large degree determine their feasibility and attractiveness.

3.3.4. Combining Data- and Control Plane Functions

An overview of experimental results and proposals for how WCs can be used to regenerate signal quality, erase and reinsert headers was presented in [85]. In addition, [28] outlines how these functions can be performed simultaneously with data plane switching, contention resolution and adaptation to transmission layer, in the WASPNET design. As mentioned

in Chapter 3.3.3.2, adaptation of *F* and *W* can be essential, particularly in B&S designs, and this can be combined with e.g. signal regeneration (based on WCs) at the node interfaces [81]. This also enables full wavelength contention resolution. As studied in 3.3.2.1, WCs are used in a number of header processing schemes. Since the WCs can perform a large part of the data- and control plane functions, it is a key enabler of efficient node designs. To limit component count, space- and power consumption it is desirable to implement as many functions as possible in each WC. However, this must be balanced with the fact that WCs tailored to specific functions, can have limited requirements to tuning range, compared to a multi-capable module.

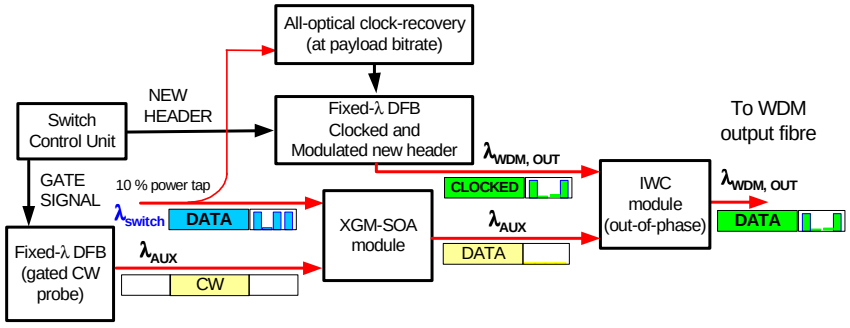

Fig. 3.9. Example of combining serial header erasure- and insertion, with 3R regeneration and wavelength domain contention resolution.

Fig. 3.9 depicts a potential design for application at the switch output interface. It combines header-erasure, header rewriting, FOWC contention resolution and 3R regeneration. The two-stage 3R regenerator is detailed in [56]. This proposal adds the header erasure and reinsertion by gating and modulating the probes of the first- and second stage WC, respectively. At the input, the signal is tapped to enable CR, and the XGM SOA based FOWC converts the packet to a wavelength used internally in the 3R regenerator. This WC provides stable power and polarisation for the second 3R stage. The RZ pulses of the second DFB is adjusted to match those of the packet, and the non-linear transfer function of the Interferometric WC, which can be a FIOWC AA-MZI, provides the 3R regeneration. Contention is resolved by ensuring that all outputs to the same fibre use a different wavelength for the second-stage DFB laser.

Other examples of combining different functionalities in WCs are studied for a hybrid design in Chapter 4.2, and for an OPS MAN ring node design in Chapter 6.2. However, the remainder of this chapter studies contention resolution assuming a generic switch matrix, and aims at reducing overall WC count through WC sharing.

3.4. Contention Resolution

This chapter discusses findings in [p20].

3.4.1. Contention Resolution Methods

Contention occurs when data units at the same wavelengths from different input fibres are switched to the same output link. Contention can be resolved in space-, wavelength- and time domain, listed below and discussed in the following:

- *Time domain*: Buffer all but one contending data units until the requested wavelength is vacant.
- *Space domain*: Either separate contending data units by transmitting them on different output fibres on the same link, or send all but one contending data units towards non-shortest-path nodes, termed "deflection routing".
- *Wavelength domain*: Convert all but one of the contending data units to a different wavelength at the same fibre.

Regarding the time-domain, near-term OPS/OBS networks cannot be store-and-forward networks, as opposed to electrical packet switching networks, as discussed in Chapter 2.6. Optical FDLs [54, 61, 86-89], or electrical buffers [18, 66] can be used to some extent, but their number and length should be minimised to limit packet misordering, noise accumulation [54, 90], space consumption [18, 19, 54, 91], and interface cost. The buffer interfaces consist of additional switch matrix port count, TWCs for WDM buffers [54, 87, 89] or O/E converters for electrical buffers [18, 66, 91], respectively.

Regarding the space domain, sending packets on different fibres between the same nodes [92], suffers from increased transmission cost (fibres, amplifiers, couplings, monitoring) that follows using multiple fibres per link. This approach is not widely studied in OPS research, but a number of studies resolve contention in the space domain, through deflection routing. However, this increases the average hop-count, which may actually reduce throughput in asynchronous networks [93]. Another drawback is the significant increase in packet delay-jitter, leading to packet misordering, which makes the egress node reassembly process more complex [p4, 91].

Regarding wavelength domain contention resolution, this method reduces the need for buffers [94], and may even enable bufferless optical packet

switches [95]. This approach requires WCs to transfer the data from the original input WDM channel onto a new output WDM channel. Contention can then be resolved by ensuring that two packets going to the same output fibre are not assigned the same output wavelength. With a FOWC at each output, as demonstrated in [96], the full potential for contention resolution in the wavelength domain is available. It has been shown that the performance increases significantly with the number of wavelengths per fibre [66, 97]. This will be further discussed below.

To reduce overall WC count for contention resolution, a Shared Per Node (SPN) TWC pool can be implemented [98], as shown in Fig. 3.10. This TWC count reduction comes at the expense of an increase in the size of the switch matrix. Hence, the SPN design is beneficial when TWCs are a main cost-factor, bulky, or have a high power consumption, compared to additional ports in the switch matrix. This is particularly relevant if the switch matrix has reached a higher maturity- and production level than WCs. Note that SPN designs are not attractive if FOWCs are anyhow needed in the switch output interface.

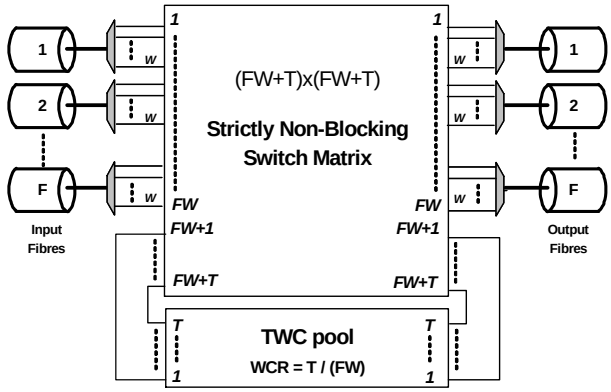

Fig. 3.10. SPN node design with *T* TWCs.

3.4.2. Performance of SPN TWC pools

This study investigates the optical packet switch performance by discrete event-driven numerical simulations in OPNET. Asynchronous operation is implemented by per-wavelength Poisson-based traffic generators, emulating serial packet scheduling by incorporating a FIFO buffer each. A negative exponential packet length distribution is chosen, for an average channel load of 0.7. In all graphs, the simulation points are interconnected by straight lines, and the 95 % confidence intervals are indicated.

We put the optical packet switch under study in a core network context, investigating both a mesh and a dense mesh context, by applying a node adjacency, *F*, of 4 and 8, respectively. Both WDM and DWDM systems are tested, varying the number of wavelengths per fibre, *W*, from 32 to 256.

The relative size of the TWC pool can be expressed by the Wavelength Conversion Ratio (*WCR*) parameter, defined as (3.2), when *T* is the number of TWCs in the pool, and the TWC pool is shared by packets from *FW* potential inputs.

$$WCR = \frac{T}{FW} \tag{3.2}$$

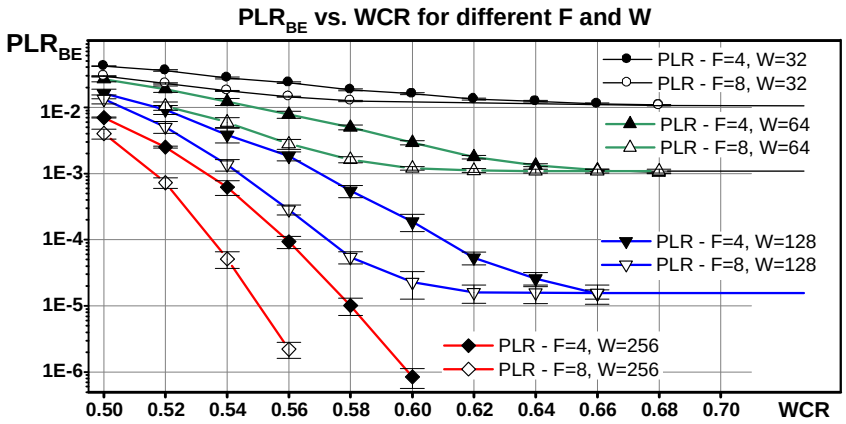

Fig. 3.11. *PLR*$_{BE}$ vs. *WCR* for average channel load of 0.7 *W* of 32-256, for *F*=4 and *F*=8.

The impact of *F*, *W* and *WCR* on the PLR is illustrated in Fig. 3.11, confirming that the size of the TWC pool has a great impact on the contention resolution performance [87, 98]. For each set of system parameters (*F*, *W* and load), there is a certain *WCR* value, *WCR*$_{th}$, above which the PLR does not decrease more than a small percentage, similar to observations in [87]. When *WCR*<*WCR*$_{th}$, TWCs are scarce, increasing the PLR. The PLR at *WCR*$_{th}$, *PLR*$_{min}$, decreases with *W*, as expected from increased statistical multiplexing gains in the wavelength domain when TWCs are no longer a limiting factor, similar to systems without TWC sharing [97].

The potential economic gains of the TWC pool increases with decreasing *WCR*$_{th}$. The value of *WCR*$_{th}$ decreases with decreasing channel load, since this reduces contention and thus the need for contention resolution

resources. Furthermore, WCR_{th} decreases with increasing F due to the increased sharing of TWC resources in the SPN design.

We conclude that the SPN pool design is an attractive way to resolve contention, when reduction of WC count is a main design target. The relative amount of TWCs needed depends mainly on channel load and fibre count. The PLR floor decreases by nearly three orders of magnitude when quadrupling the wavelength count per fibre from W=32 to W=128. Hence, DWDM systems with many channels represent attractive application for node designs resolving contention in the wavelength domain.

3.5. Overcoming Scalability Constraints

This chapter discusses the findings in [p15] and [p18].

3.5.1. Shared Per Waveband Plane Design

Fig. 3.11 shows how PLR decreases with increasing *W* and *WCR* (and with increasing *F*, for a low *WCR*) in the SPN TWC pool design. It was argued in Chapter 2.2 that WDM transmission supports wavelength counts up to 320 channels per fibre. However, as pointed out in Chapter 3.3.3.2, high port counts put quite hard demands on a number of parameters, depending on the switch architecture. Consequently, switch scalability may be prevented by e.g. limitations on TWC tuning range, AWG size, B&S splitting loss, or integration level of the SOA gate selection stage.

Parallelism is a common approach to overcome scalability constraints, and is the main idea behind the proposal for the Shared Per Waveband Plane (SPWP) switch design [p15].

As illustrated in Fig. 3.12, the switch design exploits the wavelength domain to separate the *W* channels on each fibre into wavebands with *w* channels, using passive waveband demultiplexers. The same wavebands from different input fibres are then switched in the same Waveband Plane (WP). Each WP can have its own TWC pool to resolve contention, thus the name “Shared Per Waveband Plane”.

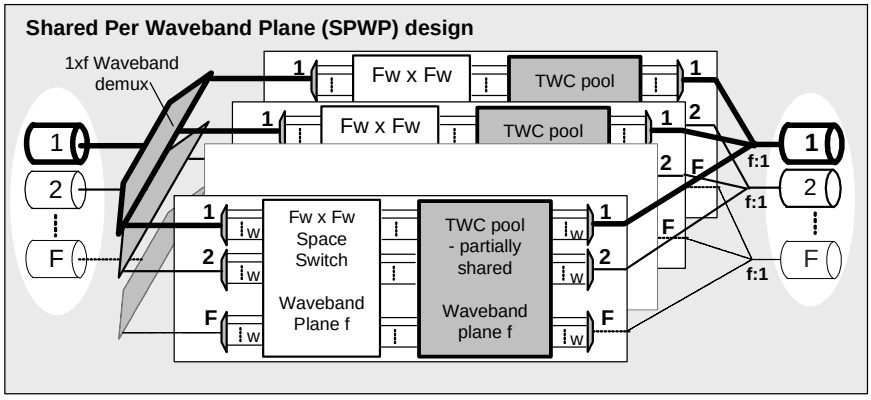

Fig. 3.12. Schematic illustration of SPWP design. The switch matrix in each WP is a SPN design, shared only by that WP.

The waveband based multi-plane design is fundamentally different from the (single) wavelength-plane design [27] and fibre-plane design [65], discussed in Chapter 3.3.3.1. In these switches, the active recombination results in similar performance as one large single-plane switch. This comes at the expense of increased switch complexity and that the dimensions of each plane are directly linked to the WDM transmission layer, through the fibre- and wavelength count. Finally, these proposals do not consider TWC sharing. A different parallel design, referring to this SPWP design has later been discussed [88].

The following list sums up main potential benefits of the SPWP design, based on discussions in [p15, p18]:

- Overcome scalability constraints, achieved by reduced optical switch maximum parameters, combined with simplified control plane scheduling, since no coordination of WP switching is assumed.
- Enable ingress-based QoS differentiation, by mapping different CoS onto different wavebands at the ingress nodes. Each WP can have different statistical PLR, by one or more of the following approaches:
  - Different wavelength channel count in the WPs.
  - Different SPWP TWC pool size in the WPs (i.e. different *WCR*).
  - Different loads in the WPs.
  - Add FDLs to the high quality WP SPWP pool.
- Enable modular capacity upgrade scenarios. When more capacity is needed in the network, one can add wavelength channels in the WDM systems, within an unused waveband, and increase capacity of the switches, by adding a new WP, without interfering with existing WPs.
- Enable hybrid networks. One can have differents WPs reserved for OPS and OCS. The switching technology of the OCS WPs can be be based on “conventional” optical switching technology, e.g. MEMS OXC, and associated control plane solutions.
- Enable an OCS to OPS migration scenario. One can first move from an OCS to a hybrid network, and ultimately to a pure OPS network (if desired), by upgrading the technology used in each WP.

Conform to the results of Fig. 3.11, the PLR of a WP with $w<W$ channels increases with decreasing w. Hence, it is important to point out that the passive recombination of the SPWP increases the average PLR of the switch. However, it may be that not all traffic requires very low PLRs. In this case the SPWP based QoS differentiation mentioned above, may be beneficial, and the design may help overcome scalability constraints and reduce overall TWC count.

3.5.2. Increasing SPWP QoS Granularity

The SPWP concept was refined by applying a QoS differentiation algorithm, internally in one of the WPs [p18]. The algorithm is the same as the one described in Chapter 5.2. This QoS approach, termed the 'SPWP+' design, thus increases the QoS granularity. In a case study of an OPS switch with two WPs, this strategy increased the number of CoS from two to three, by applying WP-internal QoS differentiation in one of the WPs. The case study has a channel load of 0.7, and the traffic matrix contains 3 CoS with PLR thresholds of $2x10^{-2}$, $2x10^{-3}$, and $2x10^{-5}$.

The four strategies listed below were compared to dimension the switch, by mapping the CoS onto a suitable combination of WPs, each with an optimised size and *WCR*.

- BE FOWC design: A BE design (single WP, w/o internal QoS differentiation) with a FOWC per output port, which must respect the most demanding PLR threshold.
- BE SPN design. A BE design (single WP, w/o internal QoS differentiation) using a SPN TWC pool, which must respect the most demanding PLR threshold.
- SPWP design. Maps each CoS onto a proper WP (each w/o internal QoS differentiation), each dimensioned to respect the corresponding PLR threshold.
- SPWP+ design. Maps the CoS with the highest- and lowest PLR threshold onto the same WP, and uses a WP-internal QoS differentiation algorithm to differentiate them. Map the third CoS onto a proper WP. Dimension both WPs and QoS parameters to respect all three PLR thresholds.

The study showed that [p18]:

- The SPWP design could not obtain the lowest PLR of $2x10^{-5}$ at the studied load.
- Compared to the BE FOWC design, the SPWP+ design enabled a 45 % reduction in WC count and a 22 % reduction in maximum switch matrix size. These benefits come at the expense of replacing FOWCs by TWCs, more complex scheduling and ~55 % increase in overall switch port count.
- Compared to the BE SPN design, the SPWP+ design obtained a similar TWC count, whilst reducing the TWC tuning range and maximum switch matrix size by 50 %. These benefits come at the expense of more complex scheduling.

Note that the traffic matrix and design parameters have a large impact on which design to prefer (and their internal parameters, such as SPWP pool

size, and QoS differentiation parameter setting). Due to increased economy of scale, all strategies perform better when W increases. For a given W and fixed PLR thresholds, the BE FOWC and BE SPN strategies are insensitive to the fraction of traffic from each CoS. The pure SPWP strategy benefits from an increased fraction of the most demanding CoS, since this increases the wavelength count of the WP containing the most critical CoS. The SPWP+ strategy, on the other hand, benefits from a decrease in the fraction of the high priority CoS.

In conclusion, the SPWP+ design enables WC count reduction, is compatible with hybrid- and migration scenarios, and reduces hardware scalability constraints. Hence, it is a promising optical element candidate, when facing hardware scalability constraints in a multiple service class optical network environment.

3.6. SPN Pools with FDLs and TWCs

This chapter incorporates [p21], published at ECOC 2004.

Replacing Shared-Per-Node Wavelength Converters by Fibre Delay Lines in an Asynchronous Optical Packet Switch

M. Nord [1, 2]

(1) Research Center COM, Technical University of Denmark, B-345V, 2800, Lyngby, Denmark. (mn@com.dtu.dk)
(2) Telenor Research & Development, N-1331, Fornebu, Norway

Abstract This study shows how a significant amount of tuneable wavelength converters can be replaced by fiber delay lines in a port-constrained asynchronous optical packet switch with a shared contention resolution pool.

A. Introduction

In Optical Packet Switching (OPS) and Optical Burst Switching (OBS) networks, the Packet Loss Ratio (PLR) and packet delay are important Class of Service (CoS) parameters. A low-complexity contention resolution scheme is a main OPS design target, to enable cost-effective and practically realisable network designs. Asynchronous operation is potentially simpler than slotted operation [99], and this is the operation mode in this study. Tuneable Wavelength Converters (TWC) can alleviate the increased blocking that follows [72]. Buffering, by use of Fibre Delay Lines (FDLs), can help resolve contention in asynchronous [54], and in slotted operation [89]. FDLs can be either single-wavelength or WDM-based. The latter uses less total fiber length, but requires couplers and demultiplexers at the FDL inputs and outputs, respectively, as well as TWCs to fully exploit their capacity. This reduces the space consumption advantage of WDM FDLs, and it increases the number of interconnects. The reduced FDL count also leads to coarser buffer time granularity in incremental FDL buffer designs. In OPS, contention resolution combining deflection routing, FDLs and a Fixed Output WC (FOWC) per switch output port has been experimentally demonstrated [96]. The total WC count can be reduced by equipping the switch with a Shared Per Node (SPN) TWC pool, exploiting the fact that many packets

do not need wavelength conversion [98]. PLR then decreases towards that of the FOWC approach with increasing pool size [p15]. The PLR can be further decreased, when *adding* FDL buffers [87]. However, this increases the pool size and thus the switch port count, which may be constrained and impractical to modify in installed systems.

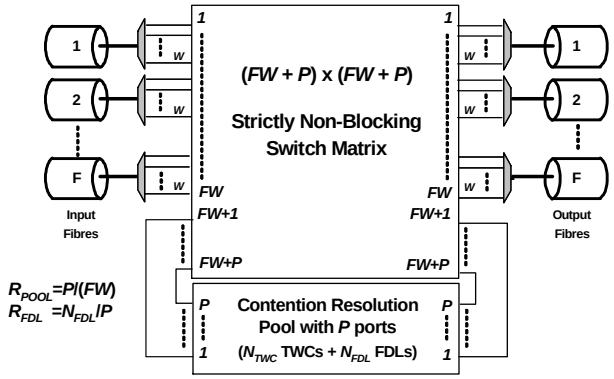

Fig. 3.13. The SPN TWC and FDL Pool node design.

This study investigates performance of a minimum hardware count approach, by *replacing* TWCs by single-wavelength FDLs, thereby reducing TWC count and maintaining the switch matrix port count. Furthermore, the scheduling is of low complexity, since it does not keep track of packet durations, as opposed to buffer output scheduling [54], and the OBS Just Enough Time (JET) scheme [87]. Suitable FDL array designs are investigated, and a simple FDL output reservation scheme, '*SoftRSV*', is proposed.

B. Node Design

Performance is studied by event-driven simulations of a single switch in a core mesh network context, assuming 4 input- and 4 output fibres (*F=4*), studied for 32 wavelengths per fibre (*W=32*). Per-wavelength channel traffic generators emulate the switch input traffic. The packets arrive according to a Poisson process, and the packet duration is negative exponential distributed. The channel load of 0.6 enables PLRs of 1.88×10^{-3}, when having a FOWC per channel. The relative size of the SPN feedback contention resolution pool, R_{POOL}, is expressed as *P/FW*, where *P* is the number of pool ports, as illustrated in Fig. 3.13. To limit the pool size, without increasing the PLR beyond 10%, *P* is set to 74, ($R_{POOL}=0.578$), resulting in a PLR of 2.04×10^{-3}.

The electronic control unit forwards packets based on incremental searches in a table containing the state of the total *FW* output fibres' output wavelengths (owl) and of the FDL array, according to the

algorithm defined in steps *i)-iv)* below. The algorithm is repeated for buffered packets, thereby allowing buffer recirculation. Each FDL can contain multiple non-overlapping packets. Step *i)* minimises the number of required TWCs, by favouring Direct Mapping (DM) [p15]. This consists of starting the search for a free owl at the requested output fibre, at the same wavelength as the packet's input wavelength (iwl). The text in parenthesis in steps *ii)* and *iii)* constitutes the *SoftRSV* scheme, aiming to increase the probability of DM for buffered packets. Note that the *SoftRSV* somewhat increases the scheduler complexity, by introducing a second state-table, to track the current number of SoftRSV's made for each of the *FW* owls.

- *i*) Direct Mapping if owl=iwl is free.
- *ii*) Otherwise use TWC if available and a free owl exists. (Avoid SoftRSV'ed wavelengths if free, non-reserved owls exist).
- *iii*) Otherwise use a FDL if there is a free FDL input. (SoftRSV owl=iwl at output fibre for the buffer duration).
- *iv*) Otherwise discard packet.

C. Results

The ratio of replaced TWCs is described by $R_{FDL}=N_{FDL}/P$. The design target of this study is to find how many TWCs can be replaced by FDLs in the pool, without suffering any PLR penalty. R_{FDL} is therefore increased from 0, to a value with equal PLR as for $R_{FDL}=0$, termed $R_{FDL,EQU}$. Fig. 3.14 illustrates the PLR vs. R_{FDL}, for an FDL design with incremental FDL length, for different basic delay units, D_{FDL}. In this case, the FDL number n has a length of n times D_{FDL}. Although beyond the scope of this study, Fig. 3.14. also shows that for limited R_{FDL}, the PLR can be decreased by more than a decade. The dotted line illustrates the PLR that would occur if the TWCs were *removed* instead of replaced. The rather sharp PLR increase proves that TWCs are actually needed at this R_{POOL} value for this load, thereby justifying the term "replacing" TWCs. In the contrary case, "adding" FDLs would be a more appropriate term.

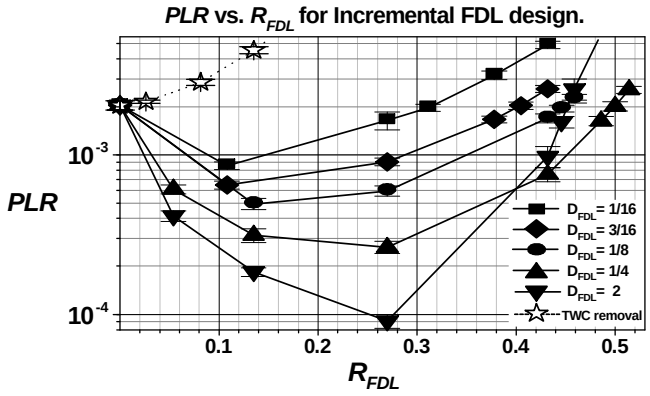

Fig. 3.14. PLR vs. R_{FDL} with 95% confidence intervals.

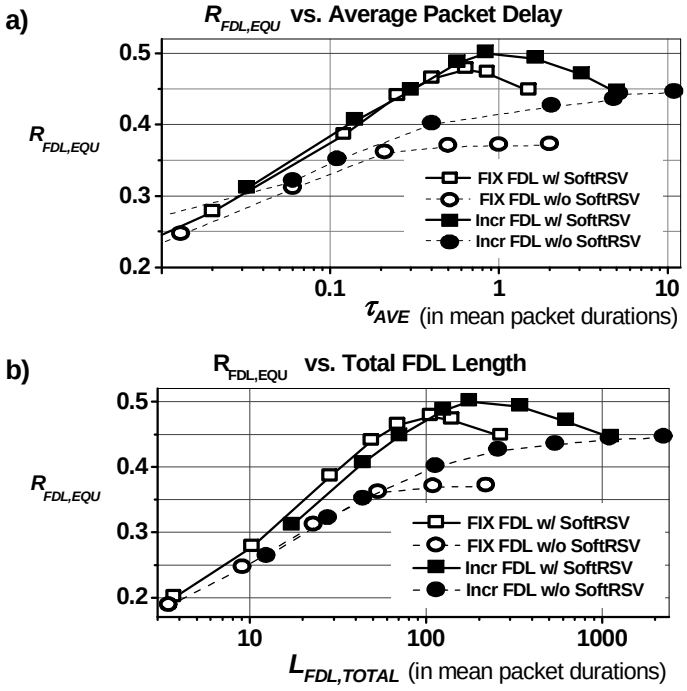

Fig. 3.15. a) $R_{FDL,EQU}$ vs. τ_{AVE} , b) $R_{FDL,EQU}$ vs. $L_{FDL,TOTAL}$.

The $R_{FDL,EQU}$ is identified for a number of D_{FDL}, both for the fixed FDL length design ("FIX FDL"), and an incremental design ("Incr FDL"). The average delay experienced by packets in the FDL buffers, τ_{AVE}, vary with $R_{FDL,EQU}$, D_{FDL} and the FDL design. This is the case also for the overall FDL length, $L_{FDL,TOTAL}$. Fig. 3.15 illustrates the switch dimensioning trade-off between the achievable $R_{FDL,EQU}$ and these parameters, with and w/o the *SoftRSV* scheme. The main trends illustrated are:

- When employing the *SoftRSV* scheme, there exist an optimum D_{FDL}.
- The *SoftRSV* scheme enables replacing ~50% of the TWCs.
- W/o *SoftRSV*, the $R_{FDL,EQU}$ increases more slowly towards a value of 0.35-0.40, after which increases only come at the expense of high τ_{AVE} and excessive $L_{FDL,TOTAL}$.
- With *SoftRSV*, the incremental design obtains a slightly higher $R_{FDL,EQU}$ than the fixed FDL design.
- With *SoftRSV*, in the region of most interest (up to the maximum value), the incremental design requires higher $L_{FDL,TOTAL}$, but yields approximately equal τ_{AVE}, as the fixed FDL design.

D. Discussion

For pure TWC pools, packets are lost either when: *i)* all wavelengths on the requested output fibre are busy, or *ii*) when a DM cannot take place and all TWCs in the pool are busy. FDLs are the only remedy of *i)*, and can also provide a solution to *ii*). Moreover, particularly when using *SoftRSV*, the use of FDLs favours DM, partly compensating the reduced TWC count. This explains that replacing TWCs by FDLs improves the PLR, even when TWCs are scarce.

Employing the *SoftRSV* scheme, there are two counter-working effects: *i)* increased buffer capacity for high D_{FDL}, and *ii*) decreased chance of packets being able to exploit their *SoftRSV* for long buffering periods, induced by high D_{FDL}. This explains the existence of an optimum D_{FDL}.

When buffer capacity is scarce, the incremental design has been shown to obtain lower PLR for the same FDL port count [89], for the same D_{FDL}, thus for much higher $L_{FDL,TOTAL}$. The latter should be low, to limit buffer space consumption. This study shows less pronounced differences, since the performance is compared for the same $L_{FDL,TOTAL}$. Hence, both designs are viable options. As an example, assuming a mean packet duration of 1 µs, choosing the fixed design with a D_{FDL} of 1 µs enables replacing ~39% of the TWCs. This can be realized with a total of ~6 km FDL length, and with a τ_{AVE} of ~0.12 µs, negligible to network propagation delay.

E. Conclusions

The *SoftRSV* algorithm enables replacing a significant fraction of TWCs with simple single-wavelength FDLs in the SPN contention resolution pool design, without suffering any PLR penalty. This increases node dimensioning flexibility, enabling cost savings, when FDLs are cheaper than TWCs. This comes at the expense of increased FDL length and packet delay.

4. Hybrid Networks

4.1. Introduction

The term “hybrid” refers to a network that supports two or more CoS that use different switching paradigms, e.g. OCS and OPS. A novel concept to realise hybrid networks, based on separating different traffic types by their relative State Of Polarisation (SOP), was proposed by S. Bjørnstad et al. at ECOC 2003 [p11], and is further detailed in [42].

This project has proposed and demonstrated an optical processing scheme to realise the segregation of the OCS and the OPS CoS. The former is switched by a Static Wavelength Routed Optical Network (SWRON) switch matrix, whilst the latter is switched by an AWGR based optical packet switch matrix. Furthermore, the scheme also separates the header from the payload, and performs the wavelength conversion required for the first stage of the packet forwarding. This work was presented at OFC 2004, and Chapter 4.2 incorporates this paper [p16].

4.2. Demonstration of Hybrid Scheme

This chapter incorporates [p16], presented at OFC 2004.

Demonstration of optical packet switching scheme for header-payload separation and class-based forwarding

M. Nord

Networks Competence Area, Research Center COM, Technical University of Denmark, 345V, 2800 Lyngby, Denmark. and
Internet Network Architecture,Telenor R&D, N-1331 Fornebu, Norway
mn@com.dtu.dk

S. Bjørnstad

Internet Network Architecture,Telenor R&D, N-1331 Fornebu, Norway and Department of physical electronics, Norwegian University of Science and Technology, 7491, Trondheim, Norway

M. L. Nielsen

Networks Competence Area, Research Center COM, Technical University of Denmark, 345V, 2800 Lyngby, Denmark

B. Dagens

Alcatel R&I, Route de Nozay, 91460 Marcoussis,France.

Abstract: We propose and experimentally verify a design realizing a novel scheme for key optical packet switching functionalities: header-payload separation, Class of Service segregation and packet forwarding.

A. Introduction

A key issue in the design of an Optical Packet Switching (OPS) node is how to efficiently combine header reading/insertion techniques with the packet forwarding [85, p7]. Moreover, how to provide Quality of Service (QoS) differentiation in statistically multiplexed networks has recently received much attention. This paper details how a novel scheme may overcome these challenges, using a single Semiconductor Optical Amplifier (SOA) based interferometer.

Serial header, sub-carrier modulation and differentiated phase shift keying are prominent header encoding techniques. Recent designs use couplers to obtain an optical copy of the header, and tunable wavelength converters (TWCs) to erase and insert headers [85, p7]. A recently proposed *orthogonal State Of Polarization* (SOP) header-payload

separation technique [p11] uses a Polarization Beam splitter (PBS) to separate headers from payloads, and a simple Polarization Maintaining (PM) coupler to realize header insertion, since the header is removed in the separation. This was proposed in combination with a scheme for physically segregating different Classes of Service (CoS). This segregation enables the OPS nodes to offer QoS differentiation, thus meeting the QoS requirements of future services with differentiated cost models, and at the same time efficiently utilizing the fibre- and switch bandwidth [100]. Specifically, offering one CoS with no packet loss for Guaranteed Service (GS), and one Best Effort (BE) CoS with a packet loss rate and delay decided by available buffer resources, will enable a wide range of services.

The physical layer packet forwarding can be realized by Array Waveguide Gratings (AWG) based designs [66, 100]. For a switch to be strictly non-blocking in asynchronous operation, and to be compatible with the wavelengths used in the WDM layer, both single- and multi-stage AWG designs require WCs both before and after the AWG, to avoid additional fast space switches [p7].

We propose to realize header-payload separation, CoS segregation and AWG-based packet forwarding, by combining the orthogonal SOP scheme with optical logic functions in a TWC. The All-Active Mach Zehnder Interferometer (AA-MZI) is a suitable TWC choice, as it is compact and has demonstrated high-performance wavelength conversion and optical processing [102]. Combining the orthogonal SOP scheme and a TWC is a good match, since the PBS used for separating the two SOPs gives stable SOPs to the AA-MZI inputs, and thus optimum performance, assuming polarization maintaining (PM) fibre inside the switch. An Automatic Polarization Controller (APC) prior to the PBS is sufficient to compensate SOP variations occurring in the transmission line [p11]. The scheme reduces cross-talk and requirements to SOP alignment, since it avoids simultaneous propagation of two modulated signals at different SOPs [p11]. The design realizing this scheme is illustrated in Fig. 4.1 a). The fact that it uses the same number of high-dynamics active components (tunable lasers and TWCs) as conventional AWG-designs need for solely doing the packet forwarding, is a main benefit. The reported proof-of-principle experiment verifies the physical viability of the design’s required TWC functionalities.

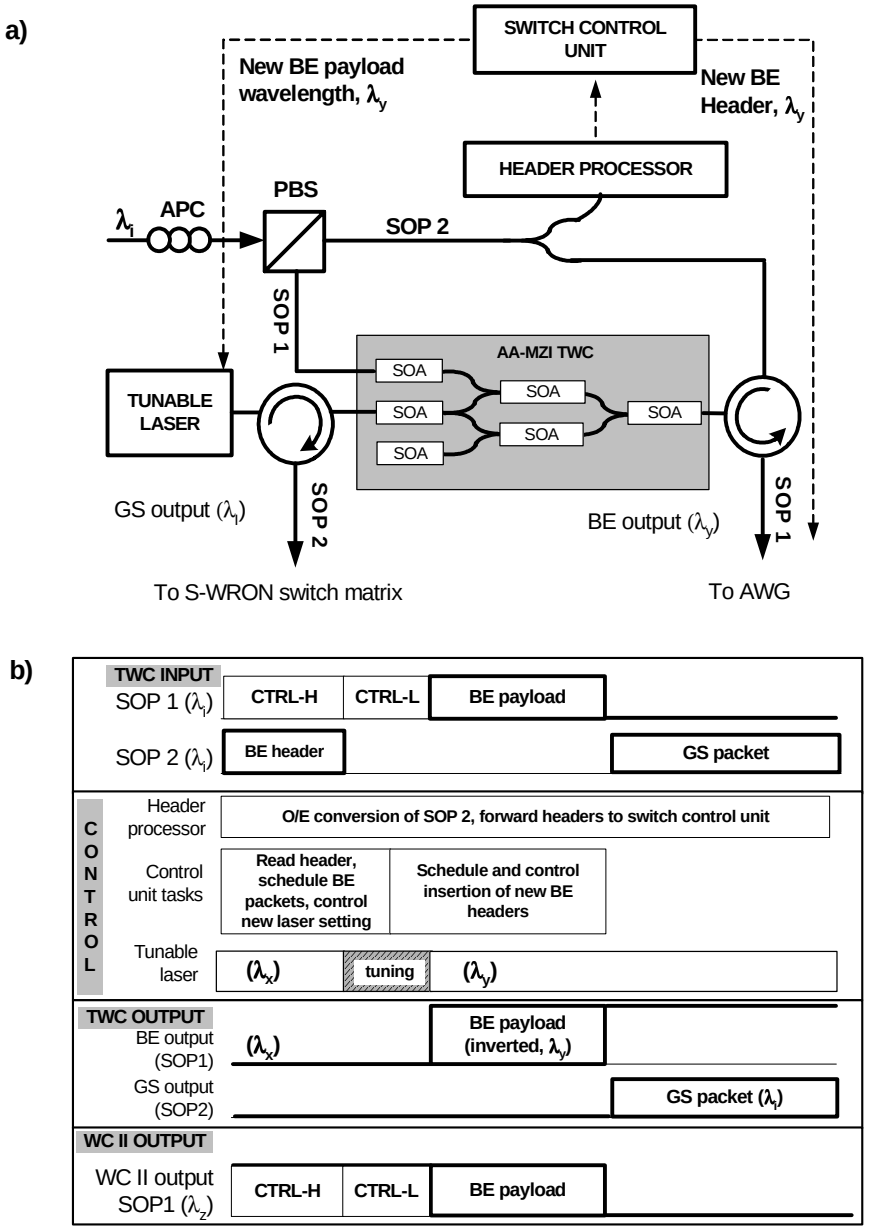

Fig. 4.1. a) The proposed design, combining SOPs and a TWC per channel. b) Detailed scheme functionality.

B. Scheme functionality and physical design

In the considered OPS system, only Best Effort (BE) packets have headers with address information, since GS packets are switched according to their wavelength, following a predetermined path through a Static-Wavelength Routed Optical Network (S-WRON) [100]. The BE payloads, BE headers and GS packets are transmitted on the two SOPs as shown in Fig. 4.1 b). In addition, in-band control signals, used to erase headers and laser output during tuning, termed CTRL-H and CTRL-L respectively, propagate at SOP 1.

At the front end of the switch, the PBS splits the incoming signal into two components, SOP 1 and SOP 2. SOP 1 enters the TWC from the left, containing either a BE packet payload or a CTRL signal. SOP 2 enters the TWC from the right, containing either a GS packet or a BE packet header. A passive splitter provides the header processor with a copy of the signals in SOP 2, where it undergoes O/E conversion. To exclusively forward packet headers to the switch control unit, the header processor can distinguish GS packets and BE headers based on a pre-amble field, e.g. by letting BE headers start with a unique bit-pattern. The TWC will extract the GS packets from SOP 2 to the "GS output", and also erase BE packet headers when a CTRL-H signal is present in SOP 1. The GS packets are then forwarded through a delay [100] towards its output fibre by the fixed switching matrix, defined by the S-WRON configuration, thereby avoiding contention. The switch control unit controls the forwarding of the BE payload in SOP 1 to the correct AWG output port via the tunable laser wavelength. Since the TWC operates out-of-phase, the CTRL-L signals in SOP 1 completes the BE payload extraction, by suppressing the tunable laser output. This is crucial in asynchronous operation, where each channel tuning would generate cross-talk via the AWG. As will be detailed elsewhere, in an analysis of the complete switch, a fixed output wavelength WC after the AWG (WC II) should also operate out-of-phase to reinvert the polarity, as illustrated in Fig. 4.1 b).

Considering the design, an important aspect of this work is that the SOA currents in the AA-MZI are kept constant during operation. This removes the need for fast control of currents to accommodate changing functionalities, and the control is thus *all-optical*. This enables upstream "in-band" control, i.e. the upstream OPS node actually controls the header-payload separation, CoS segregation and suppression of laser output during tuning in the successive node. The scheme has a number of advantages for the switch control unit: The SOP based CoS segregation reduces the workload of the control unit's header processing, since the

SOPs are maintained in the network and since GS packets are automatically switched according to the S-WRON. Furthermore, upstream control enables the control unit to identify a new header and schedule its insertion, as well as scheduling of CTRL-H and CTRL-L signals, *after* it has scheduled the BE payload. Hence, processing FDL length is reduced to the amount needed for scheduling of BE packets, if the sum of laser tuning time, connection fibre- and AWG propagation delay is sufficient for the remaining scheduling. Finally, the very fast response times of the TWC enables reduction, or even elimination, of guard times, thereby reducing packet overhead.

For a comparison of the required optical hardware, the cost of the the PBS and the APC may very well be counter-balanced by the scheme's more advanced functionality. In addition, the APC may not constitute an additional cost, as it may anyhow be required by other polarization sensitive components in the switch. Finally, using a TWC to perform header erasure eliminates dedicated optical gates for this task, and suppressing the laser during tuning removes the need for equipping each tunable laser with fast gates.

C. Experiment and results

Since multiple separations and reinsertions of signals based on their relative SOP has already been demonstrated [p11], we here focus on verifying the functionality of the TWC. The experimental set-up is depicted in Fig. 4.2 a). The tunable lasers are externally modulated, driven by 10 Gbit/s pulse pattern generators (PPG), operating with NRZ PRBS sequences of length 2^{15}-1, or by patterns emulating packets and CTRL signals. Bit Error Rates (BER) are measured by test-sets (BERT). The co-propagating wavelength conversion of SOP 1 requires spectral filtering, and the tunable Band Pass Filter (BPF) used here requires packet-by-packet tuning. However, the scheme can be implemented by a less complex solution, namely use a fixed narrow reflection filter, centered at $\lambda_{i,}$ preceeded by an isolator to remove reflections into the AA-MZI.

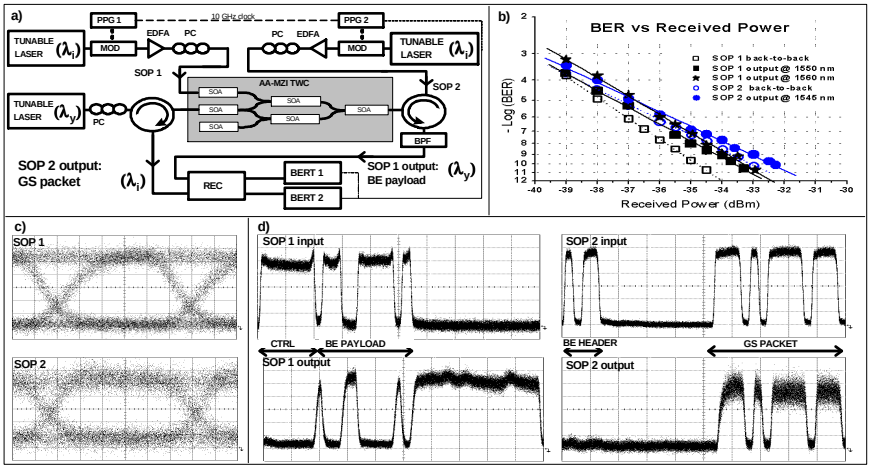

Fig. 4.2. a) Experimental set-up, b) BER vs. power curves, c) continuous mode eye diagrams, and d) emulated optical packets.

The signal quality is expressed by the BER curves of Fig. 4.2 b), revealing power penalties at BER of 10^{-9} below 2 dB, compared with back-to-back. This suggests, together with the eye diagrams of Fig. 4.2 c), that regeneration will be beneficial when traversing many network nodes. We propose to incorporate this in the WC needed after the AWG (WC II), similar to [85]. Finally, Fig. 4.2 d) shows emulated packets and control signals, demonstrating successful extraction of BE and GS packets, polarity inverting wavelength conversion of BE payloads, as well as suppression of CTRL fields and of laser output when tuning should take place. The fast response makes guard times with respect to the processing in the TWC redundant.

D. Conclusion

We detailed the novel scheme for header-payload separation, CoS segregation, packet forwarding and header insertion, and successfully demonstrated the TWC functionalities of the proposed design. Since the advanced functions can be realized with a minimum of additional components, compared to conventional AWG-based designs, the scheme is a promising path towards OPS node design with QoS differentiation.

5. QoS Differentiation

5.1. Introduction

The optical layer will benefit from supporting several QoS levels, as opposed to a single BE CoS level. As argued in Chapter 5.2, Section B, a QoS differentiated network tailored to the supported applications will not only provide sufficient QoS for the most demanding services, but also result in improved utilisation of network resources, which is economically beneficial.

Whilst OPS evaluation typically has focused on PLR, this thesis also addresses the detrimental effect of jitter. This issue is gaining increasing interest, and has recently also been addressed in [103]. This Chapter discusses QoS differentiation, as detailed in the following:

- Chapter 5.2 incorporates an article accepted for publication in IEEE Communication Magazine [p25], which discusses main types of QoS differentiation methods for bufferless OPS: Intentional Packet Dropping (IPD), Wavelength Allocation (WA) and Pre-emptive Drop Policy (PDP). The study concludes that PDP has best performance in terms of throughput, but that it also has the highest complexity. AR represents a trade-off in terms of performance and complexity. Note that WA is a special case of Access Restriction (AR), considered below.
- Chapter 5.3 incorporates a BroadNets 2004 conference paper [p20], which evaluates the AR based QoS algorithm, proposed in [p19], to improve the efficiency of the AR method, for a SPN TWC pool.
- Chapter 5.4 incorporates an OSA Journal of Optical Networking 2004 article [p22], which further develops the AR method. This study aims to benefit from a result shown in Chapter 3.6: For a port-constrained SPN pool, lowest BE PLR is obtained for a mix of TWCs and FDLs. However, using FDLs raises the question of whether all traffic can support the induced jitter. The article thus studies performance as a function of increased CoS granularity, with different PLR and jitter requirements.

5.2. QoS Differentiation Methods

This chapter incorporates [p25], accepted for publication in IEEE Communication Magazine in 2005.

Evaluation of QoS Differentiation Mechanisms in Asynchronous Bufferless Optical Packet Switched Networks

Harald Øverby [1], Martin Nord [2,3], Norvald Stol [4].

(1,4) Department of Telematics, Norwegian University of Science and Technology, O.S. Bragstadsplass 2E, N-7491 Trondheim, Norway
(2) Research Centre COM, Technical University of Denmark, DK-2800 Lyngby, Denmark
(3) Telenor R&D, N-1331 Fornebu, Norway
[1]Tel: +47 73 59 43 50, Fax: +47 73 59 69 73, Email: haraldov@item.ntnu.no
[4]Tel: +47 73 59 21 33, Fax: +47 73 59 69 73, Email: norvald@item.ntnu.no

Abstract. Existing Quality of Service (QoS) differentiation schemes for today's IP over point-to-point Optical Wavelength Division Multiplexed (WDM) networks take advantage of electronic Random Access Memory (RAM) to implement Active Queue Management (AQM) algorithms in order to isolate the service classes. Since practical optical RAM is not available, these techniques are not suitable for a future all-optical network. Hence, new schemes are needed to support QoS differentiation in Optical Packet Switched networks (OPS). In this article we first present an overview over existing QoS differentiation mechanisms suitable for asynchronous bufferless OPS. We then compare the performance of the presented schemes, as well as qualitatively discussing implementation issues, in order to evaluate the mechanisms. In particular, we present an evaluation framework, which quantifies the throughput reduction observed when migrating from a best-effort scenario to a service-differentiated scenario. Our study shows that pre-emption based schemes have best performance, but also the highest implementation complexity.

A. Introduction

During the last decade we have experienced an explosive growth of the Internet traffic in the core networks. This growth is mainly caused by the increasing number of Internet users, combined with increased access network capacity. For instance, the International Telecommunication Union (ITU) reported a 14.7 % growth in the number of Internet users

from 2002 to 2003 [104], while the number of DSL subscribers increased from 36 millions to 64 millions in the same period [105]. Future traffic growth must be accompanied by a corresponding growth in the core networks, in order to avoid a capacity bottleneck. The most promising transmission technology to satisfy such a demand is Wavelength Division Multiplexing (WDM), which enables capacities of several Tbps in a single optical fibre [4].

Today, WDM is utilized in a point-to-point architecture, which means that all channels on each fibre are terminated in all network nodes, for switching in the electronic domain. Hence, the signals undergo optical-electrical (O/E) and electrical-optical (E/O) conversions when entering end leaving the switch, respectively. This approach has several drawbacks, such as high cost, due to extensive use of expensive O/E and E/O converters, and lack of data transparency [4]. Moreover, electronic technology faces technological limits when it comes to handling line speeds above 40 Gbps, which will result in complex and expensive switch constructions to accommodate the expected increase in traffic [25]. Hence, today's point-to-point WDM networks are predicted to evolve into all-optical network architectures that avoid O/E/O conversions of data, such as Wavelength Routed networks (WR) [4], Optical Burst Switched networks (OBS) [33, 106] and Optical Packet Switched networks (OPS) [25]. In WR, end-to-end lightpaths are set-up between communicating end nodes in the core network. In OPS and OBS, packets/bursts are routed hop-by-hop in the optical domain from an ingress node to an egress node in the core network. Since OPS and OBS benefit from statistical sharing of link resources, they show better utilization of network resources than WR, and are thus promising candidates for the future core network. However, a commercial deployment of OPS and OBS requires advances in several key enabling technologies, such as optical wavelength conversion [107] and scalable switch matrices with fast switching times [4].

In order to enable a successful deployment of real-time and mission critical applications, Quality of Service (QoS) differentiation should be provided in future OPS [108]. That is, the current best-effort service may not offer adequate QoS to the most demanding applications such as real-time video, interactive gaming and tele-medicine. Furthermore, although some applications need better QoS than can be provided by the best-effort service, other applications (e.g. e-mail and file-transfer applications) are satisfied using the best-effort service [108]. Hence, a QoS differentiated network tailored to the different applications will not only provide sufficient QoS to demanding services, but also result in an improved utilization of network resources, which is demanded by economics [18].

Existing QoS schemes for point-to-point WDM networks (with electronic switches) are not suitable for OPS. This is because the existing QoS schemes rely on Active Queue Management (AQM) algorithms to differentiate between the service classes, e.g. by giving high priority traffic strict priority over low priority traffic, which means that low priority traffic is buffered until all high priority traffic has been processed and the link is idle [110]. In electronic switches this is feasible due to the existence of low-cost, high capacity electronic Random Access Memory (RAM). However, buffering in the optical domain is today only available through Fibre Delay Lines (FDLs), where packets are delayed by being transmitted on a fixed length optical fibre. Using FDLs as a replacement for electronic RAM to perform AQM is not feasible due to several reasons:

Packets delayed in an FDL can only be retrieved when it leaves the FDL after a predefined time, i.e. random memory access is not possible with FDL. Hence, complex processing is needed in order to perform AQM on FDLs. FDLs can only delay packets for a limited amount of time, depending on the length of the FDL and on the allowed number of re-circulations, since each circulation detoriates signal quality [99]. This, combined with the bulkiness and cost introduced by FDLs, limits the capacity of an FDL buffer. Hence, new QoS schemes, which take advantage of intrinsic properties of the WDM layer is required, i.e. the service classes must be differentiated without the use of electronic RAM.

In recent research, many proposals for providing QoS differentiation in OPS have been made [54, 111-114, p19]. When considering these mechanisms, it is important to clearly distinguish synchronous and asynchronous OPS, because a given QoS differentiation mechanism is in general not suitable for both architectures. In synchronous OPS, fixed-sized packets arrive at a core switch in synchronized time-slots, where complex synchronisers compensate for delay variations occurring between packets. In asynchronous OPS, packets can arrive at a core switch at any instant, and there is no need for synchronization between the input ports, thereby avoiding complex optical synchronization technology.

In this article we present an overview of existing QoS differentiation mechanisms, suitable for asynchronous bufferless OPS. It is our aim to show the differences in performance and complexity of the various QoS differentiation mechanisms. In particular, we introduce a quantitative framework for measuring the throughput penalty experienced when introducing QoS differentiation in asynchronous OPS.

B. Contention resolution in OPS

Before we move on to discussing the various QoS differentiation mechanisms, we briefly addresses how contention can be resolved in OPS, since the choice of contention resolution architecture highly influences the mode of operation of the QoS differentiation schemes. In asynchronous OPS, contention occurs when a packet is destined for a wavelength that is currently occupied transmitting another packet. The arriving packet will be dropped unless some contention resolution mechanism is utilized. The contention resolution mechanisms proposed in recent literature can be grouped into three domains [115]:

- **Wavelength domain:** Contending packets are converted to idle wavelengths on the same fibre using wavelength converters. This technique does not cause additional delay, nor reordering of the packets.
- **Time domain:** Contending packets are delayed and scheduled for transmission a later point in time when the wavelength is (hopefully) available. This technique results in an additional delay and may result in reordering of packets. It is important to note that using the time domain for contention resolution in this manner is fundamentally different from buffering using electronic RAM today's store-and-forward networks: In the latter, all packets are buffered, although resources are available, while in the former, only contending packets are buffered.
- **Space domain:** Contending packets are transmitted on the same wavelength on another idle output fibre, which leads to another node than originally intended. Hence, the packets may follow non-optimal paths toward its destination. This technique potentially results in a large additional delay, which increases both the probability- and magnitude of packet reordering.

As shown in [87], both the wavelength- and time domain can effectively reduce the packet loss ratio (PLR) in the case of contention. However, since utilizing the time domain implies the use of FDLs or electronic RAM [114], the switch complexity and hardware cost increases. Also, due to the added delay from the buffers, packets may experience an increased end-to-end delay and potential reordering of packets within a stream, which is unfavourable for e.g. high quality streaming services and TCP connections [p22]. Regarding the space domain, the authors of [115] show that only a limited reduction in the PLR can be achieved. Hence, due to the drawbacks and limitations associated with contention resolution in both the time- and space domain, we focus on an optical

packet switch architecture that exclusively utilizes the wavelength domain for contention resolution (see Section C.1 for details).

In such a bufferless OPS architecture, the most significant QoS parameter is the PLR [p19, p22, 54, 87, 97, 111-114] that results from network layer contention. Since there are no contention resolution buffers at intermediate nodes, the end-to-end delay is governed by the propagation delay, and possibly the packet assembly delay in the ingress router, and will not be addressed further in this article.

C. QoS differentiation in asynchronous bufferless OPS networks

QoS can be provided based on a per-flow-, or on a per-class classification of the traffic, which is analogue to the IETF IntServ and DiffServ approaches, respectively [116]. In this article we focus on per-class classification of the traffic, due to the scalability problem associated with per-flow classification of the traffic in large networks.

For the per-class architecture, the QoS guarantees can be expressed as relative to other service classes (relative guarantees) or within absolute bounds (absolute guarantees). With relative guarantees, QoS parameters of a certain service class are given relative to another service class, e.g. PLR for low priority traffic / PLR for high priority traffic = 10^2. With absolute guarantees, QoS parameters of a certain service class are given upper bounds, e.g. PLR for high priority traffic < 10^{-4}. This article focuses on relative guarantees, but it should be noted that the presented QoS differentiation schemes may be extended to provide absolute guarantees as well, as shown in [111] for the Preemptive Drop Policy (PDP).

In order to isolate the PLR between the service classes in asynchronous bufferless OPS by utilizing the WDM layer, three different mechanisms may be used: Access Restriction, Preemption, and Packet dropping. Further on we describe these mechanisms in detail, and present QoS differentiation schemes utilizing these mechanisms.

C.1. System model

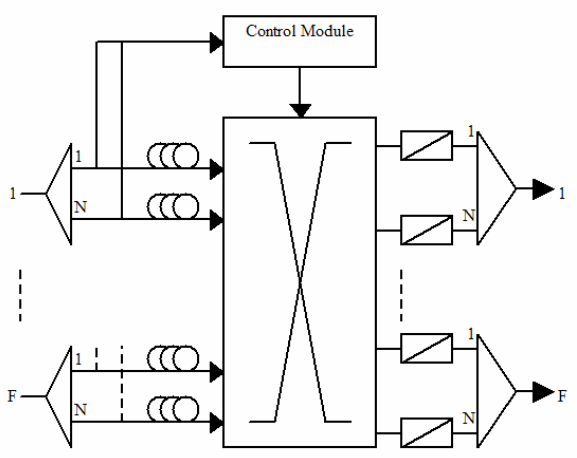

Fig. 5.1. Generic optical packet switch.

We consider a generic optical packet core switch with F input/output fibres and N wavelengths per fibre, as illustrated in Fig. 5.1. The switch has a full-range wavelength converter placed at each output wavelength. As will be become evident in the next sections, we consider two scenarios:

Best-effort scenario: All packets belong to the same service class and are treated equally, which results in all packets having the same PLR.

Service differentiated scenario: There are two service classes in the network, where service class 0 is given priority over service class 1. The service classes are isolated by using one of the considered QoS differentiation schemes presented in Section C.2 - C.4.

We assume a uniform traffic pattern, which means that we can restrict our study to consider a single output fibre (which may be any output fibre), denoted as the tagged output fibre. When a packet arrives to the optical packet switch, the packet header is extracted and converted to the electronic domain for processing, while the packet payload is delayed using input FDLs, as illustrated in Fig. 5.1. Regarding the tagged output fibre, let the term P_{be} denote the PLR in the best-effort scenario, and the terms P_0 and P_1 denote the PLR for service class 0 and 1 in the service differentiated scenario, respectively. Also, let S_0 and S_1 denote the relative share of class 0 and class 1 traffic, respectively. The throughput is defined as G_{be}=1-P, for the best-effort scenario, and G_{sd}=1-(S_0P_0+S_1P_1), for the service differentiated scenario. Finally, denote the class isolation as I=P_1/P_0, to quantify the relative PLR difference between the service classes. Table 5.1 summarizes the parameters used.

The numerical results presented in the next sections have been obtained using the analytical models presented in [111, 112, 117], which are based on discrete-time Markov chains.

Table 5.1.

Parameter	Description
F	Number of input/output fibres in the switch.
N	Number of wavelengths per fibre.
A	Normalized system load.
S_0	Relative share of class 0 traffic.
P_{be}	Packet loss rate in the best-effort scenario.
G_{be}	Throughput in the best-effort scenario.
P_0	Packet loss rate for class 0 traffic in the service differentiated scenario.
P_1	Packet loss rate for class 1 traffic in the service differentiated scenario.
G_{sd}	Throughput in the service differentiated scenario.
n	Number of wavelengths reserved to class 0 traffic (WA).
p	Probability of successful preemption (PDP).
d	Probability of dropping a class 1 packets (IPD).

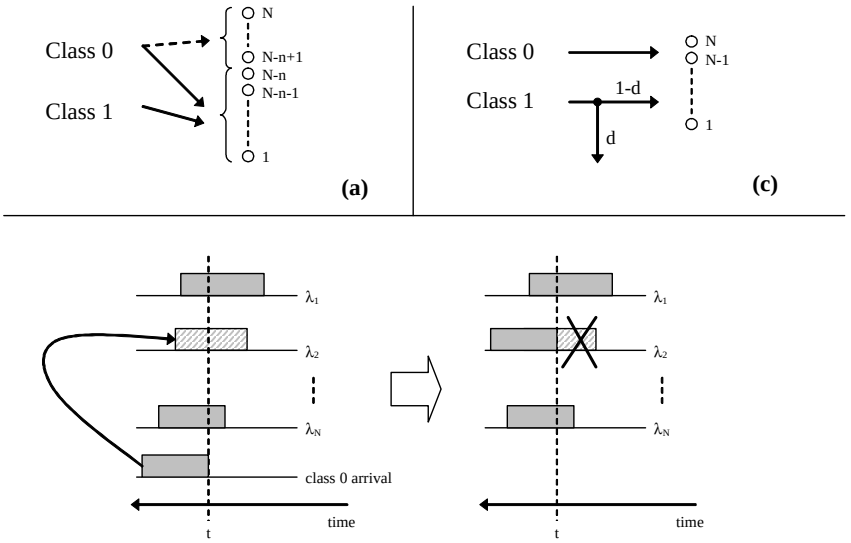

Figure 5.2. The mode of operation for the WA (a), the PDP (b) and the IPD (c). In (a), both service classes have access to wavelengths 1 to *N-n*, while only class 0 packets have access to wavelengths *N-n*+1 to *N*. In (b), a class 0 packet preempts a class 1 packet currently in transmission on wavelength λ_2. In (c), a proportion *d* of class 1 traffic is dropped before reaching the tagged output fibre.

C.2. QoS differentiation schemes based on access restriction: The Wavelength Allocation algorithm (WA)

With access restriction, a subset of the available resources (may be wavelengths, wavelength converters, buffering space, etc.) is exclusively reserved for high priority traffic. This means that low priority traffic has fewer resources available than high priority traffic, which results in a lower PLR for high priority traffic, compared to low priority traffic.

An example of a QoS differentiation scheme based on access restriction is the Wavelength Allocation algorithm (WA). Here, $n<N$ wavelengths at the tagged output fibre are exclusively reserved for class 0 traffic [112], as illustrated in Fig. 5.2 a). That is, as long as less than N-n wavelengths at the tagged output fibre are occupied, both new class 0 and class 1 arrivals are accepted. In the opposite case, only class 0 arrivals are accepted, whilst class 1 arrivals are dropped, resulting in a lower PLR for service class 0 than for service class 1. The class isolation (I) may be controlled by adjusting the number of wavelengths reserved for service class 0 (n). Fig. 5.3 shows how the PLR for class 0 traffic decreases, and the PLR for class 1 traffic increases, with increasing n.

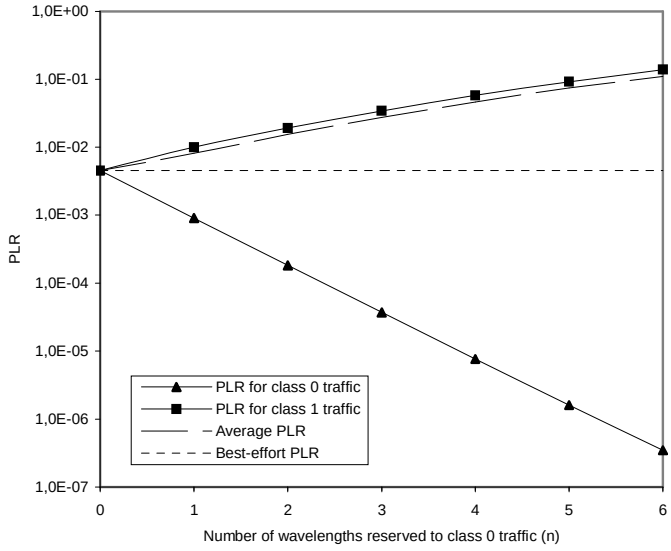

Fig. 5.3. The PLR as a function of the number of wavelengths reserved for class 0 traffic (n) for the WA. N=16, A=0.5, S_0=0.2.

C.3. QoS differentiation schemes based on preemption: The Preemptive Drop Policy (PDP)

With pre-emptive techniques, all free resources are available to all traffic. However, when all resources are taken, a high priority packet may take over (preempt) a resource currently occupied by a low priority packet, which is then (at least partially) lost. On the other hand, a low priority packet cannot preempt any other packet. Hence, on average less resources are available to low priority packets than to high priority packets, resulting in a lower PLR for high priority traffic.

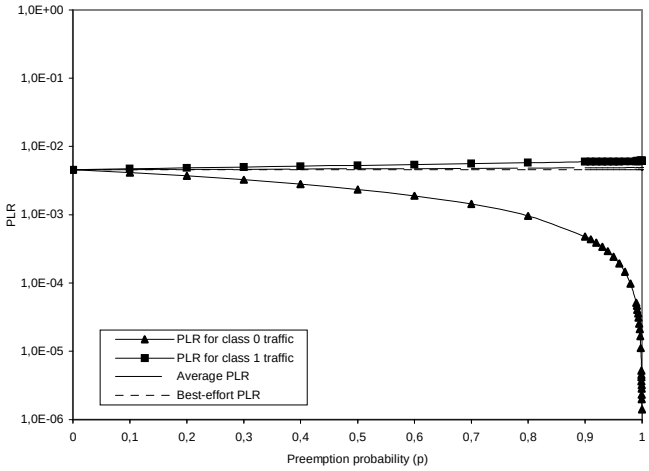

Fig. 5.4. The PLR as a function of the preemption probability (*p*) for the PDP. *N*=16, *A*=0.5, S_0=0.2.

In the Preemptive Drop Policy (PDP) [111], a class 0 packet may preempt a class 1 packet currently occupying a wavelength, when all N wavelengths at the tagged output fibre are occupied, as illustrated in Fig. 5.2 b). This means that a class 1 packet is lost instead of a class 0 packet, which intuitively results in a lower PLR for class 0 traffic relative to class 1 traffic. If there are only high priority packets occupying the wavelengths, preemption is not possible, and the arriving class 0 packet is lost. The design parameter p denotes the probability of preemption, and can be used to control the class isolation. That is, when all wavelengths at the tagged output fibre are occupied, and a class 0 packet arrives, there is a probability p that preemption takes place given that there are class 1 packets currently in transmission. Hence, with p=0 one expects the PLR for class 0 and class 1 traffic to be equal, while the maximum class isolation is obtained for p=1. In the latter case, class 0 traffic is lost only when a class 0 arrival finds all output wavelength occupied transmitting class 0 packets. Fig. 5.4 shows the PLR as a function of the preemption

probability (p). We confirm that the PLR of both class 0 and class 1 are equal for p=0, before decreasing and increasing, respectively, with increasing p.

C.4. QoS differentiation schemes based on packet dropping: Intentional Packet Dropping (IPD)

With packet dropping, low priority traffic is dropped with a certain probability before attempting to seize a resource. This results in an increased PLR for low priority traffic, but also a decreased PLR for high priority traffic, since the total system load on the resource decreases.

A packet-dropping scheme for OBS, Intentional Packet Dropping (IPD), has been proposed in [113], but this scheme may well be used for OPS as well. Here, class 1 packet arrivals are dropped with a probability d before reaching the tagged output fibre, as illustrated by Fig. 5.2 c). This has two effects: First, the PLR for class 1 traffic increases, since packets are dropped with a probability d (in fact, $P_1 \geq d$). Second, the PLR for class 0 traffic decreases since the system load on the tagged fibre decreases compared to the best-effort scenario. Hence, the parameter d may be used to control the class isolation between, e.g. for d=0 we expect an equal PLR for the service classes, while for d=1 we expect the maximum class isolation. In particular, for the latter case, we have that $P_1=d=1$.

Fig. 5.5 shows the PLR as a function of the dropping probability (d). We see that the PLR for class 0 traffic decreases and the PLR for class 1 traffic increases as the parameter d increases.

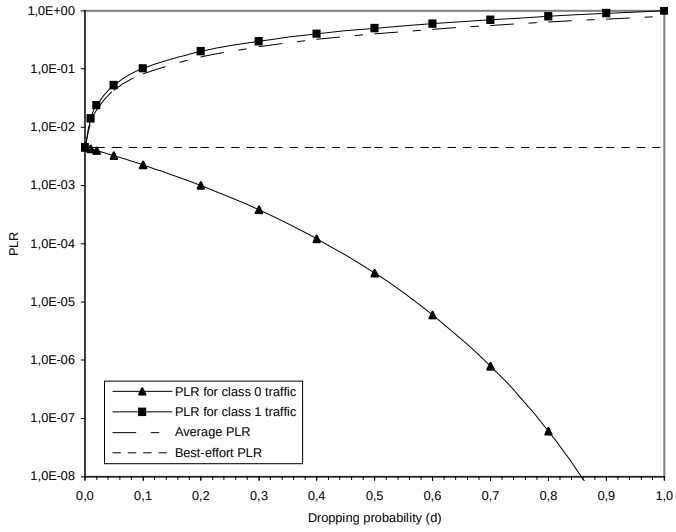

Fig. 5.5. The PLR as a function of the dropping probability (d) for the IPD. N=16, A=0.5, S_0=0.2

D. Comparison study of QoS mechanisms

A crucial issue regarding deployment of QoS differentiation in asynchronous OPS is the associated decrease in the throughput. This is due to the non-optimal utilization of resources resulting from utilizing the WDM layer to differentiate between the service classes, as studied e.g. in [111, p22]. Clearly, this drawback should be minimised. In this section we therefore present a general evaluation framework in order to quantify this effect (Section D.1), as well as a comparison study of the above-presented QoS differentiation schemes (Section D.2).

D.1. Comparison framework

We consider the case where the network migrates from the best-effort scenario to a service differentiated scenario with two service classes. This means that the throughput changes from G_{be} to G_{sd}. Denote S as the relative decrease in throughput when introducing QoS differentiation:

$$S = \frac{G_{sd}}{G_{be}} = \frac{1-\left(S_0 P_0 + S_1 P_1\right)}{1-P_{be}} \tag{5.1}$$

For instance, S=0.80 means that the throughput is 80 % of the throughput prior to the introduction of QoS differentiation. Hence, the ideal situation is to have S=1.0, which means that employing QoS differentiation does not influence the throughput. In this case, there is a one-to-one mapping between lost class 0 and class 1 packets, i.e. each class 1 packet that is lost due to QoS differentiation actually prevents a class 0 packet from being lost. However, as we will see in the next section, in practice we often find situations where S<1, which means that there is a non-optimal utilization of the network resources in the service differentiated scenario. The proposed comparison framework is general and may be applied to evaluate the throughput penalty of QoS differentiation schemes suitable for asynchronous OPS with buffering as well as synchronous OPS.

D.2. Numerical evaluation

We evaluated the performance of the WA, PDP and the IPD, using the proposed evaluation framework. Fig. 5.6 shows their throughput penalty S as a function of the isolation degree, I, for system loads of 0.2, 0.5 and 0.8. First, we observe that $S_{PDP} \geq S_{WA} \geq S_{IPD}$ for all considered scenarios, which indicates that the PDP has the least reduction in the throughput, followed by the WA and the IPD. In particular, we see that $S_{PDP} \approx 1.0$ for all considered scenarios, which indicates that the use of the PDP does not reduce the throughput significantly. However, for the WA and IPD, we see that the value of S is well below 1.0 when the system load is 0.5 and 0.8, and that IPD suffers the most.

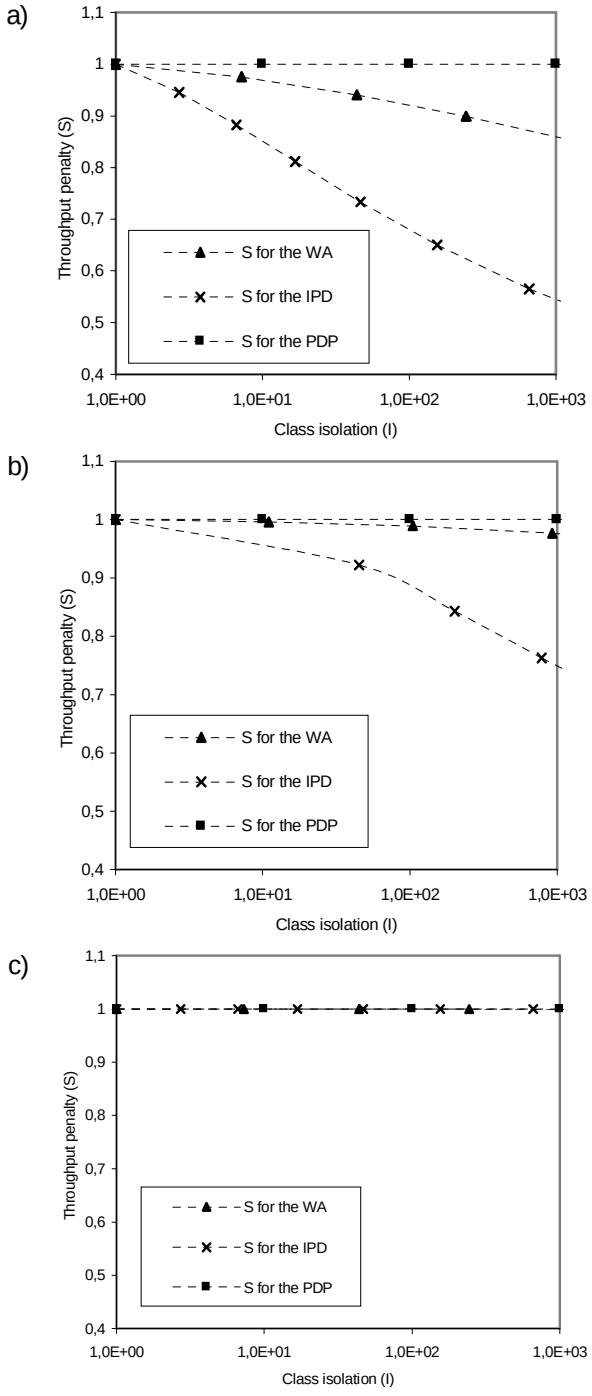

Fig. 5.6. Throughput penalty (*S*) as a function of the class isolation (*I*) when N=16, S_0=0.2, for: a) A=0.8, b) A=0.5, and c) A=0.2.

The reason for the observed differences between the schemes is that in the WA and IPD, packets are dropped although wavelengths are idle, while in the PDP all wavelengths are shared amongst all arrivals. Moreover, since the WA drops packets only when the tagged output fibre is highly strained, it shows better performance than the IPD, which drops packets independently of the state of the tagged output fibre. On the other hand, for a low system load of 0.2, there is no significant throughput reduction in neither case: $S_{WA} \approx S_{IPD} \approx S_{PDP} \approx 1.0$. Hence, for sufficiently low system loads, the initial PLR is so low that the desired isolation, although achieved for a relatively large change in the PLR of both service classes, does not significantly reduce the throughput.

E. Implementation issues

To complete the evaluation of the QoS differentiation mechanisms, we qualitatively discuss their implementation complexity. We make a clear difference between hardware complexity and scheduling complexity. Increased hardware complexity stems from additional hardware resources needed to manipulate optical packets in order to realize the QoS differentiation scheme, whilst increased scheduling complexity results from additional electronic processing associated with implementing the QoS differentiation scheme.

Note that best-effort schemes requires a switch matrix, as discussed in Section C.1, and a scheduler which tracks the state of all output wavelengths, including the remaining duration of allocated packets.

Regarding the IPD, no additional hardware is required, as class 1 packets are randomly dropped before reaching the output fibre. When it comes to scheduling, the IPD has the same complexity as in the best-effort scenario, since no additional state information about the output ports is needed.

Regarding the WA, no additional hardware is required. However, when it comes to scheduling, the switch must compare the number of occupied wavelengths at each output fibre with n, to be able to drop class 1 packets when there are N-n or more wavelengths occupied.

For the PDP, the output wavelength state information must also include the service class of the packet, to be able to pre-empt only class 1 packets. An improvement of the PDP is achieved by preempting the latest class 1 arrival, to minimise the “wasted bandwidth”. This requires including information about when the currently switched packets arrived. Regarding hardware complexity, additional hardware is required to erase

the part of the preemptied packet that has already been transmitted, to minimise the bandwidth utilization in downstream nodes.

Table 5.2.

	Scheduling complexity	Hardware complexity	Performance
IPD	Low	Low	Poor
WA	Medium	Low	Medium
PDP	Medium-High	Low-Medium	Good

F. Conclusions

This article has provided an overview of existing QoS mechanisms for asynchronous bufferless OPS. These schemes are fundamentally different from the schemes utilised for store-and-forward networks, since electronic RAM is not feasible to implement in the optical domain. We have evaluated the overall reduction in the throughput as QoS differentiation is introduced in asynchronous bufferless OPS. Based on the proposed evaluation framework, we have shown that the PDP has the best performance, followed by the WA and the IPD. This difference is more accentuated when the switch is highly strained, which arguably is also the scenarios in which QoS differentiation is needed the most. However, regarding implementation complexity, the PDP is the most complex followed by the WA and the IPD. These findings are summarized in Table 5.2.

ACKNOWLEDGEMENTS

The authors wish to thank Telenor R&D for supporting this work, and Andreas Kimsaas at the Department of Telematics for valuable advice.

5.3. Access Restriction in TWC SPN Pools

This chapter incorporates an IEEE BroadNets 2004 conference paper [p20] (the proceedings contains a shorter poster version).

Performance analysis of a low-complexity and efficient QoS differentiation algorithm for bufferless optical packet switches with shared wavelength converters in asynchronous operation

Martin Nord (mn@com.dtu.dk)

Research Centre COM, Technical University of Denmark, B345V, Lyngby, Denmark & Telenor R&D, 1331 Fornebu, Norway.

Abstract. This paper discusses the influence of node adjacency, fibre wavelength count, overload situations and potential improvements to a low-complexity Quality of Service differentiation scheme with high efficiency, in terms of overall packet loss to obtain a given service-class isolation, suitable for asynchronous operation.

A. Introduction

A.1. Class of Service (CoS) Requirements

It is not possible to foresee the requirements of future applications and services. However, it is unlikely that all future traffic requires very low PLRs. A widespread belief is that the IP protocol will further gain in popularity to constitute a network layer supporting a wide range of applications. Some applications, such as transport of Voice over IP (VoIP), may be satisfied with a PLR of around 1-10 % [p4]. For file transfers, satisfactory TCP performance imposes a limit to the maximum PLR for a given end-to-end delay, since the product of packet loss rate and the square of the throughput-delay product should be less than one, where the throughput is measured in packets/second on a per TCP connection basis [118]. This is valid for random loss, but has been proposed as a design rule to find order of magnitudes for acceptable PLRs of TCP transfers over OPS in a US scenario, with an end-to-end network delay of 50 ms [119]. Assuming a 1 Mbit/s TCP connection using 1500 Bytes packets results in a maximum network PLR of 5.4 %. This corresponds to a 6-node network with a PLR of 1 % each, assuming constant PLRs at each node. For half the TCP connection throughput and

the same delay, a 5 % PLR at each packet switch in a 5-node network is then acceptable. Other applications may require network PLRs in the 10^{-4} and 10^{-6} range [p4], like MPEG-2 video codec streaming.

To illustrate the proposed QoS differentiation scheme, the client traffic of the OPS transport network is here divided into two CoS. Since delay in the switches is negligible to transmission delay in our bufferless network, the CoS are differentiated solely on the PLRs [114, 120, 121], with CoS1 and CoS2 calling for PLRs in the range of 10^{-5} and 10^{-2}, respectively. The fraction of traffic in each CoS is equal on average, which puts quite hard demands on the network, since generally, overall performance decreases with the fraction of high priority traffic [114].

A.2. QoS differentiation in OPS and OBS

Future IP networks may support QoS differentiation in a scalable way, e.g. by implementing the relative-CoS-priority IETF DiffServ approach, based on per-hop behaviour [57, 108]. The asynchronous and variable packet length nature of the Internet, makes an asynchronous, VLP OPS core transport network with QoS differentiation support a good server layer candidate [54].

An approach to QoS differentiation could be to apply access restriction at the ingress nodes, where having packets in electrical form enables quite advanced functionality. However, this approach would require an accurate view of the network state, to avoid unnecessary packet discards at the ingress. This is not readily available in networks with a high traffic pattern dynamics.

A different approach, similar to the DiffServ approach, is to map the packets at the ingress onto data units with a particular CoS encoded in the control information. The packets are then not treated differently (thus not unnecessarily discarded) until they actually encounter a potential congestion situation in one of the core optical packet switches. QoS differentiating by Wavelength Allocation (WA) schemes, also termed Access Restriction (AR) schemes, have been applied to FDL buffer wavelengths in asynchronous- [54] and in slotted [89] operation, to electronic buffers inputs in asynchronous operation [114], as well as to TWC access in a bufferless approach with asynchronous operation [112] and full wavelength conversion capability.

A.3. Outline and rationale for this study

The work in this paper is a part of a study on scalability constraints and QoS differentiation in asynchronous VLP OPS networks. The study aims at obtaining a high functionality/cost ratio, by making hardware savings to limit complexity and by keeping a low scheduling complexity. Similar

to [95], contention is resolved exclusively in the wavelength dimension. Using Shared Per Node (SPN) TWCs, a QoS differentiation algorithm, restricting access to both TWCs and output interfaces (wavelength channels) was proposed to improve efficiency in terms of the ratio of overall loss and obtained isolation [p19], compared to single-resource algorithms. This algorithm enables a more flexible QoS-aware dimensioning in the Shared Per Waveband Plane (SPWP) design, resulting in potential overall hardware savings [p18].

This study extends our former works by investigating three performance aspects of the QoS algorithm's performance:

- Potential improvements to the algorithm.
- Impact of node adjacency.
- Impact of temporarily (low-dynamic) overloads situations.

In this work we consider a single node, as in [p19] but this could also be a Waveband Plane, as in [p15], assuming that there are several independent parallel planes. Different from most existing work, but similar to [114], the performance of QoS differentiation algorithms is investigated for systems with rather high wavelength count.

B. Quality of Service Differentiation by Access Restriction

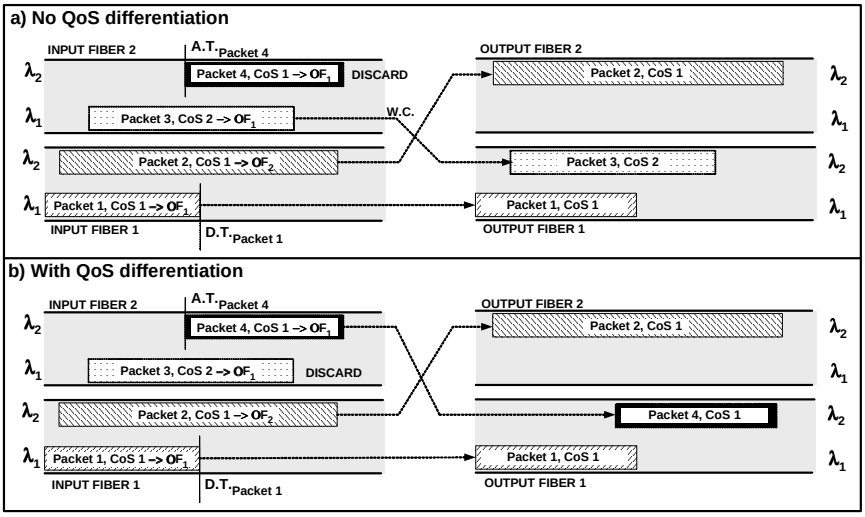

Fig. 5.7. Case of *F*=2, *W*=2, with traffic with contention that *cannot* be resolved in the wavelength domain. QoS differentiation saves a CoS1 packet on the expense of a CoS2 packet.

The principle of QoS differentiation by differentiating the access to network resources is illustrated through an example in Fig. 5.7, showing incoming traffic with contention, meaning that more than one packet contend for the same output wavelength on the same output fibre. In this example, Packet 1 and Packet 2 are being switched from Input Fibre 1, to Output Fibre 1 and to Output Fibre 2, respectively. When Packet 3 arrives on Input Fibre 2, its own wavelength, λ_1, is currently occupied on the requested output fibre. The contention can be resolved in the wavelength domain, as shown in Fig. 5.7 a), by converting the packet to λ_2, when switching it to Output Fibre 1. However, in this case, a later arriving packet cannot be switched to this output fibre, since all wavelengths then are occupied at its arrival time.

When Packet 4 is a CoS1 packet, it could be given priority over Packet 3, which is a CoS2 packet. Fig. 5.7 b) illustrates an example, where the control unit only allows the CoS2 packets to be allocated if there is more than one free wavelength at its requested output fibre. This increases the probability of successful allocation of future CoS1 packets, and Packet 4 can now be allocated, at the expense of the discarded Packet 3.

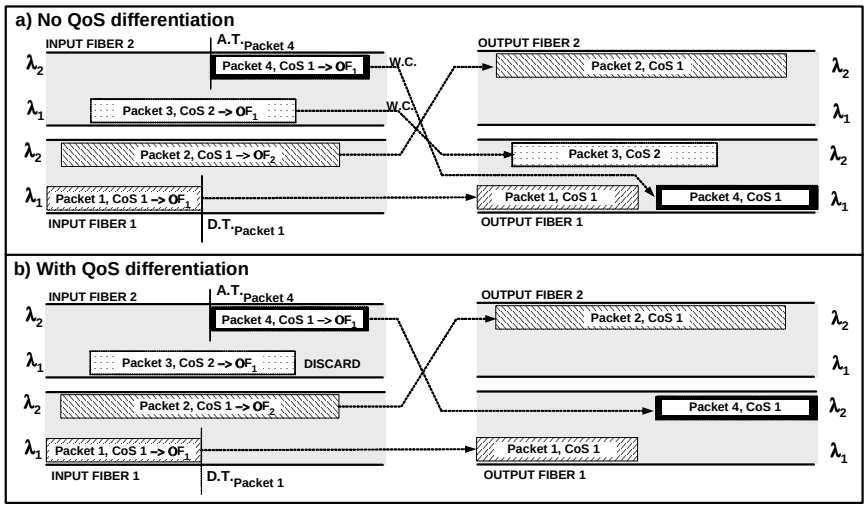

Fig. 5.8. Case of *F*=2, *W*=2, with traffic with contention that *can* be resolved in the wavelength domain. QoS differentiation induces a loss of a CoS2 packet, which does not save a CoS1 packet.

In slotted operation, with fixed packet length and synchronous packet arrival, the control unit knows all the packets to be switched within the time-slot, and the algorithm can hence ensure that CoS2 packets are only discarded, when they prevent a CoS1 packet from being discarded. Moreover, no voids are formed, and the overall PLR in the case of QoS differentiation, $PLR_{WITH\_QoS\_DIFF}$, will remain the same as the best-effort PLR, PLR_{BE}, i.e. overall PLR when not applying QoS differentiation. In asynchronous operation, however, the control unit does not know future packet arrivals when allocating a packet. This could be accomplished by introducing input FDLs to get a "time-window" or "horizon" in which it studies the arrivals, before allocation resources to earlier packets. However, this increases the computational effort of the control unit's allocation algorithm. In this paper, the control unit differentiates access to node resources, based on current allocations only and without studying their duration, as opposed to some offset-based OBS schemes [p4]. This can be a low complexity approach, particularly suited for asynchronous operation with variable length packets, and thereby helping to alleviate the potential OPS/OBS control plane bottleneck [p10, 80].

The potential drawback is an increase in overall PLR, when either no CoS1 packet exploits the vacancy left by a discarded CoS2 packet, or when a certain time passes before the CoS1 packet arrives. This "void" leads to decreased utilisation and will increase the probability of discards of later arriving packets. The situation in which all contention can be

resolved with wavelength conversion is illustrated in Fig. 5.8 a), since the already allocated Packet 1 on Output Fibre 2 departs before Packet 4 arrives, i.e. the Arrival Time of Packet 4 is larger than the Departure Time of Packet 1 ($A.T._{Packet\ 4} > D.T._{Packet\ 1}$). However, applying the same QoS differentiation algorithm as above still triggers the discard of Packet 3, even if Packet 4 does not need this for future allocation, as illustrated in Fig. 5.8 b). Hence, contrary to the situation in Fig. 5.7 b), it does not decrease PLR_{CoS1}, whilst still increasing PLR_{CoS2}.

QoS differentiation through Access Restriction (AR) thus obtains the desired effect of increasing the ratio of PLR_{CoS2} and PLR_{CoS1}, but it comes at the expense of an increase in PLR, i.e. $PLR_{WITH\_QoS\_DIFF} \geq PLR_{BE}$. We propose to quantify this effect by introducing an isolation factor, $PLR_{ISOLATION}$, and a penalty factor, $PLR_{PENALTY}$ [p19], defined in (5.2) and (5.3), respectively. The efficiency of different QoS algorithms, or their parameter setting, can then be evaluated by $PLR_{PENALTY}$ vs. $PLR_{ISOLATION}$ graphs.

$$PLR_{ISOLATION} = \frac{PLR_{CoS2}}{PLR_{CoS1}} \tag{5.2}$$

$$PLR_{PENALTY} = \frac{PLR_{WITH\_QoS\_DIFF}}{PLR_{BE}} = \frac{0.5(PLR_{CoS1} + PLR_{CoS2})}{PLR_{BE}} \approx \frac{PLR_{CoS2}}{2PLR_{BE}},\ when\ PLR_{CoS1} << PLR_{CoS2} \tag{5.3}$$

C. Performance of QoS algorithm

C.1. Context

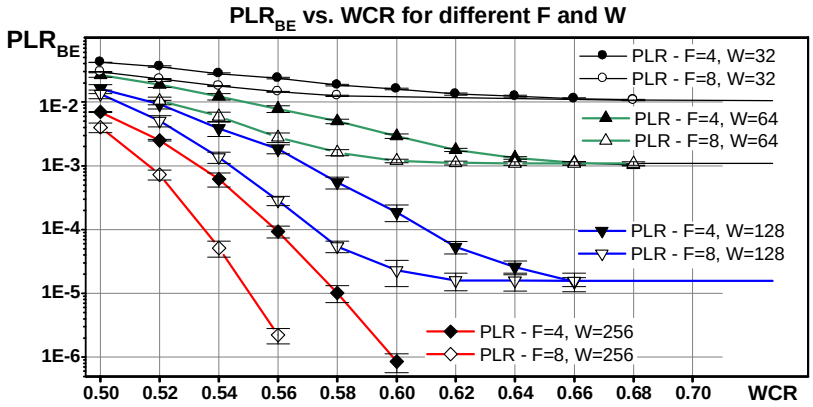

Fig. 5.9. *PLR* $_{BE}$ vs. *WCR* for average channel load of 0.7 *W* of 32-256, for *F*=4 and *F*=8.

The case study is that of the SPN TWC pool in asynchronous operation, with exponential packet length distribution and a system load of 0.7. The performance is illustrated in Fig. 5.9, and the node design is illustrated in Fig. 3.10.

Section C.2 - C.5 evaluates the QoS algorithm for *F*=4, with *W*=64 and *W*=128, to allow high performance of the wavelength domain contention resolution. Section C.6 investigates the effect of increased node adjacency (*F*=8), and Section C.7 studies overload situations. The *WCR*, being the ratio of TWC count and *FW*, is chosen in each case to reach a PLR_{BE} close to 10^{-3}, in order to obtain a significant (~40%) TWC count saving, while still be able to accommodate an increase in PLR_{CoS2}. For *F*=4, the *WCR* values are then chosen to be 0.625 and 0.57 for *W*=64 and *W*=128, respectively. This dimensioning results in PLR_{BE} values of $(1.6\pm0.1)\text{x}10^{-3}$ and $(1.1\pm0.1)\text{x}10^{-3}$, respectively.

C.2. QoS differentiation algorithm

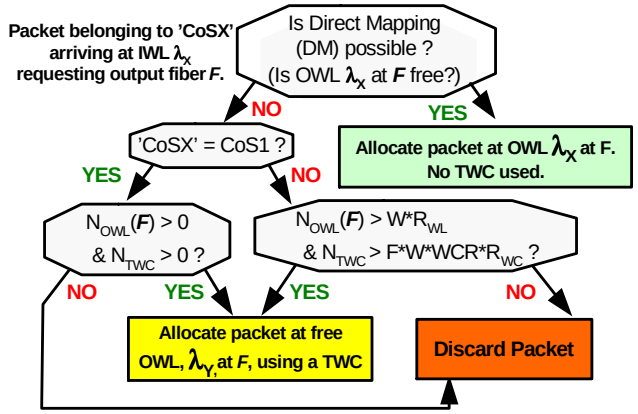

Fig. 5.10. QoS differentiation algorithm with Direct Mapping (DM) preference.

The QoS differentiation algorithm is formulated in Fig. 5.10. The ratio of reserved TWC in the pool is denoted R_{WC}, and the ratio of reserved wavelengths on any given output fibre is denoted R_{WL}. The number of free TWCs in the pool, and the number of free wavelength on output fibre F, are denoted, N_{TWC} and $N_{OWL}(F)$, respectively. Depending on the setting of R_{WC}, and R_{WL}, the algorithm can give the three cases listed below:

- A BE algorithm, i.e. no QoS differentiation when R_{WC}=R_{WL}=0 (thus $PLR_{ISOLATION}$=$PLR_{PENALTY}$=1).
- A *one-dimensional* QoS differentiation algorithm, when either R_{WC}>0 or R_{WL}>0 (exclusively).
- A *two-dimensional* QoS differentiation algorithm, when both R_{WC}>0 and R_{WL}>0.

C.3. One-dimensional approaches

One-dimensional, or single-resource based, AR algorithms are evaluated in Fig. 5.11. In the case of a pure wavelength conversion reservation scheme ("pure WC reservation"), the reservation parameter is R_{WC}, exclusively. In the case of a pure output wavelength reservation scheme ("pure WL reservation"), the reservation parameter is R_{WL}, exclusively.

Fig. 5.11 a) illustrates that for approximately the same PLR_{CoS2}, the pure WL-reservation scheme achieves lower PLR_{CoS1} than pure WC-reservation scheme in the case of W=64. Fig. 5.11 b) illustrates that for W=128 the situation is reversed; the WC-reservation scheme achieves much lower PLR_{CoS1} than the pure WL-reservation scheme, for the same PLR_{CoS2}. Furthermore, for W=128, only a very small fraction of TWCs

needs to be reserved to achieve a very low PLR_{CoS1}. These differences reflect the increased relative scarceness of TWCs at W=128, since the PLR_{BE} vs. WCR curve is very steep in the region of WCR=0.57.

Fig. 5.11 c) and d) better illustrate these algorithms' efficiency by plotting penalty vs. isolation graphs. The values of $PLR_{ISOLATION}$ (5.2) are calculated using the values of PLR_{CoS1} and PLR_{CoS2} that are plotted in Fig. 5.11 a) and b). The $PLR_{PENALTY}$ (5.3) values for W=64 and W=128 are calculated relative to the value of PLR_{BE} achieved for the same W values, W=64 and W=128, respectively. For W=64 the pure WL-reservation scheme cannot achieve an isolation factor of 100 without exceeding a penalty factor of 5. For W=128, the pure WC reservation scheme achieves a much higher isolation for a quite low penalty.

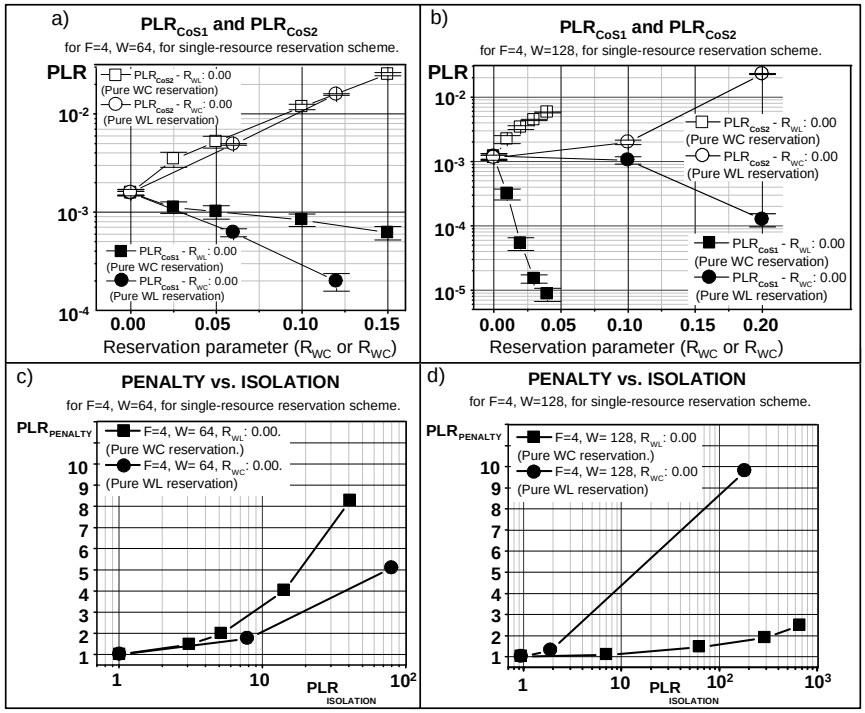

Fig. 5.11. PLR_{CoS1} and PLR_{CoS2} as a function of R_{WL} (for pure wavelength reservation) and of R_{WC} (for pure WC reservation), for a) W=64, and b) W=128. c) and d) show resulting Penalty vs. Isolation curves for pure WC reservation and pure wavelength reservation, for W=64 and W=128, respectively.

C.4. Two-dimensional approaches

To evaluate the two-dimensional QoS differentiation algorithm, the access threshold is a combination of R_{WC} and R_{WL}, effectively forming an (R_{WC}, R_{WL}) access threshold "duplet".

To quantify the performance of this algorithm, the parameter space is scanned by setting the values in the duplet equal to the values used for R_{WL} and R_{WC} in the above one-dimensional approaches.

To illustrate the PLR of both CoS in this parameter space, Fig. 5.12 a) and b) plot PLR_{CoS1} and PLR_{CoS2} for each of the three selected R_{WL} values for increasing R_{WC} values. These values are used to calculate the values for the penalty vs. isolation curves of Fig. 5.12 c) and d). Recall that the pure WC reservation corresponds to the case of R_{WL}=0; the figure confirms that these two curves overlap. Furthermore, the pure WL reservation corresponds to the case with case of R_{WC}=0; hence this curve overlaps with the left-most point on each of the two-dimensional curves.

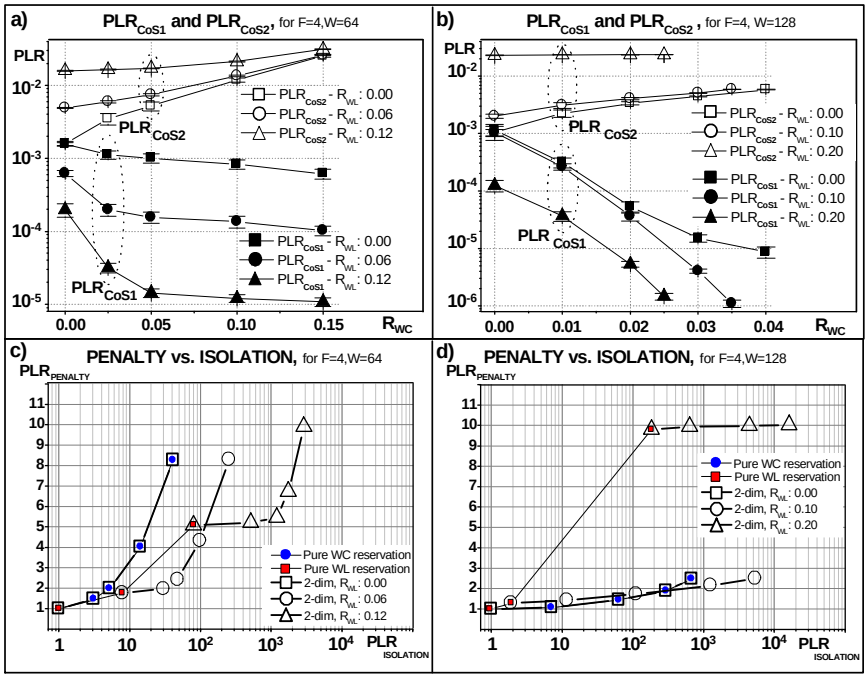

Fig. 5.12. a) and b) show PLR_{CoS1} and PLR_{CoS2} for two-dimensional algorithm. Resulting Penalty vs. Isolation curves are shown in c) and d), showing one-dimensional approaches for comparison.

As illustrated in Fig. 5.12 c) and d), for W=64 and for W=128, the two-dimensional approach has a lower or equal penalty than the one-

dimensional approaches, given the same isolation. E.g. for W=128, more than a decade improvement in isolation is obtained using the two-dimensional algorithm. Overall, higher wavelength counts decreases the penalty of high isolations, thereby enabling very low PLR_{CoS1}, when correctly setting R_{WC} and R_{WL}. Hence, we conclude that a two-dimensional approach is more efficient and flexible than one-dimensional approaches. Furthermore, the performance scales well with W.

C.5. Direct Mapping Preference

A peculiarity of the algorithm is that it allows CoS2 packets to be allocated even if either of the thresholds is violated, provided that the packet can find its own wavelength free at the requested output fibre. Whilst this makes sense from a pure TWC utilisation point of view, since the packet does not consume any TWCs anyway, the effect is more complex with respect to wavelength utilisation. In fact, this direct wavelength mapping preference of the algorithm decreases PLR_{CoS2}, but the increased utilisation may cause rejection of some CoS1 packets, increasing PLR_{CoS1}. In theory, an increased amount of directly mapped packets increases the probability that a later arriving packet from the same input fibre, going to the same output fibre will find its own wavelength free, thereby also avoiding use of a TWC for these packets. This increased number of direct mappings can be exploited to reduce the *WCR* or to decrease the overall PLR, i.e. $PLR_{WITH\_QoS\_DIFF}$, for the same *WCR*.

Fig. 5.13 compares the penalty vs. isolation graphs of the algorithm with and without such Direct Mapping (DM) preference. It confirms that for the BE case (R_{WC}=R_{WL}=0), there is no difference between the algorithms. For W=64 the two algorithms have very similar performance. But for high isolation ratio the algorithm without DM preference is slightly better. For W=128, the situation is reversed; the algorithm with DM preference has the lower penalty. This difference is attributed to the combination of the scarceness of TWCs in this system at W=128, making DM preference more attractive, and the increased size of the system, effectively decreasing the probability that all wavelengths are occupied on the output fibre, thereby minimising this source of CoS1 packet loss.

Although not studied here, it is likely that the benefit of the algorithm with DM preference increases with decreasing F. This is because a higher portion of packets with the same input-output fibre pattern will follow a directly mapped CoS2 packet within its packet duration. This increases the benefit of the reduced probability that these packets find their own wavelength being occupied. The algorithm with DM preference is used in

the remainder of this study, but note that the interplay of *F*, *W*, *WCR* and the desired $PLR_{ISOLATION}$ governs which performs better in a rather complex manner.

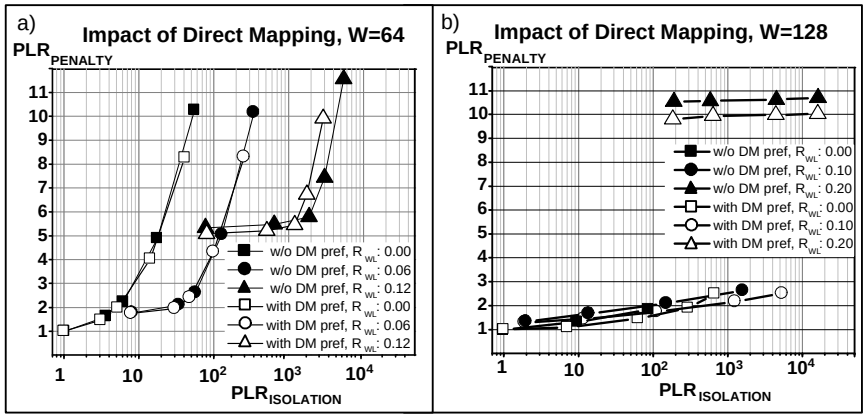

Fig. 5.13. Comparison of penalty vs. isolation graphs with and without DM preference for a) *W*=64, and b) *W*=128.

C.6. Node Adjacency

The algorithm is applied to systems with *F*=8, for *W*=64 and for *W*=128. Increased sharing of TWCs enables a lower *WCR* for the same PLR_{BE}. In our case, *WCR*s of 0.57 and 0.54 were chosen, enabling PLR_{BE} of $3x10^{-3}$ and $1x10^{-3}$, respectively. The PLR and penalty vs. isolation curves are plotted in Fig. 5.14, to study to what extent the system behaviour changes. The increased size of the system gives smaller penalties for the same isolation, but otherwise the behaviour is quite similar, with the two-dimensional algorithm again outperforming the single-dimensional one. Hence, we conclude that the algorithm scales well with *F*.

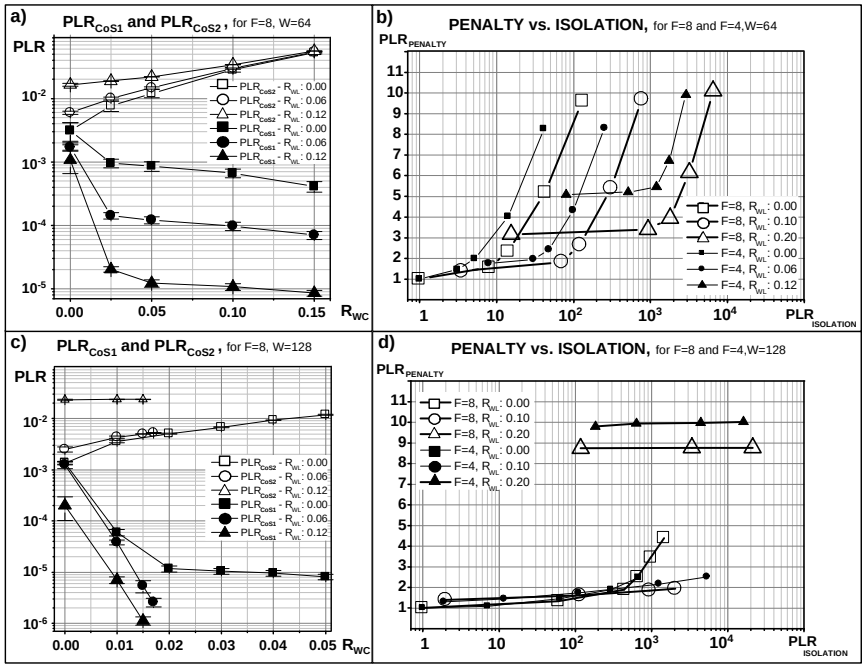

Fig. 5.14. Impact of node adjacency. PLR curves and Penalty vs. Isolation curves for W=64 and W=128, for F=4 and F=8. Penalty is expressed with respect to the PLR_{BE} of that F and W.

C.7. Load variations

The input load (as measured e.g. by observing what portion of the *FW* inputs are occupied at a given moment) at an optical packet switch constantly fluctuates during simulations, but the average load is constant. In a real network, in periods of increased client layer traffic, the average load of the optical network may increase, unless a strict access policy is applied. Even so, it may also be that a certain node in the network experiences higher average load than the network average, in periods where a large portion of packets is to be switched by this node. Such low dynamic load increase scenarios are here termed *overload* situations, and modelled as a stationary state, by simply increasing the average load for the whole duration of a simulation.

On a BE network, increased network load increases congestion and thus PLR, as discussed below. A given PLR can only to a certain degree be maintained by increasing the *WCR* [p18], which in practice calls for time-consuming and costly hardware modifications. In contrast, when the PLR_{CoS1} is of main concern, and when a QoS differentiation scheme is applied, such changes could be tolerated by monitoring input traffic and modifying the parameter setting of the QoS differentiation scheme, similar to [122].

Fig. 5.15 plots the robustness (ability to maintain a low PLR_{CoS1}) and efficiency for overload situations of 5 % and 10 % increase, i.e. average channel loads of 0.735 and 0.77, respectively. Since the fraction of CoS1 packets remain 0.50, both the number of CoS1 and CoS2 packets increase. Fig. 5.15 confirms that PLR_{CoS1} can be maintained, when reserving more resources. This robustness comes at the expense of a sharp increase in PLR_{CoS2}.

Fig. 5.16 plots the penalties, relative to PLR_{BE} of the situation *with no overload.* Hence, the lowest-penalty point of each overload series, corresponds to the penalty induced by the traffic increase only, i.e. the penalty that would occur for the same overload without QoS differentiation. The penalty factors were 4.3 and 12.1 for 5 % and 10 % overload, respectively, for W=64, and 8.4 and 27.3, respectively, for W=128. These numbers show that systems with a limited number of TWCs are quite sensitive to load variations. Comparing the additional penalty with respect to these penalty values for each series, assesses the cost of the QoS differentiation. This penalty factor is below 2, for isolations as high as 10^3 and 10^4, for W=64 and W=128, respectively. Hence, after the initial penalty resulting from the overload, the additional penalty of introducing QoS differentiation to ensure low PLR_{CoS1} is modest.

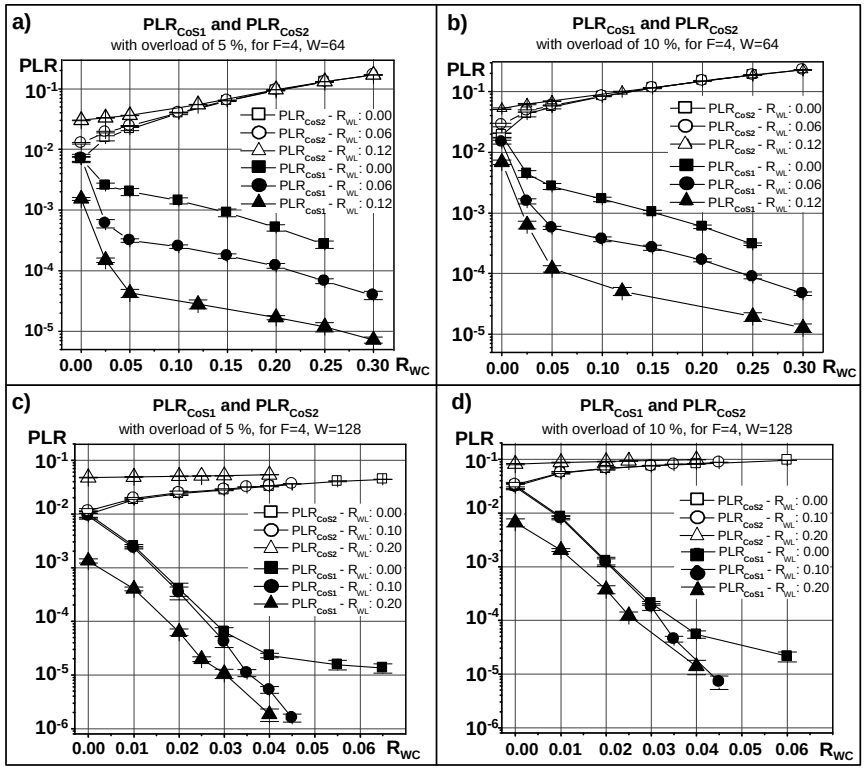

Fig. 5.15. PLR curves for overload situations of 5% and 10% for *F*=4, *W*=64 and *W*=128.

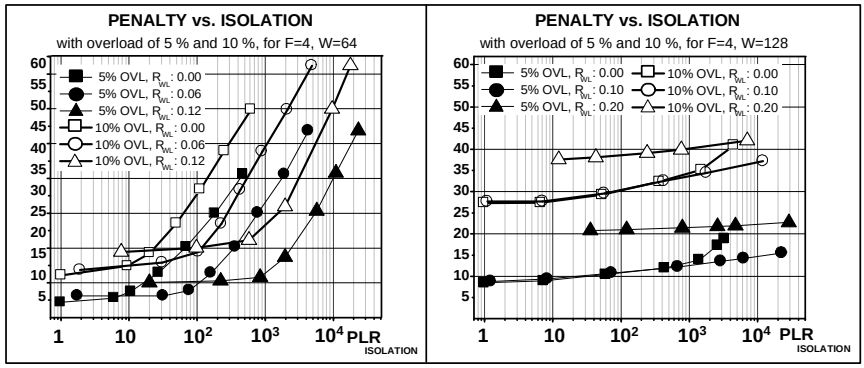

Fig. 5.16. Penalty vs. Isolation for overload situations of 5% and 10% for *F*=4, *W*=64 and *W*=128.

D. Discussion

Our results show that the two-dimensional QoS differentiation scheme enables offering a CoS with low PLR to the client layer, with a smaller increase of the PLR of the other CoS, (and thus overall PLR) than comparable single-dimensional systems.

It was also shown that a variation in the scheme, no longer allowing CoS2 packets (that can be directly mapped) to violate reservation thresholds, could further improve the penalty and isolation ratio when TWCs are not scarce.

The performance of the scheme increases with increasing F and W, but hardware realisation issues may impose limits on the scalability, due to maximum number of switch matrix port counts, and tunability of TWCs [p15].

Overload situations were studied, and it was found that the reservation parameters should be adjusted to maintain a certain PLR for the CoS1 packets. Decreases in load would lead to decreased PLR_{BE}, and is not investigated here. However, it is intuitive, that also in this situation the parameters of the QoS differentiation scheme would need to be adjusted to optimally exploit this lower load. Therefore, when the average load seen by a network node varies, the capability of signalling such changes by the management system, or distributed load monitoring with associated parameter adjustment, are needed to make the QoS differentiation scheme as efficient as possible. This is not needed in a system without QoS differentiation, and thus represents an additional cost of any AR scheme.

The choice of Poisson arrival will for most systems yield a better performance than more bursty traffic patterns [121]. On the other hand, the fraction of high priority packets is as high as 0.5. Lowering this would significantly ease the system constraints, enabling lower $PLR_{WITH\_QoS\_DIFF}$ and/or hardware savings in terms of WCR.

E. Conclusion

The proposed two-dimensional AR based QoS differentiation scheme provides an efficient means of reaching lower PLRs than what could otherwise have been reached in a Best-Effort scenario, provided that not all CoS require low PLRs. At the same time, its compatibility with the Shared Per Node design achieves ~40 % TWC count reduction.

The performance of the scheme increases with increasing number of wavelengths per fibre and increasing node adjacency, enabling lower PLRs or reduced TWC count, relative to the total switch capacity.

Whilst the PLR_{BE} of this SPN design increases significantly during overload situations, the QoS differentiation scheme itself suffers moderate additional penalties. Hence, a very low PLR of the high-priority CoS can be maintained, if a significant increase in the PLR of the other CoS can be accommodated. However, this calls for a method of adjusting the QoS differentiation parameters.

The low complexity of the algorithm facilitates avoiding the electronic scheduling bottleneck. Hence, it is particularly attractive for asynchronous OPS networks with short average packet durations. However, the algorithm can also be adapted to OBS networks.

5.4. Access Restriction in TWC+FDL SPN Pools

This chapter incorporates the OSA Journal of Optical Networking 2004 article [p22].

Packet Loss Rate- and Jitter Differentiating QoS Schemes for Asynchronous Optical Packet Switches

M. Nord [1,2] and H. Øverby [3]

(1) Research Centre COM, Technical University of Denmark, DK-2800 Lyngby, Denmark.
(2) Telenor R&D, N-1331 Fornebu, Norway.
(3) Department of Telematics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.
mn@com.dtu.dk

Abstract. We propose access restriction based Quality of Service differentiation schemes, suitable for an asynchronous optical packet switch with a contention resolution pool that contains both tunable wavelength converters and fiber delay lines. The schemes aim at obtaining a high degree of packet loss rate isolation, for a low increase in overall packet loss rate, at the same time respecting the jitter tolerance of each Class of Service. Numerical simulations quantify how the performance depends significantly on the jitter tolerance of the traffic in general, and of the highest priority Class of Service in particular.

A. Introduction

Optical Packet Switching (OPS) is a network architecture with the potential to offer huge bandwidth to core telecommunication transport networks [18, 19, 25, 34, 64, 99]. Recently, the issue of Quality of Service (QoS) differentiation in OPS has been addressed [p4, p19, 54, 89, 112, 114]. The main rationale for QoS differentiation is to respect the QoS requirements of a wide range of services, without dimensioning the whole network to comply with the most demanding one, as would be required in a Best Effort (BE) network paradigm. This is in line with one of the main reasons for introducing optical statistical multiplexing; namely efficient usage of resources, demanded by economics [18, 19]. QoS differentiation in the optical layer should be compatible with the relative-CoS-priority Differentiated Services approach in the IP layer [123], in order to facilitate a transition from today's Best Effort (BE) Internet to a QoS aware network [54, 57, 119].

In line with [p19, 54, 64, 89, 112, 114, 119], we assume an OPS network in asynchronous operation, to avoid optical synchronisers, and we use variable length packets, to provide a good match with Internet traffic. We emphasise that the QoS differentiation schemes should be compatible with OPS node designs that offer a good trade-off between performance and complexity. Therefore, we extend our work in [p19] by proposing QoS schemes that are suitable for an optical packet switch design with a shared contention resolution pool of limited size. In contrast to previous works, the pool contains Fibre Delay Lines (FDLs), in addition to Tuneable Wavelength Converters (TWCs) in order to minimise the overall Packet Loss Rate (PLR). However, using FDLs raises the issue of jitter tolerance of traffic, which we address by including maximum jitter in the CoS specification, in addition to PLR.

The remainder of this article is organised as follows: Section B discusses OPS QoS requirements. Section C describes the simulation model and the switch design. Section D outlines the QoS differentiation principle and the performance parameters. Sections E - G detail the proposed QoS differentiation schemes, which are compared and discussed in Section H, before drawing a conclusion in Section I.

B. QoS differentiation in an IP-over-OPS network concept

The most important performance related QoS parameters in a statistically multiplexed network are PLR, delay and delay jitter [57, 124]. Typically, an application communicates through a stream, so the PLR becomes the average ratio of the number of lost- and incoming packets, belonging to the stream. Delay becomes the average time spent by a packet in the flow to traverse the network, and the delay jitter quantifies variations in this delay.

The requirements of the optical layer depend on the higher layer network-, transport- and application layers. It is not possible to foresee the exact QoS requirements of future protocols and applications. However, the end-to-end performance targets of some of today's Internet based services, discussed below, may give an order-of-magnitude estimate of future performance target values.

The delay tolerance ranges from around 10 ms for PC interactive games [125], 150 ms for Voice over IP (VoIP) [124] and up to 10 s for streaming services [124]. Since the time spent in the optical packet switches, even with FDL buffering, is almost negligible to the propagation delay in the network [126], QoS differentiation of delay is not appropriate. Instead, any application-imposed delay limit should be handled at the routing level.

When it comes to acceptable network PLR, there is a multiple orders-of-magnitude mismatch, motivating a PLR-differentiation in the network. On the one hand, TCP performance [119], audio streaming services, Real-Time (RT) interactive video and VoIP [124], as well as computer oriented video streaming [125], accept a network PLR around 10^{-2}. On the other hand, some MPEG-2 based online gaming- and TV oriented streaming services require PLR in the 10^{-5} range [125].

Jitter has also been proposed incorporated into the DiffServ framework, through proportional jitter between different CoS [127]. An important aspect of jitter in packet networks is that it may cause packet misordering, whenever maximum jitter is larger than minimum packet duration. Several applications and protocols are sensitive to jitter or packet sequence integrity. Examples include:

- VoIP calls for jitter below 1 ms [124].
- High quality streaming services requires << 1 ms of jitter [124].
- Jitter disturbs the transmission of reference clock cells in MPEG-2 transmissions [128].
- Misordering of TCP segments leads to waste of bandwidth, unnecessary reduction of transmission rate, and even increased burstyness [129].
- Allowing IP packets to be fragmented over several optical packets calls for reassembly of IP packets at the OPS egress nodes, which can be more complex when packets are reordered.

Jitter and misordering can be compensated in network edge devices [128], or in transport protocol- or application level dejittering buffers [127]. In lossy networks, one then needs to determine whether a packet has been discarded in the network, or if it has simply been delayed. Bounding jitter enables making this decision in a short time, reducing the complexity if this process, and in turn reducing the buffer sizes [127]. Since the OPS network typically constitutes a part of the end-to-end path, limiting OPS network jitter will leave larger margins to the remaining end-to-end path. Better yet is to offer jitter free operation of the OPS network. This enables network designers to more freely design the protocol stack, by only accounting for the jitter in the electrical networks and its interfaces, if the path is not optical end-to-end.

In an OPS core network without deflection routing, only the FDLs induce jitter that can cause packet reordering. However, FDLs should be applied when possible in port-constrained SPN designs, to benefit from the PLR reduction they bring [p21]. This study therefore proposes a solution where the OPS ingress nodes aggregate packets from jitter tolerant streams onto a jitter tolerant CoS, which has access to the FDLs and thus may experience a jitter, whilst jitter intolerant streams are mapped onto a jitter free CoS, which do not have FDL access.

C. Optical packet switch modelling, design and dimensioning

C.1. Modelling

We assess performance by use of discrete event-driven simulations in OPNET, considering a single core optical packet switch in asynchronous operation. The strictly non-blocking switch matrix used in the packet switch is a generic design, which, being out of the scope of this study, is not modelled in detail. The performance then depends on the contention resolution pool design, as discussed below. Table 5.3 shows the node parameters of our study. To limit the parameter space, whilst showing the differences in performance of our proposed schemes, we study a fixed case with 4 input fibres- and 4 output fibres (being quite representative of a core mesh network), 32 wavelengths per fiber (typical WDM channel count using only C-band EDFA amplifiers), at an offered normalised system load of 0.6 (putting quite hard demands on the switch). At 10 Gbit/s channel rates, this represent a load of 768 Gbit/s. It should be noted that the schemes would work for other parameters as well. E.g. a load increase could be handled by adjusting QoS threshold parameters, cf. to QoS differentiation in a bufferless switch [p20]. However, this results in an increased PLR of the low priority CoS, to maintain the same PLR of the high priority CoS, unless the size of the pool is increased sufficiently.

The incoming packets are modelled by independent packet generators at each input wavelength, according to a Poisson arrival process, which is in accordance with recent measurements of the Internet core network [130]. Future work will address the impact of bursty traffic at the core switches, which in general tends to increase the PLR. The packets are subject to FIFO buffering in each packet generator before being sent to the packet switch input, to emulate output clocking of the upstream switch. The packet duration is negative exponential distributed. The mean packet duration (m.p.d.) is abstracted in the model, and is the time unit of reference, i.e. the FDL delay is expressed with respect to the m.p.d. For reference, most work on OPS assume a m.p.d. of 1-2 μs, which gives packet sizes of 1.25-2.5 kB at 10 Gbit/s channel rates. The graphs show 95 % confidence intervals obtained by the method of 10 simulated batch means. The packets' output fibre destinations, as well as their CoS (for QoS differentiation schemes), are uniformly distributed.

C.2. Optical packet switch design

Contention can be resolved in Space- [93], Time- [61], and Wavelength-domain [95], or a combination of these [96]. In this study we combine the

two latter methods, and thus avoid the ms-range jitter resulting from deflection routing. For wavelength domain contention resolution, it is not necessary to provide a Fixed Output-wavelength Wavelength Converter (FOWC) per output of the switch matrix. Instead, one can equip the packet switch with a feedback-based Shared Per Node (SPN) contention resolution pool, containing Tunable WCs (TWCs) [p19, 98], or a combination of TWCs and FDLs [p21, 87], as illustrated in Fig. 5.17 a).

We have assumed FDLs that can contain multiple packets at any wavelength, as long as they do not overlap in time. As opposed to WDM FDLs, which can contain multiple time overlapping packets on different wavelengths [54, 89], they do not require TWCs nor multiplexers to fully exploit the buffer capacity. Although we then need more FDLs for the same buffer capacity, our efficient buffer scheduling scheme still enables us to respect the space consumption imposed constraint of not using more than a few tens of FDLs [19, 54].

Using the node- and resource usage parameters defined in Table 5.3, Fig. 5.17 b) details the BE scheduling algorithm. The scheduling includes use of a hereby proposed “soft reservation” buffer scheduling algorithm, termed *SoftRSV+*, which is an enhanced version of our earlier proposed *SoftRSV* algorithm [p21].

It aims at reducing the need for TWCs, and works as follows: Consider a packet arriving from a fiber at a certain input wavelength (IWL), which must be buffered either due to lack of free output wavelengths (OWL) at the requested output fiber, or due to lack of free TWCs (in case a wavelength conversion is required). The scheme consists of reserving the OWL=IWL at the buffered packet’s requested output fiber, by incrementing a SoftRSV counter for that OWL. When a packet later arrives from an input fibre the scheduler preferrably chooses its fiber output OWL equal to its IWL. If this is taken, but there is both a free TWC and a free OWL on the output fiber, the scheduler picks the free OWL with the least number of SoftRSVs (preferrably ‘0’), in order to maximise the probability that buffered packets will not need a TWC for switching at the output of the FDL buffer.

Note that the reservation is ‘soft’ in the sense that non-buffered packets may use a SoftRSV’ed OWL, either to avoid use of TWCs (when OWL=IWL is free, but SoftRSV’ed), or when all free OWLs are SoftRSV’ed. However, if the switched packet is of shorter duration than the total remaining time the packet may spend in the FDL buffer, the buffered packet may still benefit from its SoftRSV, since the SoftRSV counter is only decremented when the buffered packet is succesfully switched or discarded.

Table 5.3. Overview of the parameters used in this study. 'CoSX' indicates an arbitrary CoS.

Node Design	Parameter Description	Parameter range
F	Number of input/output fibres	F=4
W	Number of wavelengths per fibre	W=32
A	Normalised system load	A=0.6
P	Number of contention resolution pool ports ($P = P_{TWC}+P_{FDL}$)	P=74
P_{TWC}	Number of TWCs in the contention resolution pool	P_{TWC}=[46, 74]
P_{FDL}	Number of FDLs in the contention resolution pool	P_{FDL}=[0, 28]
D_{FDL}	Delay of an FDL, relative to mean packet duration (m.p.d.)	D_{FDL}=[1, 3] m.p.d.
B_{MAX}	Max number of buffer circulations, before packet is discarded	B_{MAX}>0
Resource Usage	**Parameter Description**	**Parameter range**
$B_{COUNTER}$	Number of buffer circulations of a buffered packet	$0 \le B_{COUNTER} \le B_{MAX}$
$N_{OWL}(F)$	Number of free OWLs on output fibre F	$N_{OWL}(F)$=[0, W]
N_{TWC}	Number of free TWCs in the contention resolution pool	N_{TWC}=[0, P_{TWC}]
N_{FDL}	Number of free FDL input ports in the contention resolution pool	N_{FDL}=[0, P_{FDL}]
QoS Scheme	**Parameter Description**	**Parameter range**
$AR_{OWL,'CoSX'}$	Access Restriction threshold of 'CoSX', w.r.t. number of free OWLs	$AR_{OWL,'CoSX'}$=[0, W]
$AR_{TWC,'CoSX'}$	Access Restriction threshold of 'CoSX', w.r.t. number of free TWCs	$AR_{TWC,'CoSX'}$=[0, P_{TWC}]
$AR_{FDL,'CoSX'}$	Access Restriction threshold of 'CoSX', w.r.t. number of free FDLs	$AR_{FDL,'CoSX'}$=[0, P_{FDL}]
CoS terminology	**CoS Type Description**	
PLR_{BE_JF}	PLR of a BE (non PLR differentiated) scheme, when the CoS is Jitter Free (no FDLs)	
PLR_{BE_JT}	PLR of a BE (non PLR differentiated) scheme, when the CoS is Jitter Tolerant (with FDLs)	
PLR_{JF_CoS}	The PLR of the Jitter Free CoS when the access to the FDLs is differentiated	
PLR_{JT_CoS}	The PLR of the Jitter Tolerant CoS when the access to the FDLs is differentiated	

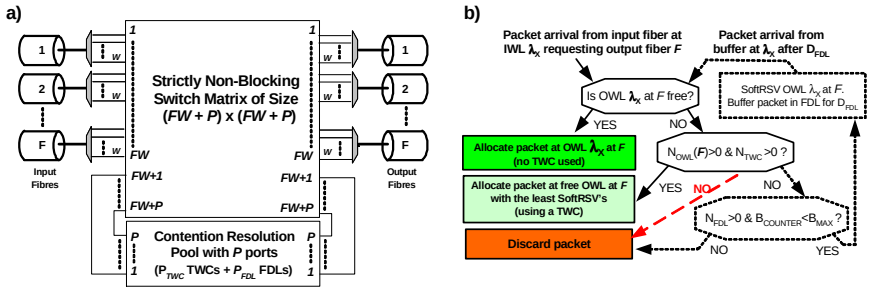

Fig. 5.17. a). The OPS node contains a generic switch matrix and a SPN contention resolution pool. b) The BE scheduling algorithm for a contention resolution pool with TWCs and FDLs. Dotted elements are replaced by the red, stippled line when P_{FDL}=0.

C.3. Switch dimensioning

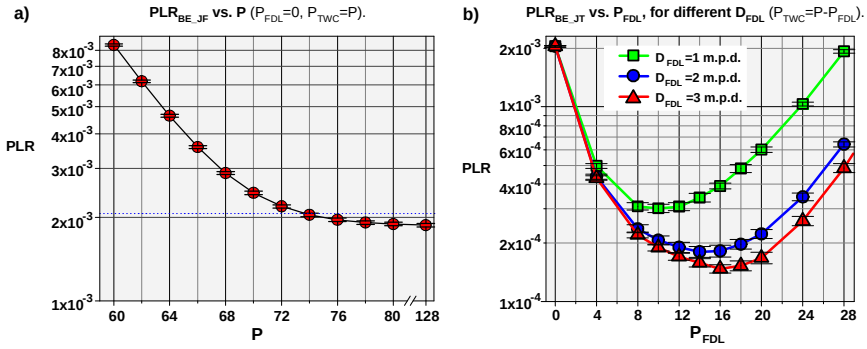

Fig. 5.18. a) PLR_{BE_JF} vs. P for bufferless switch. Note the break along the primary axis. Dotted line indicates a 10% PLR increase from FOWC case. b) PLR_{BE_JT} vs. P_{FDL} and D_{FDL}, for P=74.

A small contention resolution pool port count, P, minimises hardware resources such as switch matrix port count, and the sum of TWC- and FDL count. On the other hand, low P gives rise to blocking, when packets cannot be switched to free output wavelengths (OWLs) due to lack of TWCs, or when a free FDL cannot be found. To dimension P, we consider a pure TWC contention resolution pool (P=P_{TWC}). The PLR of this Best Effort (BE) Jitter Free switch, $PLR_{BE\_JF,}$ is shown as a function of P in Fig. 5.18 a). As in [p21, 87], P is chosen to induce a 10 % PLR increase, compared to the case of the FOWC design. This results in P=74, which has a PLR_{BE_JF} of $(2.054\pm0.022)\text{x}10^{-3}$. We fix P in the remainder of the study to enable a fair comparison of QoS schemes.

The PLR of this Best Effort (BE) Jitter Tolerant switch, $PLR_{BE\_JT,}$ is shown vs. the number of FDLs in the contention resolution pool, P_{FDL}, (P_{TWC}=P-P_{FDL}) for different delay units, D_{FDL}, in Fig. 5.18 b). The simulations confirm the existence of an optimum choice of both P_{FDL} and D_{FDL} to reach the minimum PLR_{BE_JT} [p21]. Optimum P_{FDL} signifies that with a port-constrained contention resolution pool, although TWCs are very efficient to resolve contention, some FDLs are needed to better resolve those types of contention that TWCs cannot resolve, i.e. lack of free OWLs on the output fibre. However, not all contention should be resolved in the time domain. The optimum D_{FDL} reflects a trade-off between increased buffer capacity with increased D_{FDL} on the one hand, and decreased time granularity and decreased effects of *SoftRSV+* scheme on the other hand [p21]. For clarity, we only show the curves with D_{FDL} up to 3 m.p.d, which yields near-optimum performance, whilst limiting total FDL length.

For D_{FDL}=3 m.p.d., the minimum PLR_{BE_JT} is situated at P_{FDL}=16, which respects the space constraints discussed in Section C.2. However, each FDL circulation induces signal quality degradation [99], and increases the maximum jitter. Limiting the number of buffer circulations combats these drawbacks. By simulations, we assessed the influence of the maximum number of buffer circulations, and we found that beyond B_{MAX}=3, PLR_{BE_JT} does not decrease further. This parameter is maintained in this study, resulting in a maximum jitter for an optical packet switch of 9 m.p.d. This results in a maximum jitter below 0.1 ms in a network with a m.p.d. of 1 μs, even for the very unlikely case of a packet experiencing a maximum delay of $B_{MAX} \times D_{FDL}$ in all nodes along a 10-node long path. The minimum PLR for a BE switch with P_{FDL}=16 is termed $PLR_{BE\_JT\_MIN}$=1.50×10^{-4}, and is used as a reference value in the penalty definition (5.6).

D. Quality of Service differentiation by Access Restriction

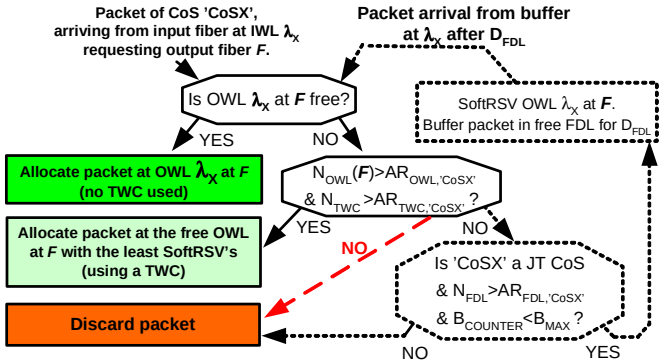

Fig. 5.19. QoS differentiation algorithm. Dotted elements are replaced by the red, stippled line for bufferless switches.

OPS approaches to QoS differentiation can roughly be divided into dropping-based, Access Restriction (AR)-based, and pre-emptive techniques [p4]. A quantitative comparison of these methods is out of the scope of this article, but will be addressed in an upcoming overview article [p25]. In short, the pros and cons of the three techniques can be summed up qualitatively:

- The dropping based technique discards low-priority packets randomly, in order to decrease PLR of the high-priority CoS, which results in a high overall PLR. The random discard policy gives it a low scheduling complexity. The optical hardware complexity is low since it only requires being able to discard certain packets arriving at the inputs.
- The AR based technique results in an improved performance since packets are only dropped when the switch is strained. The scheduling is more complex, since it has to take current resource allocation into account before deciding whether a low priority packet should be discarded when arriving at the input. The optical hardware complexity is similar to that of the dropping based approach.
- The lowest overall PLR results from the pre-emptive approach, which only discards a low-priority packet, when a high-priority packet would otherwise be lost. However, this technique requires more complex scheduling, since it should track the time-line of allocated packets to choose the best packet to pre-empt. Furthermore, pre-emption calls for being able to detect and erase pre-empted packets.

We believe the AR approach to be a good compromise between performance and complexity. In this study, we further develop our AR-

based approach [p19], by adapting the QoS algorithms to an optical packet switch with FDLs. The algorithm is detailed in Fig. 5.19, using the QoS and CoS parameter definitions in Table 5.3. As pointed out in earlier work [p19], satisfying performance of the AR method requires correctly setting the relevant access threshold parameters and their values. In the ideal case, any loss of the lower priority CoS packet should be rewarded by avoidance of loss of a higher priority CoS packet. However, in asynchronous operation, as opposed to slotted operation, the scheduling is done without knowing future effects of the allocation. Indeed, the mismatch of static AR threshold settings with the statistical nature of packets' arrival time, duration, CoS and requested output, results in sub-optimum resource usage. A good operation point reserves enough resources for the high priority CoS to ensure a low PLR of this CoS. On the other hand, it should not reserve excessive resources, to avoid an unnecessarily increase in the PLR of the low-priority CoS, when being deprived of accessing a high portion of switch resources. This point will be evidenced by a minimum increase in total PLR, thus $PLR_{PENALTY}$, for the desired difference in the isolation ratio of the PLRs, $PLR_{ISOLATION}$, as defined in (5.4)-(5.6), in which 'CoSX' and 'CoSY' denotes the two considered CoS. According to (5.4), $PLR_{ISOLATION}$ is equal to or larger than unity. (5.6) quantifies the cost of the scheme, in terms of the overall PLR, compared to what can be achieved when FDLs are accessible for both CoS, in the BE case, i.e. when no PLR differentiation is desired. Note that the penalty also includes the effect of some traffic being jitter free, thus being deprived of FDL access, since (5.6) uses the $PLR_{BE\_JT\_MIN}$ as denominator, as explained in Section C.3.

$$PLR_{ISOLATION}('CoSX','CoSY') = \max(PLR_{CoSX'}, PLR_{CoSY'}) / \min(PLR_{CoSX'}, PLR_{CoSY'}) \quad (5.4)$$

$$PLR_{OVERALL}('CoSX','CoSY') = 0.5(PLR_{CoSX'} + PLR_{CoSY'}) \quad (5.5)$$

$$PLR_{PENALTY} = (PLR_{OVERALL} / PLR_{BE\_JT\_MIN}) \quad (5.6)$$

The penalty and isolation parameters are essential to quantify the performance of different QoS differentiation schemes. The higher the isolation, and the lower the penalty, the more efficient the scheme is. To exemplify the values of the parameters, consider two CoS, with $PLR_{ISOLATION}$=100 and $PLR_{PENALTY}$=10. Then the PLR of the low-priority CoS would be ~20 times that of $PLR_{BE\_JT\_MIN}$, thus ~$3x10^{-3}$, whilst the PLR of the high-priority CoS would be a ~100 times lower than the low-priority CoS, thus equal to ~$3x10^{-5}$, which is a reduction of a factor of 5 compared to $PLR_{BE\_JT\_MIN}$. In Sections E - G, the proposed schemes are assessed, before comparing their performance in Section H.

E. QoS by AR in bufferless OPS nodes: Jitter Free Scheme

In bufferless OPS nodes (P=74=P_{TWC}), the jitter tolerance does not have to be considered. The drawback is that compared to the $PLR_{BE\_JT\_MIN}$ reference value, which benefited from FDLs, this design suffers a penalty even before introducing QoS differentiation of the PLR, i.e. $PLR_{PENALTY}$=(13.66±0.14) for $PLR_{ISOLATION}$=1. To achieve QoS differentiation of the PLR of two jitter free CoS, termed JF_CoS1 and JF_CoS2, one applies access restriction on OWLs and TWCs only for the JF_CoS2 packets. In the algorithm, this means having AR_{OWL,JF_CoS2}≥0 and AR_{TWC,JF_CoS2}≥0, whilst AR_{OWL,JF_CoS1}=AR_{TWC,JF_CoS1}=0.

This scheduling algorithm is detailed in Fig. 5.19, with the simplification of replacing the dotted boxes by a "discard state", indicated by the red, stippled line. This scheme decreases PLR_{JF_CoS1} at the expense of an increased PLR_{JF_CoS2}. Fig. 5.20 a) plots resulting $PLR_{PENALTY}$ vs. $PLR_{ISOLATION}$ curves. Increasing the AR thresholds for JF_CoS2 increases the $PLR_{ISOLATION}$, but also $PLR_{PENALTY}$. The values of the AR parameters are not the main point here, but many values are included to make the point of [p19]: to obtain a certain $PLR_{ISOLATION}$ with a minimum $PLR_{PENALTY}$ an optimum choice of both TWC and OWL AR thresholds should be made. However, a high $PLR_{ISOLATION}$ can only be obtained for a high $PLR_{PENALTY}$.

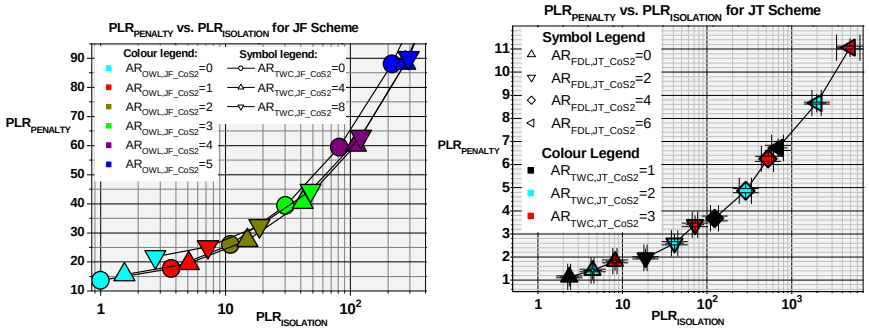

Fig. 5.20. a) $PLR_{PENALTY}$ vs. $PLR_{ISOLATION}$ for the JF Scheme. Confidence intervals are omitted for clarity, but are well within the symbol size. b) $PLR_{PENALTY}$ vs. $PLR_{ISOLATION}$ for JT Scheme.

F. QoS differentiation in OPS node with FDL buffers: Jitter Tolerant Scheme

For QoS differentiation of two jitter tolerant CoS (JT_CoS), JT_CoS1 can have a lower PLR than JT_CoS2, by applying AR thresholds >0 for

JT_CoS2 packets. However, both JT_CoS1 and JT_CoS2 packets can use FDLs, provided that they respect the AR thresholds.

The performance will depend on P_{FDL}, and on the three AR thresholds, AR_{FDL,JT_CoS2}, AR_{TWC,JT_CoS2} and AR_{OWL,JT_CoS2}. A parametric simulation scan, varying the numbers of FDLs in the pool, as well as the three AR parameters, showed that P_{FDL}=16 gave low penalties compared to other P_{FDL} values, over a wide isolation range, and this value is maintained in this section. More parametric scans were conducted, showing that AR_{OWL,JT_CoS2} was the least efficient AR parameter; it generally gives higher penalty for a given isolation than what can be obtained by modifying the other two AR parameters. To limit the parameter space, we thus maintain AR_{OWL,JT_CoS2}=0 in this section, which would also simplify the AR parameter setting in a real switch. Fig. 5.20 b) shows that e.g. a $PLR_{ISOLATION}$ of (690±109) is obtained by reserving 6 out of the 16 FDL inputs, and by reserving 1 out of the 58 TWCs for a $PLR_{PENALTY}$ of (6.7±0.1). Moreover, $PLR_{ISOLATION}$>5000 can be obtained for $PLR_{PENALTY}$<12.

G. QoS in OPS nodes with FDL buffers: Partially Jitter Free Schemes

In the two former QoS differentiation schemes, both CoS were either jitter free or jitter tolerant. The Partially Jitter Free (PJF) schemes aim at offering jitter free CoS (JF_CoS) and jitter tolerant CoS (JT_CoS) simultaneously. Hence, in a network with one JF_CoS and one JT_CoS, the service provider should be able to offer a PLR of the JF_CoS that is either equal, higher or lower than that of the JT_CoS. This can be realised by the four Partially Jitter Free (PJF) schemes discussed in Section G.1 – G.4, and summed up in Table 5.4. All of them operate with only the JT_CoS having access to the FDL buffers. There are then three parameters that govern the performance: P_{FDL}, AR_{TWC} and AR_{OWL}. In addition, Section G.5 proposes a PLR- and jitter decoupled PJF scheme, termed PJF_DCP Scheme, where packets are mapped onto one of 4 CoS, depending on the desired PLR level and jitter tolerance.

Table 5.4. Summary of Partially Jitter Free (PJF) Schemes.

PJF Scheme	Relative PLR	P_{FDL}	JF_CoS AR settings		JT_CoS AR settings	
			$AR_{TWC,\ JF\_CoS}$	$AR_{OWL,\ JF\_CoS}$	$AR_{TWC,\ JT\_CoS}$	$AR_{OWL,\ JT\_CoS}$
BE_PJF Scheme	$PLR_{JF\_CoS} \sim PLR_{JT\_CoS}$	>0	=0	=0	≥0	≥0
PJF Scheme 1	$PLR_{JF\_CoS} > PLR_{JT\_CoS}$	>0	=0	=0	=0	=0
PJF Scheme 2	$PLR_{JF\_CoS} > PLR_{JT\_CoS}$	>0	≥0	≥0	=0	=0
PJF Scheme 3	$PLR_{JF\_CoS} < PLR_{JT\_CoS}$	>0	=0	=0	≥0	≥0

G.1. BE_PJF Scheme

Given a P_{FDL}, and with all AR parameters set to '0', the PLR of a JT_CoS will be lower than that of a JF_CoS. When the QoS differentiation should be based only on jitter tolerance, one should balance the PLR of the JT_CoS with that of the JF_CoS, to obtain a single PLR parameter, termed $PLR_{BE\_PJF.}$

To this end one applies AR on TWCs and on OWLs for the JT_CoS packets. This increases PLR_{JT_CoS} and reduces utilisation of node resources, which in turn lowers PLR_{JF_CoS}. A parametric scan was conducted to identify optimum choice of parameters to minimise $PLR_{PENALTY}$. Choosing P_{FDL}=12, AR_{TWC,JT_CoS}=4 and AR_{OWL,JT_CoS}=3 gives a difference in PLR_{JF_CoS} and PLR_{JT_CoS} below 3 %, hence $PLR_{ISOLATION}$~1. This was achieved for $PLR_{PENALTY}$=(4.3±0.07), which corresponds to a penalty reduction by more than a factor of 3, compared to that of the BE_JF scheme, c.f. Section E. Hence, the BE_PJF Scheme offers an attractive approach of lowering the PLR through use of FDL buffers, when no PLR differentiation is required, but when only some of the traffic tolerates jitter.

G.2. PJF Scheme 1

PJF Scheme 1 exploits the PLR differentiation obtained when only allowing FDL access to the JT_CoS. No other AR parameters are used, and therefore P_{FDL} governs $PLR_{ISOLATION}$, as shown in Fig. 5.21 a). As expected, at P_{FDL}=0, both CoS have the same PLR, confirmed by $PLR_{ISOLATION}$=1, and the penalty is that of bufferless nodes, i.e. $PLR_{PENALTY}$=(13.66±0.14). Increasing P_{FDL} increases the buffering capacity of JT_CoS packets, which decreases PLR_{JT_CoS}, and also the $PLR_{PENALTY}$. However, since P_{TWC} is reduced accordingly, and since the JF_CoS packets more often will find the OWLs occupied by a JT_CoS packet, the JF_CoS is penalised. Initially, the benefit of buffering lowers overall PLR, but at P_{FDL}=3 this trade-off yields a minimum $PLR_{PENALTY}$=(8.64±0.13), for $PLR_{ISOLATION}$=(10.11±0.21). Further increasing P_{FDL} increases $PLR_{ISOLATION}$, but at the expense of an increased $PLR_{PENALTY}$.

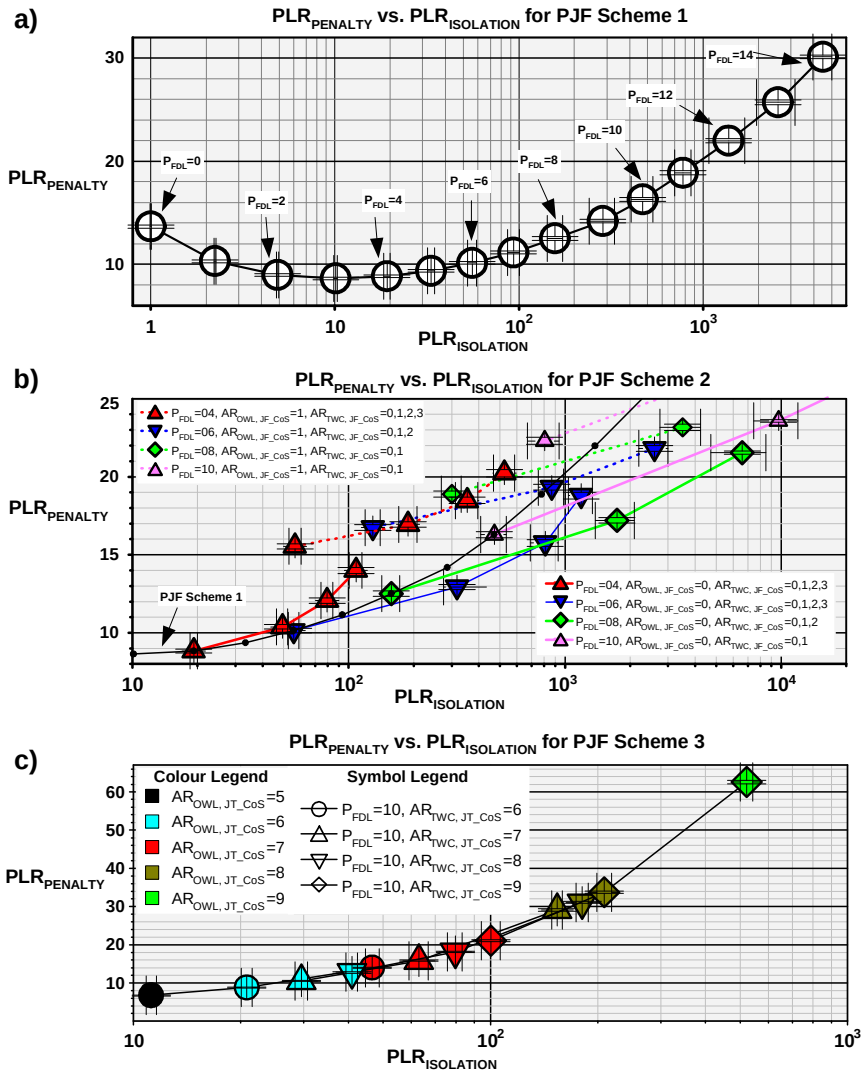

Fig. 5.21. *$PLR_{PENALTY}$* vs. *$PLR_{ISOLATION}$* for: a) PJF Scheme 1, b) PJF Scheme 2 for *P_{FDL}*=4-10 and PJF Scheme 1 for comparison (black line), c) PJF Scheme 3 for *P_{FDL}*=10.

G.3. PJF Scheme 2

PJF Scheme 2 enables the network operator to vary *$PLR_{ISOLATION}$* by fixing a moderate value of *P_{FDL}*, and then increase *$PLR_{ISOLATION}$* by applying AR on OWLs and TWCs for JF_CoS packets. Hence, this sceme is not dependent on hardware modification, as opposed to PJF Scheme 1.

A simulation scan revealed that the best parameter range was for AR_{TWC,JF_CoS} values of 0 to 3, and AR_{OWL,JF_CoS} values of 0 and 1, depending on P_{FDL}. Fig. 5.21 b) plots the results for AR parameters within this range. The results are plotted for P_{FDL} values of 4 to 10 by steps of 2; in each series the $PLR_{ISOLATION}$ value increases with the incremented AR_{TWC,JF_CoS} values (starting at '0'). We confirm that the PJF Scheme 2 curves intersect with the PJF Scheme 1 curve (indicated by black line), at the corresponding P_{FDL} value, when all AR thresholds are 0. For $P_{FDL}\geq 6$, the curves with $AR_{OWL,JF\_CoS}=1$ suffer from a higher penalty than what can be obtained by maintaining $AR_{OWL,JF\_CoS}=0$ and instead increment AR_{TWC,JF_CoS}. Hence, in practice, only the AR_{TWC,JF_CoS} parameter is needed to control the isolation degree, which simplifies operation. Using e.g. $P_{FDL}=8$, enables a wide range of isolation degrees by adjusting AR_{TWC,JF_CoS} only, although the isolation granularity is somewhat limited.

G.4. PJF Scheme 3

In the PJF Schemes 1-2, we have $PLR_{JF\_CoS}\geq PLR_{JT\_CoS}$. In contrast, applying sufficiently high AR thresholds to TWCs and OWLs for JT_CoS packets will give a JF_CoS with lower PLR than that of the JT_CoS, in spite of its lack of FDL access. The JF_CoS then becomes a 'super-priority CoS', in which packets have no jitter and the lowest PLR of the two CoS.

The performance of PJF Scheme 3, was studied for different values of P_{FDL}. It was found that $P_{FDL}=10$ enables a very good performance for $PLR_{ISOLATION}$ in the 10-300 range, by choosing suitable values for AR_{TWC,JT_CoS} and AR_{OWL,JT_CoS}. It was to be expected that this optimal FDL count is lower than $P_{FDL}=16$ (optimum for the BE_JT case) since high values of P_{FDL} penalises the JF_CoS through reduced TWC count. In turn this would require AR on an excessive amount of resources for JT_CoS packets. The results are plotted in Fig. 5.21 c). We observe that to obtain the same $PLR_{ISOLATION}$, a higher number of TWCs and OWLs should be reserved, compared to PJF Scheme 2.

G.5. PJF_DCP Scheme: Decoupling jitter and PLR.

Jitter tolerance and PLR are orthogonal values for the clients, in that the client should ideally be able to choose these independently. This PJF_DCP scheme is a partially jitter free scheme, that decouples jitter and PLR, to enable offering delay-jitter and PLR as orthogonal QoS parameters. We assume that two PLR thresholds are needed, each offered as a jitter free and a jitter tolerant CoS. The CoS names and properties are given in Table 5.5. One seeks to obtain the case in which:

$$(PLR_{JT\_CoS1} \sim PLR_{JF\_CoS1}) < (PLR_{JF\_CoS2} \sim PLR_{JT\_CoS2}) \quad (5.7)$$

Although strictly speaking there now are four CoS, there are only two PLR levels, so that (5.4)-(5.6) applies for the calculation of $PLR_{ISOLATION}$ and $PLR_{PENALTY}$. For each level, we only consider the value of PLR values of (PLR_{JF_CoS1}, $_{JT\_CoS1}$), and of (PLR_{JF_CoS2}, PLR_{JT_CoS2}) that give the worst-case for the calculation of $PLR_{ISOLATION}$ and $PLR_{PENALTY.}$

Table 5.5. CoS and AR thresholds for PJF_DCP Scheme, (P_{FDL}=11).

CoS	Delay-jitter	AR_{TWC}	AR_{OWL}	AR_{FDL}	PLR
JF_CoS1	0	**0**	**0**	N/A	$(1.14\pm0.03)\times10^{-4}$
JT_CoS1	max 3 m.p.d.	4	4	**0**	$(1.21\pm0.05)\times10^{-4}$
JF_CoS2	0	2	2	N/A	$(3.72\pm0.05)\times10^{-3}$
JT_CoS2	max 3 m.p.d.	4	4	4	$(3.45\pm0.05)\times10^{-3}$

In addition to P_{FDL}, a total of three AR parameters can be used to differentiate the PLR for each CoS, resulting in 12 AR threshold values to determine. However, AR_{FDL} thresholds are Non Applicable (N/A) for JF_CoS1 and JF_CoS2. Moreover, intuition suggests setting some of the JT_CoS1 and JF_CoS1 AR thresholds to '0', as indicated in bold in Table 5.5. This reduces the parameter space to 7 values. Nevertheless, the target of obtaining the lowest overall PLR, given a desired ratio of the two PLR levels, becomes a complex optimisation problem. This is out of the scope of this article, but a simulation study showed a parameter setting, indicated in Table 5.5, proving the feasibility of the DCP_PJF Scheme, since resulting PLR values respect (5.7). A $PLR_{ISOLATION}$ of 28.6 was reached for $PLR_{PENALTY}$ of 12.3.

H. Comparison and Discussion

H.1. Comparison of the schemes

Fig. 5.22 sums up the performance of all proposed schemes, by plotting a selected set of $PLR_{PENALTY}$ vs. $PLR_{ISOLATION}$ curves from the above presented results. Since the aim is to compare performance between the schemes, we omit a detailed legend for clarity, and refer the reader to the corresponding section of each scheme for the detailed parameter setting.

First, we study the jitter-free, partially jitter-free and jitter tolerant BE schemes, namely BE_JF, BE_PJF and BE_JT, which all have a $PLR_{ISOLATION}$ of 1. The penalties of the BE_JF and the BE_JT schemes vary by more than a decade whilst the BE_PJF scheme is situated in between. Clearly, the less tolerant the traffic is to jitter, the worse the performance. This can be expected, since FDLs are essential to minimise PLR in our SPN based node design with limited pool size, cf. Fig. 5.18 b).

When it comes to the QoS schemes with PLR differentiation, we limit the region of interest to below the threshold penalty of 33, corresponding to a PLR value of the lowest priority CoS of ~1 %.

Let us first consider the extreme cases of the JF Scheme and the JT Scheme: The former has the highest penalty for all isolation values, and the 1 % low priority PLR threshold limits the isolation degree to around 20. In contrast, the JT Scheme has a decade decrease in penalty, throughout the studied isolation range. Furthermore, it can reach isolation ratios above 5000 without violating the penalty threshold value.

PJF Schemes 1-3 has roughly the same penalties for isolation ratios from 10 and up to 30, after which the performance of PJF Scheme 3 detoriates. Still, it outperforms the JF Scheme. Hence, employing FDLs enables a PLR reduction even when the traffic that should have the lowest PLR does not tolerate jitter. However, its poor performance compared to PJF Scheme 1 and PJF Scheme 2, shows that offering such a ‘super priority CoS’, i.e. a jitter free CoS with low PLR, is more costly than letting the CoS with the lowest PLR be the jitter tolerant CoS. Comparing PJF Scheme 1 and PJF Scheme 2, the latter has lower penalty at high isolation rates, due to the increased flexibility enabled by the AR thresholds. This flexibility can also be exploited to adjust the isolation range while maintaining P_{FDL} fixed, i.e. not having to replace any TWCs by FDLs. Such physical intervention in a switch is unattractive from a network operator’s point of view, since manual labour is costly, and since it may disrupt network operation for a non-negligible time.

Finally, the PJF_DCP Scheme has the highest CoS granularity, effectively operating with 4 CoS. Being able to freely choose between the two offered PLR levels, and between jitter tolerant and jitter free switching, comes at the expense of a relatively high penalty. Since half of the jitter free traffic should have a low PLR, it can be expected that the penalty is above PJF Scheme 1 and PJF Scheme 2. We attribute the scheme's increased penalty compared to PJF Scheme 3 to its higher CoS granularity, although the scheme should be studied over an increased isolation range to draw decisive conclusions.

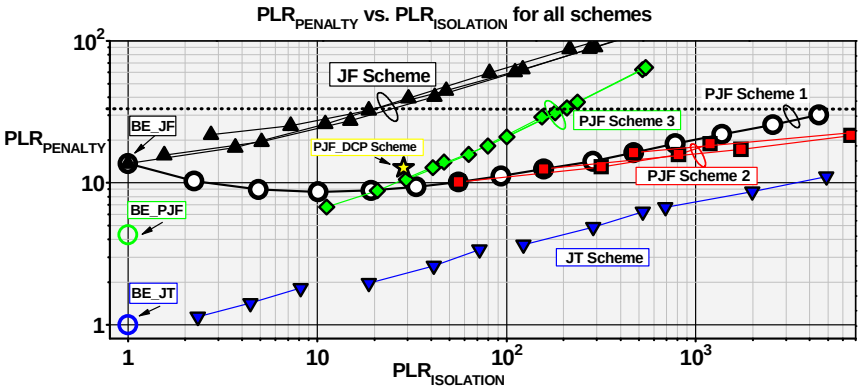

Fig. 5.22. *$PLR_{PENALTY}$* vs. *$PLR_{ISOLATION}$* for selected values of all proposed schemes. The dotted line indicates PLR of 1% for the lowest priority CoS.

H.2. Discussion

Penalty and isolation parameters are suitable for comparing the relative performance of the schemes, and their individual parameter setting. However, most network designers are interested in the PLR values that can be offered to both CoS. Therefore, we plot the PLR values of the lower-priority CoS vs. the PLR of the higher-priority CoS in Fig. 5.23. Note that these values can be adjusted by adjusting the size of the contention resolution pool, or other parameters in this study. Hence, these values cannot be more than a mere example, but they nevertheless highlight how AR can be applied to provide efficient PLR differentiation. Section B suggested that the most demanding network PLR should be ~10^{-5}, which means that the PLR in a single node should be well below that, depending on the path hop count. Taking the example of a single node high-priority CoS PLR of $5x10^{-6}$, Fig. 5.23 shows that the resulting PLR of the low-priority CoS can be below 1 % only for the JT Scheme, the PJF Scheme 1 and PJF Scheme 2. Hence, in this example, these PLR thresholds can only be met when either both CoS, or the high-priority

CoS exclusively, is jitter tolerant. Fig. 5.23 also highlights the benefit of QoS differentiation: Even the BE_JT scheme cannot reach PLR values below 10^{-4}, whilst all QoS schemes that tolerate some jitter can have a high priority CoS with a PLR as low as $6x10^{-5}$ without violating the 1 % PLR threshold of the lower priority CoS.

It is not possible to foresee the QoS requirements of applications and protocols at the time of OPS network implementation, potentially still many years down the line. Moreover, the impact of the surrounding network must also be taken into account, and we do not know what it will look like. However, the general trends provides some key results that we believe to be of general interest for optical networking research: The work supplements that of [p21], in showing that the use of simple FDL buffers enables a significant performance improvement in an optical packet switch with a port-constrained SPN contention resolution pool, compared to the bufferless approach, also when QoS differentiation on PLR and on jitter is offered. Moreover, the relative strong difference in performance between the different schemes highlights the strong impact jitter tolerance has on OPS performance, and that it is intimately related to the PLR differentiation in a QoS differentiation paradigm.

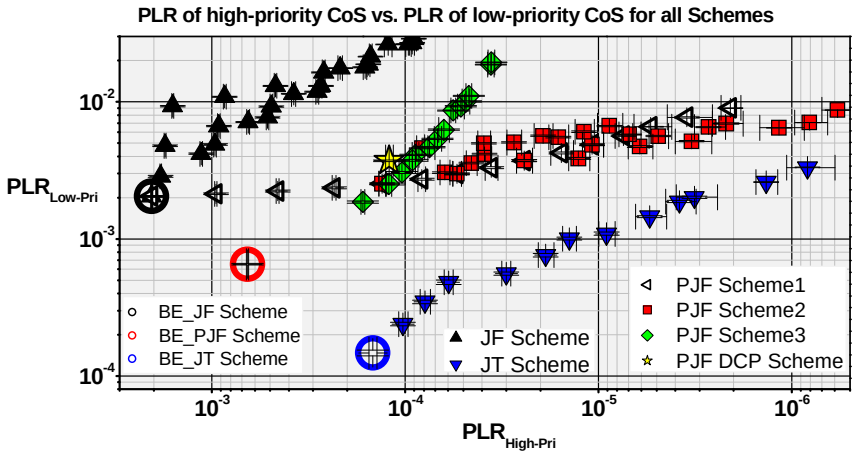

Fig. 5.23. The PLR of the lower priority CoS vs. the PLR of the higher priority CoS, for all studied values for all QoS schemes. Note the inverse scale of the x-axis.

I. Conclusion

Future higher-layer networks being served by an optical layer may benefit from being able to choose between CoS that are jitter free and CoS that tolerate a bounded jitter, in addition to PLR differentiation. We have proposed and evaluated several Access Restriction (AR) based QoS differentiation schemes enabling such two-dimensional QoS differentiation for use in a SPN optical packet switch.

The study shows that very large isolation values can be obtained, but that overall PLR detoriates with reduced jitter tolerance of the traffic, quantified to a decade decrease in overall PLR, for PLR isolation values ranging from 1 to above 10^4. Moreover, when having a jitter free CoS and a jitter tolerant CoS, overall PLR increases by a factor of ~2-4 in the isolation range from 100-700, when offering a 'super-priority CoS' with low-PLR and jitter-free operation, as opposed to a low-PLR, jitter-tolerant CoS. Finally, a decoupled scheme with increased CoS granularity also detoriates performance. Still, all these schemes are better than the QoS scheme that does not employ FDLs. These properties suggest that both the PLR and jitter properties of the network's expected traffic matrix should be carefully analysed before dimensioning the optical packet switch and selection of a QoS differentiation scheme.

The potential benefit of applying QoS differentiation is to support a wider range of services, which in turn may increase the operator's income, provided that the cost of implementing the QoS differentiation is sufficiently low. Since our schemes are of relatively low-complexity, and since they are suitable for asynchronous OPS switches minimising contention resolution hardware resources, we believe they are attractive candidates to realise a future optical statistically multiplexed network.

6. Metro Networks

6.1. Introduction

The Metropolitan Area Network (MAN) is increasingly seen as a potential application of OPS. Both IST DAVID and Virtual Department 2 (VD2) of e-Photon/One address this segment. An interesting point is that since this segment experiences less aggregation than the core of the network, the traffic is expected to be more self-similar in this area. Covering a geographically limited area, the metro area may also be a realistic segment to introduce OPS [26].

- Chapter 6.2 incorporates a Photonics in Switching 2003 conference paper [p14]. It reports a novel node design, which uses AA-MZIs both to switch and wavelength convert the packets to be forwarded
- Chapter 6.3 incorporates an OSA JON 2005 article [p23]. It investigates the performance of an interconnected Optical Packet Switched Ring Network (OPSRN) when applying the proposed "*Asynchronous Insertion Priority Scheduling with Insertion Threshold*" (*AIPSwIT*) MAC protocol, which enables support of VLP.
- Chapter 6.4 incorporates an article submitted to Elsevier Journal on Optical Switching and Networking [p26]. This article addresses fairness in OPSRNs, by extending the *AIPSwIT* to also include this feature. First, it highlights the good performance of the OPSRN, by comparing it with a Static Wavelength Routed Optical Network, for uniform traffic. Then, for unbalanced traffic matrices, it highlights a fairness-throughput trade-off. Nevertheless, the study shows that the combination of a flexible node architecture and the *AIPSwIT* MAC protocol supports high loads even for a quite unbalanced traffic matrix.

6.2. Demonstration of Ring Node Designs

This chapter incorporates a Post-Deadline paper from the Photonics in Switching conference 2003 [p14].

Novel strictly non-blocking Node Designs for asynchronous OPS MAN

M. Nord [1,2], S. Bjørnstad [2,3], M.L. Nielsen [1], B. Dagens [4]

(1) Research Center COM, Technical University of Denmark, B-345V, DK-2800 Lyngby, Denmark.
(2) Telenor R&D, N-1331, Norway.
(3) Norwegian University of Science and Technology, 7491, Trondheim, Norway
(4) Alcatel R&I, Route de Nozay, 91460 Marcoussis, France

Abstract: We propose novel designs for strictly non-blcking ring nodes and ring interchangers, suited for asynchronous optical packet switched networks. The designs enable ring interconnection and interface to a wide area network. Combining optical multicast, full bandwidth sharing, wavelength conversion and space reuse maximises link utilisation. We demonstrate viability of forwarding functions by proof-of-principle experiments.

A. Introduction

Metropolitan Area Networks (MANs) aim to interconnect different access networks and high-end users, possibly crossing multiple MANs or even Wide Area Networks (WANs). Optical Packet Switching (OPS) appears to be a good candidate for MAN applications [131]. Ring architectures are prominent candidates for OPS MANs, minimising overall fibre length and reducing node complexity, compared to mesh networks, which requires switching between a higher number of fibres. Work on MAN OPS networks have focused on slotted operation [25]. To avoid complex synchronisers, minimise packet overhead and increase freedom in packet assembly, we here consider asynchronous, variable length packets.

The paper is organised as follows: Section B discusses design issues for single-ring and multiple-ring OPS MAN networks. Section C describes Ring Node and Ring Interchanger designs. Section D describes set-up, reports experimental results and compares with an existing design. The study is concluded in Section E.

B. Design of OPS MAN networks

B.1. Ring network parameters

Table 6.1 identifies four ring features that have significant influence on ring throughput and node complexity. The design choices made in this paper are indicated in bold.

Table 6.1. Main throughput-related OPS MAN ring features.

Ring Feature	Lower throughput	Higher throughput
Space Reuse	No	**Yes**
Directionality	**Uni-directional**	Bi-directional
Transfer type	**Unicast only**	Multicast enabled
Full BW sharing	NO	**Yes**

Space reuse increases link utilisation by enabling nodes to reuse ring wavelengths, instead of them being reserved for a whole round on the ring. In a slotted OPS MAN unidirectional ring network, link wavelength count was reduced by a factor of around 2-3, depending on traffic [133].

In *bi-directional rings*, the possibility of choosing the shortest path between two nodes may improve bandwidth utilisation. However, to avoid collisions between counter-directional packets requires either centralised scheduling, which increases delay, or separation of the two directions, increasing node and/or transmission layer complexity by dictating use of space switches or bandwidth partitioning.

Full *link bandwidth sharing* increases statistical multiplexing gains, as opposed to bandwidth partitioning for waveband concepts or bidirectional networks. To exploit this requires Wavelength Conversion (WC). Whilst slotted operation enables non-blocking Ring Nodes by per-slot reallocation of wavelength used for packet insertion, asynchronous operation requires wavelength conversion also of forwarded packets to avoid internal blocking.

Multicast enables traffic from a single source to be sent to multiple destinations. This forwarding paradigm reduces the total number of packets transmitted, thereby saving link bandwidth. It is particularly interesting for distribution of bandwidth intensive services such as video conferencing, Video on Demand (VoD) and online gaming.

Fig. 6.1 represents a unidirectional ring network, employing multicast and space reuse. Different packet transfer scenarios are distinguished by different line patterns, and wavelength allocation on each link is denoted by λ *: [source] -> [destination(s)]*. For simplicity, only one multicast and

two unicast packet transfers are illustrated, using links with only two wavelengths. We will consider the benefit of WC through an example: Any node may start inserting a packet (here: transfer [E]->[A]), at an idle wavelength. When a packet using the same wavelength arrives (here: multicast transfer [C]->[D,A]), node E must use WC to avoid internal blocking during forwarding. Hence, the only condition for successful packet transfer is having a free wavelength (but not necessarily the same) at each link between source and destination(s).

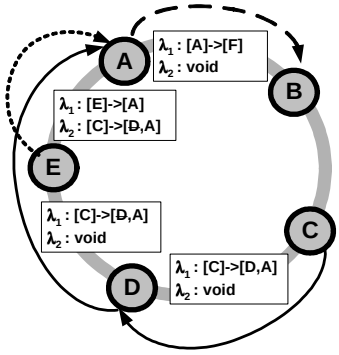

Fig. 6.1. Intra-ring unicast and multicast transfers.

B.2. Ring interconnection and MAN/WAN interface

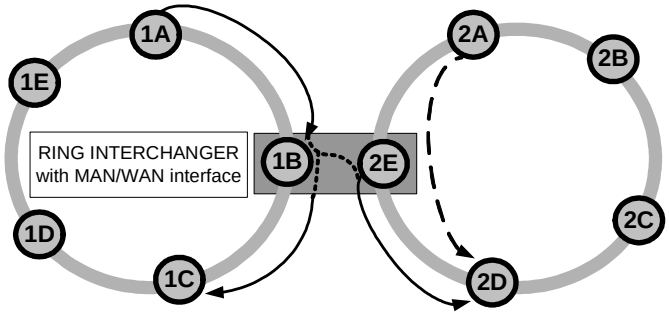

Fig. 6.2. Interchanger enables inter-ring connections.

To increase network size requires both ring interconnection and an interface to the WAN. This can be achieved by a Ring Interchanger with MAN/WAN interface. Fig. 6.2 illustrates how a multicast packet from node A on ring 1 now can reach node D on ring 2, as well as node C on ring 1. The MAN/WAN interface is assumed to include a O/E/O conversion, enabling 3R regeneration and adaptation to new signal format and bitrate. Furthermore, this enables monitoring and policing of WAN ingress and egress traffic, required to realise service level agreements between different operators.

C. Node design

C.1. Design Rationale

Introducing advanced ring features, requires *active node* architectures, in general increasing node complexity. However, space reuse, multicast and full bandwidth sharing decrease the required number of wavelengths per fibre. This reduces WDM transmission layer CAPEX and OPEX (by decreasing size and/or number of transmitters, receivers EDFAs, multiplexers, couplers, performance monitors, dispersion compensators) and it decreases the size of the switches in the nodes, ring interchangers and MAN/WAN interfaces. Hence, active nodes are beneficial whenever these savings outweigh the additional cost of introducing more active components in the nodes.

We propose active node designs that combine WC and switching, using an All-Active Mach Zehnder Interferometer (AA-MZI) and associated tunable lasers as the only active components, forming a Tunable Wavelength Converter (TWC). The TWC configuration is such that it keeps a copy of the input signals, enabling multicast. The high integration level of the AA-MZI enables compact devices.

This study only considers the data plane, and assumes that the devices are configured by an OPS control unit that implements the desired scheduling policy, based on packet control information read on a control channel [25].

C.2. Ring Node Design

For comparison, we show in Fig. 6.3 a) an existing design of an active node without optical WC capability and without waveband separation [133]. The design is intended for slotted operation, and synchronisers will thus be required at some or all ring nodes. The novel Ring Node design in Fig. 6.3 b) is suitable for asynchronous operation and has full WC capability. It works as follows: The TWCs drop all input packets from the ring to the receiver (Rx) array, which selects which packets to send to the access network interface. If the packet should be forwarded on the ring, the TWC laser is tuned to the wavelength to be used at the next link, avoiding internal blocking with existing packet transfers (forwarded or inserted); otherwise it is erased by turning off the laser, or by tuning to a specific “dump” wavelength, as discussed in D.2 Then, all packets to be forwarded are coupled with the inserted packets from the transmitter (Tx) array, which forms the output interface of the access network. It can be realised by fixed lasers and modulators. We compare design component counts in D.4.

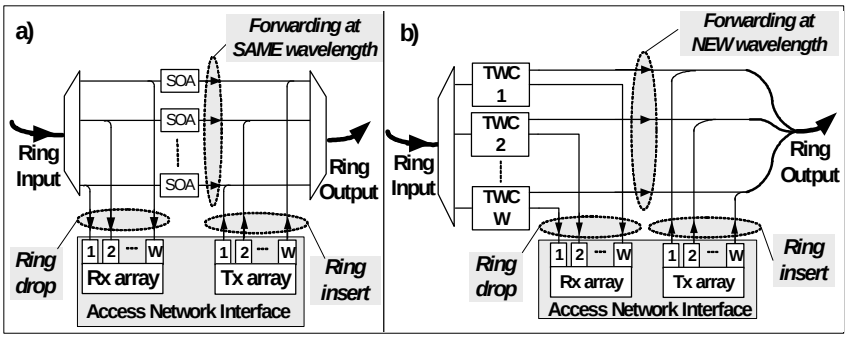

Fig. 6.3. a) Ring Node w/o WC, b) Proposed Ring Node w/ WC.

C.3. Ring Interchanger design

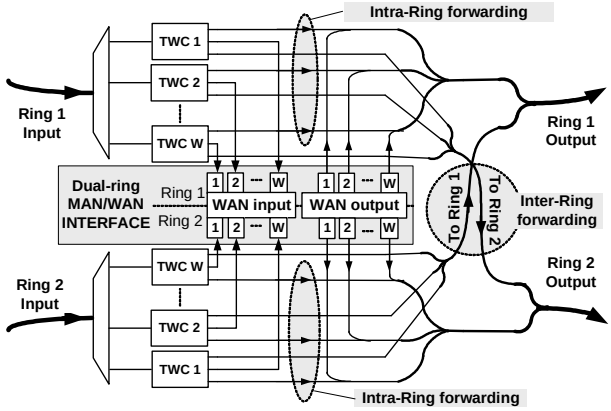

Fig. 6.4. Proposed Ring Interchanger MAN/WAN interface.

The proposed ring interchanger is illustrated in Fig. 6.4. In addition to the drop output and the Intra-Ring packet forwarding output, the TWC should also be able to output a third copy of the input packet at a freely selectable wavelength, for strictly non-blocking Inter-Ring forwarding. Such a novel TWC design is depicted in Fig 6.5 b). Note also that access network interface is replaced by the MAN/WAN interface, but this does not change its optical interface.

D. Experiment and Results

The experiments deal with the operation of the AA-MZI, being the critical part of the node design. We hence emulate the appropriate scenarios for operation both for the ring node and the ring interchanger. An important feature of these designs is that all SOA currents in the AA-MZI are maintained during operation, even for changing functionalities, the AA-MZI control is thus *all-optical*.

D.1. Experimental Set-up

Fig. 6.5 illustrates the experimental set-up. The transmitter module emulating ring input uses tunable external cavity lasers (ECL) and MZI modulators (MOD), driven by 10 Gbit/s NRZ pulse pattern generators (PPG), with PRBS word lengths of 2^7-1. ECL 2 and 3 provide continuous wave (CW) light for the AA-MZI inputs, needed for intra- and inter ring forwarding, respectively. The signal quality in terms of bit error rate (BER) is measured by the preamplified receiver (REC). Polarisation controllers, amplifiers and attenuators, used for optimisation at beginning of experiment are omitted for clarity.

For the Ring Interchanger experiment, depicted in Fig. 6.5 b), a Tunable Bandpass Filter (TBF) was used to select between Inter-Ring signal or Drop signal. In a final Ring Interchanger design, one should instead exploit that the input wavelength is fixed, by replacing the TBF by a Fixed Reflection Filter (FRF). Hence one receives the Ring Drop signal at an added arm of the circulator, C1, and the Inter-Ring signal after the FRF. Note that this prevents using same Inter-Ring output and input wavelength.

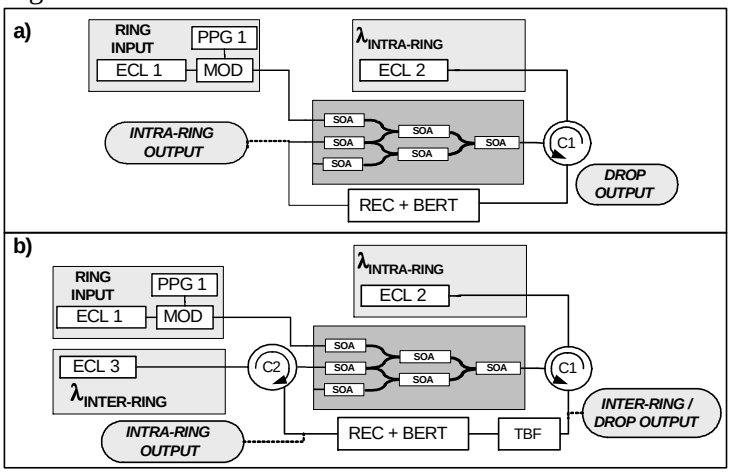

Fig. 6.5. Experimental set-up of a) Ring Node, b) Interchanger.

D.2. Ring Node experiment

In this experiment, we have a ring input signal at 1545 nm, which is received and measured at the Drop output, and that can be forwarded to the Intra-Ring output at a the ECL 2 wavelength, here set to 1550 nm. BER of drop signal is measured both when input data is simultaneously forwarded and not, termed Ring Drop I and Ring Drop II, respectively. The results are expressed by the BER curves and eye diagrams of Fig. 6.6.

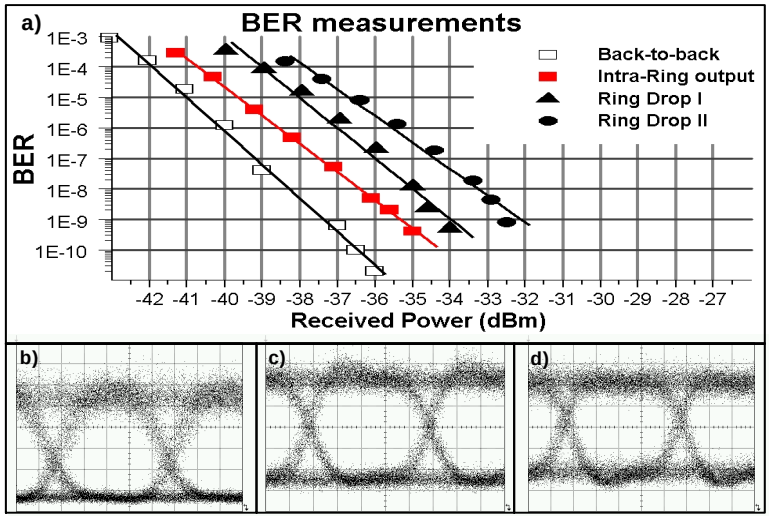

Fig. 6.6. a) BER curves. Eye diagrams of b) Intra-Ring output, c) Ring Drop I and d) Ring Drop II.

Comparing sensitivities at BER of 10^{-9} with back-to-back receiver sensitivity reveals penalties of around 2, 3 and 5 dB, for the Intra-Ring, Ring Drop I and Ring Drop II, respectively. These results indicate that to limit penalty to 3 dB, the design should be improved by always having ECL 1 on. Packets can instead be erased by tuning ECL 1 to a “dump” wavelength, which is subsequently filtered out at the switch output, thus shared by all TWCs.

D.3. Ring Interchanger experiment

Sensitivities, resulting from the Ring Interchanger experiment, of the all three outputs simultaneously on are expressed in Table 6.2, and eye diagrams are depicted in Fig. 6.7 (with x-axis as ground level). Penalties up to 5 dB suggest that regenerators may be needed in the ring. However, all results are obtained with AA-MZIs not originally designed for the applications described here, and signal quality might be improved by optimising the design for this purpose.

Table 6.2. Sensitivities at BER of 10^{-9}.

Output signal	Sensitivity
Drop @ λ_1=1545.0 nm	-36.1 dBm
Intra-Ring output @ λ_1=1547.5 nm	-32.0 dBm
Inter-Ring output @ λ_1=1555.0 nm	-35.4 dBm

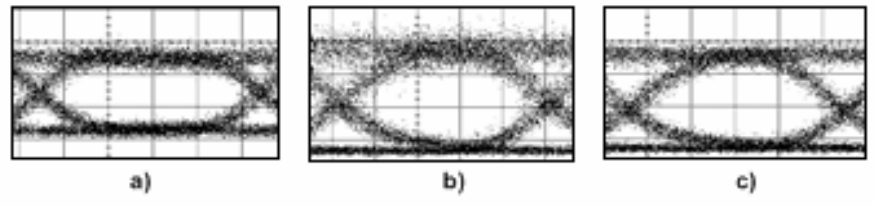

Fig. 6.7. Eye diagrams of a) Ring-Drop, b) Intra-Ring, and c) Inter-Ring.

D.4. Hardware comparison

Table 6.3 sums up the components needed for the active Ring Node w/o WC, active Ring Node w/ WC and the Ring Interchanger, using a component *count (port)* notation. To compare identical capacities, the Ring Interchanger only includes components belonging to one ring.

The Ring Node's compatibility with multicast and asynchronous operation mainly comes at the expense of replacing *W* SOA gates with *W* AA-MZIs and tunable lasers. For the passive components, a total of *W* coupler ports can be removed, but *W* circulators must be added. On the other hand, the design avoids complex packet synchronisers, required in slotted operation. The combined ring interchanger and MAN/WAN interface requires only *W* additional lasers, FRFs, circulators and a *W*:1 coupler, compared to the Ring Node w/ WC.

Table 6.3. Component count of discussed designs.

Comp. Design	Tx	Rx	SOA	AA-MZI	Tun. Las.	Mux	Coupler	FRF	Circ
Ring Node w/o WC	*W*	*W*	*W*	0	0	2 (*W*:1)	2*W* (2:1)	0	0
Ring Node w/ WC	*W*	*W*	0	*W*	*W*	1 (*W*:1)	*W* (2:1) +1 (*W*:1)	0	*W* (3)
Ring Interchanger	*W*	*W*	0	*W*	2W	1 (*W*:1)	*W* (2:1) +2 (*W*:1)	*W*	*W* (3) +*W* (4)

E. Conclusion

The proposed designs enable strictly non-blocking, OPS MAN multi-ring networks in asynchronous operation with WAN interface. Full space reuse, bandwidth sharing and multicast minimises link CAPEX and OPEX. The added functionality is obtained by replacing synchronisers and SOA gates by AA-MZI's and associated tunable lasers. The viability of the concepts were verified by demonstrating the required AA-MZI functionalities. Further studies should focus on logical performance and performance/complexity trade-offs.

6.3. Supporting VLP in OPS Metro Rings

This chapter incorporates the OSA Journal of Optical Networking 2005 article [p23].

Distributed MAC Protocol for Optical Packet Switched Ring Network Supporting Variable Length Packets

M. Nord

Research Center COM, Technical University of Denmark, 345V, DK-2800 Lyngby, Denmark
and Telenor R&D, N-1331 Fornebu, Norway
mn@com.dtu.dk

S. Bjørnstad

Telenor R&D, N-1331 Fornebu, Norway
and Norwegian University of Science and Technology, N-7491, Trondheim, Norway

M. L. Nielsen

Research Center COM, Technical University of Denmark, 345V, DK-2800 Lyngby, Denmark

Abstract. We propose a distributed medium access control protocol for an asynchronous, optical packet switch architecture, suitable for an efficient, scalable and flexible interconnected ring network. We compare the complexity of our proposal with existing techniques to support variable length packets and with node architectures that enable spatial wavelength reuse. Simulations quantify the throughput increase enabled by the MAC protocol, and show that moderate hardware resources are sufficient to offer low Packet Loss Rates with low maximum delay and delay jitter. Finally, we show that the network efficiently supports bursty traffic.

A. Introduction

Optical Packet Switching (OPS) is an excellent candidate for the future Metropolitan Area Network (MAN), which will be much more dynamic and demanding than today's networks [131]. The MAN is a critical network segment, subject to an emerging gap between the high-speed local networks and the very high-speed backbone networks, termed the "metro gap" [132]. The Optical Packet Switched Ring Network (OPSRN) architecture may overcome this gap, when the following design criteria are respected [132]:

- *Efficient* use of wavelength resources through spatial wavelength reuse: The destination nodes remove packets from the ring, freeing bandwidth for other transfers, thus increasing throughput.
- A *scalable* network: This requires ease of upgradeability and that the node count is independent of the WDM channel count, which is closely connected to the node architecture.
- A *flexible* network: to support varying traffic loads and packet formats, in particular support of Variable Length Packets (VLP), which increases the range of acceptable protocols and applications that can be supported by the network.

To meet these requirements, we propose an *Asynchronous Insertion Priority Scheduling with Insertion Threshold* (*AIPSwIT*) MAC protocol and associated node architecture. It is a flexible design that enables spatial wavelength reuse for VLP. The remainder of this article is organised as follows: Section B describes the network architecture, reviews existing VLP techniques and introduces our *AIPSwIT* MAC protocol. Section C compares the complexity of the architecture with existing proposals. Section D analyses network performance as a function of hardware resources. Section E concludes the article.

B. Optical Packet Switched Ring Network Design

B.1. Network overview

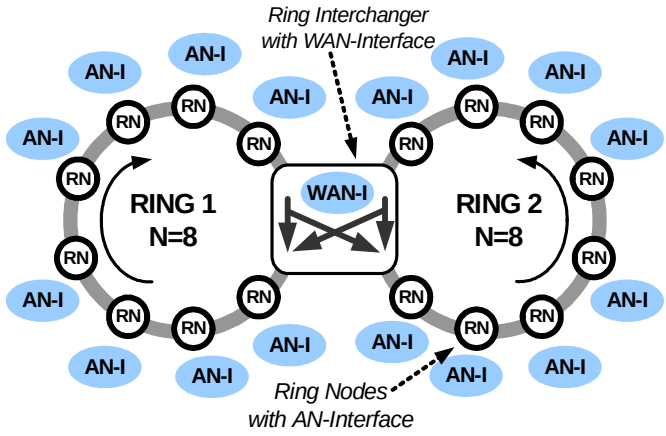

Fig. 6.8. The studied MAN architecture: two interconnected rings with AN- and WAN interfaces.

Fig. 6.8 depicts the network architecture assumed in this study: a MAN network with two unidirectional WDM rings Each ring has 8 Ring Nodes (N=8), interconnected by a Ring Interchanger. Each Ring Node serves an Access Network, connected by an electrical Access Network Interface (AN-I). The Ring Interchanger enables communication across a WAN through its electrical Wide Area Network Interface (WAN-I). The ring network signal path is purely optical. O/E conversions are needed only at network interfaces, enabling electrical ingress buffers with random access times for inserting traffic on the ring, and facilitating adaptation between AN and WAN signal protocol, format and bitrate.

The use of two rings, instead of a single ring with 2N Ring Nodes and a node for the WAN-I, reduces mean hop distance, H, from 8.5 to 6.61 for uniform traffic. This reduces required network link bandwidth and the mean end-to-end delay. The benefit increases when the ratio of traffic going to the WAN-I increases, e.g. H=5.4 with the traffic matrix assumed in Section D.2. However, the ring interchanger needs space switching functionality. This is not needed in Ring Nodes, which either insert packets from their input interface, forward packets on the ring, or drop packets to their output interface.

B.2. Distributed MAC Protocols supporting Slot Reuse and VLP

In OPSRNs, contention can be avoided whilst supporting VLP, by using a *reservation protocol* [134], or a *Multi-Token Interarrival Time (MTIT)* protocol [135]. *Reservation protocols* require significant resources to realise the communication between the nodes. Furthermore, they add a delay, equal to the sum of the ring's Round Trip Time (RTT) and the reservation processing delay, to the packet's propagation delay. The RTT is ~100-1000 µs for rings with lengths of 20–200 km. This prevents "immediate access", which we consider to be achieved when the ingress buffer delay is well below ring RTT. On the other hand, the *MTIT* proposal has low access delay, but throughput is reduced by not employing spatial wavelength reuse [132].

To enable immediate access, spatial wavelength reuse and low complexity scheduling, we study *distributed* MAC protocols. The main candidates are "empty-slot" protocols, which switch fixed-length packets synchronously [132]. By giving priority to packets in-transit, contention is resolved by only inserting packets from the ingress buffer when there are free slots on a suitable wavelength on the ring. This wavelength availability control information can be obtained from either in-band packet headers, or from an out-of-band control channel. In this study we do not specify any particular method, since it is a design choice that does not intrinsically impact performance. Note that in both cases, *processing FDL* is required to delay the optical data packets whilst the control information is processed.

These MAC protocols are typically *a posteriori* schemes, where each of the node's accessible ring wavelengths are associated with a Virtual Output Queue (VOQ) in the ingress buffer [134]. The wavelength availability and scheduling policy govern which VOQ to insert packets from. This increases the node's chance of inserting packets on available slots, and it avoids Head Of Line (HOL) blocking. However, such schemes require faster scheduling than *a priori* schemes, which selects a packet to insert before knowing wavelength availability [132]. Different techniques, listed below, have been proposed to extend empty-slot protocols to support VLP, and these are compared with *AIPSwIT* in Section B.3:

- *Preemption*: The scheduler starts inserting a packet on an empty slot, but aborts the insertion attempt if the packet is not fully inserted before the channel contains a non-empty slot [136]. This calls for use of optical gates in the transmission path to erase the pre-empted packet at the next downstream node, to avoid bandwidth wastage

[136]. Moreover, the scheme is unfair to long packets, since nodes with long packets have less access to the ring bandwidth.

- *Segmentation And Reassembly* (*SAR*): Segments packets with length above a slot into fixed length segments, and reassembles them at the destination. However, segments from a packet may be interleaved by segments from other packets [137], and each receiver should have a VOQ per source node to reassemble the segments [138]. Moreover, since each segment has a header, or MAC frame, *SAR* suffers from increased overhead [136]. Again, the scheme is unfair to long packets, since nodes with long packets have higher overhead.
- *SAR-On Demand* (*SAR-OD*): Combines pre-emption and *SAR* to reduce overhead and packet length unfairness, by only segmenting a packet when a non-empty slot interrupts its insertion [138]. Still, the nodes must handle the complex *SAR* procedure which ideally should be avoided [134, 139].
- *Multiple Slot Sizes* (*MSS*): Operates a slotted ring with multiple slot sizes per wavelength, suitable to the expected ingress traffic [136]. However, performance suffers from HOL blocking (within a destination VOQ), when the first packet in a VOQ does not fit the size of the free slot. *MSS* aggravates the problem of global slot synchronisation and detection of slot boundaries [134]. An efficient implementation depends on an accurate prediction of the packet length distribution, and since the RTT is the upper bound of all slot sizes on a wavelength, MSS has a limited packet length well below the RTT.
- *Look-ahead:* Increases scheduling horizon to equal the maximum packet length, by increasing the *processing FDL* length correspondingly. Input packets are sorted into VOQs by a pre-classification scheme [139], based on their packet length. However, acceptable *processing FDL* length and scheduling complexity limits maximum packet length.
- *FDL based register insertion technique (FDL Reg. Ins.):* Gives priority to inserted packets, using a set of switchable FDLs to delay in-transit packets during contention [137]. However, the coarse granularity of the FDLs decreases bandwidth efficiency [134]. The bulkyness of FDLs may limit the maximum packet length, and use of optical switches in the set makes it a complex component.

Note that slotted operation prevents the MAC to completely fill up all accessed slots when the input packets are VLP. Hence, the Slot Filling Ratio (SFR) is sub-optimal, which reduces bandwidth efficiency.

B.3. Proposal for MAC protocol with Insertion Priority and Insertion Threshold

We here describe our *Asynchronous Insertion Priority Scheduling with Insertion Threshold* (*AIPSwIT*) MAC protocol. It operates asynchronously and gives priority to packets currently being inserted, thereby enabling "unlimited" VLP length without any discrimination of longer packets. The MAC protocol exploits the wavelength domain for contention resolution, similar to the principle used in mesh networks, as studied by [95]. To this end, it converts packets to be forwarded, which find their own channel occupied, to other free wavelengths on the next hop. Simultaneously, the MAC protocol optimises the probability of finding such a free wavelength, by only inserting packets when the number of free wavelengths on the next-hop link is above an insertion threshold.

Table 6.4. Comparison of empty-slot MAC protocols supporting VLP and *AIPSwIT* MAC protocol.

MAC	"Unlimited" VLP length	Packet length fairness	MAC overhead	SFR	SAR complexity	Insertion complexity	Forward Complexity
SAR	Yes	No	High	Sub-Opt.	High	Low	Low
PreEmption	Yes	No	Low	Sub-Opt.	N/A	Low	High (SOA)
SAR-OD	Yes	No	Medium/Low	Sub-Opt.	High	Low	Low
MSS	No	Possible	Low	Sub-Opt.	N/A	Medium	Low
Look Ahead	No	Possible	Low	Sub-Opt	N/A	High	Low
FDL Reg. Ins.	No	Yes	Low	Sub-Opt.	N/A	Low	High (FDL)
AIPSwIT	Yes	Yes	Low	N/A	N/A	Low	High (TWC)

Regarding the performance, the insertion threshold principle has earlier proven beneficial for output scheduling from electrical contention resolution buffers in OPS *mesh* networks [140]. We show in Section D how it greatly increases network throughput. The scheduling complexity is low, since the allocation of wavelength both for forwarding and insertion of packets is based on first-fit searches in the scheduler's wavelength allocation table, and since the inserting decision is based on a simple counter - threshold comparison. Moreover, the *AIPSwIT* MAC can operate *a priori* without HOL blocking, since the packet to be inserted is strictly FIFO based, and does not depend on which wavelengths that are free. This alleviates processing time requirements. Finally, note that the insertion threshold can also be employed to ensure fairness, which is the focus of an ongoing study, beyond the scope of this article.

Table 6.4 sums up the qualitative comparison of the empty-slot VLP adaptations and *AIPSwIT*. The former suffer from a combination of: *i)* not supporting unlimited VLP length, *ii)* discrimination of long packets, *iii)* sub-optimal SFR, *iv)* increased MAC overhead, and, *v)* increased electrical or optical hardware complexity for insertion, forwarding or reception of packets. In contrast, *AIPSwIT* avoids *i)-iv)*, but does require a TWC for forwarding. However, as we will see in Section C, such an active device is needed also to enable spatial wavelength reuse in a flexible network design, and we combine these two functionalities to reduce component count. This partly compensates the drawback of using a TWC.

C. Node architectures

C.1. Generic Ring Node Architectures

The following discussion on node architecture uses the conventional: FT^{i}- TT^{j}- FR^{m}- TR^{n} notation scheme [132], meaning Fixed Transmitter, Tunable Transmitter, Fixed Receiver, Tunable Receiver. The indices denote the number of each type used in each Ring Node, omitted only when the index is '1'.

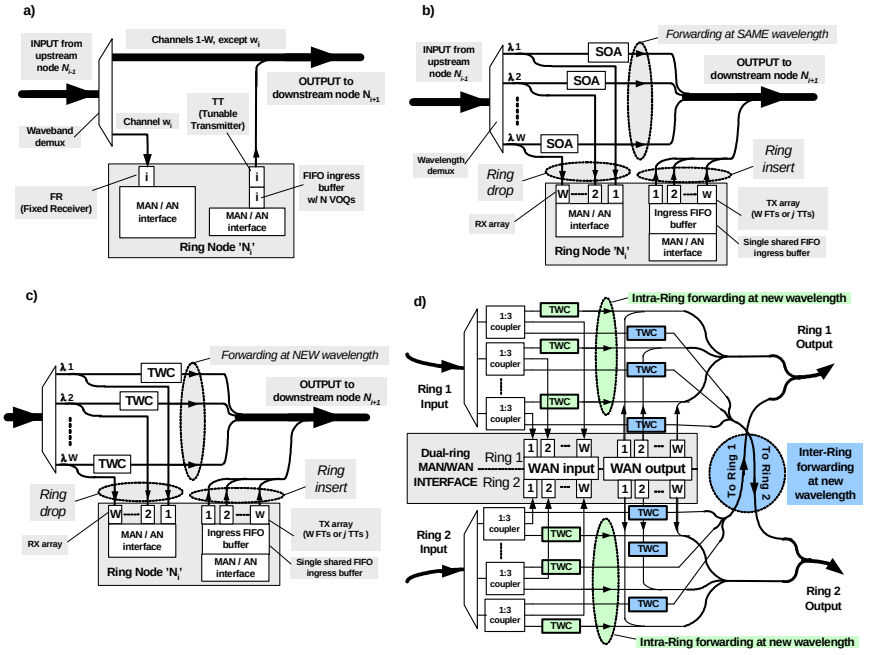

Fig. 6.9. Generic node designs of: a) TT-FR Ring Node, b) SOA gate based FT^{W}-FR^{W} Ring Node, c) TWC based FT^{W}-FR^{W} Ring Node, d) Proposed Ring Interchanger.

Many empty-slot protocols assume a TT-FR node architecture with the number of WDM wavelengths, *W*, being equal to the number of nodes, *N* [132], such as [137, 141]. This enables simple optical demultiplexing hardware to receive the channel, since all nodes are the termination point of one WDM wavelength. The resulting generic node design is illustrated in Fig. 6.9 a). The architecture enables spatial wavelength reuse, since all packets are removed from the ring at their destination, when the wavelength is terminated in the fixed receiver. However, the constraint on the relation between network channels and number of nodes prevents network scalability. Furthermore, the TT-FR architecture only enables the

node to insert packets on one channel at the time, and to receive packets on a single, fixed channel. To insert a packet, the scheduler tunes the TT to the wavelength of the (non-empty) VOQ associated with the wavelength of a free slot, according to the MAC scheduling policy. However, if the VOQ of all free slots are empty, the node cannot insert any packet, even though other VOQs may contain packets. This static architecture thus limits bandwidth sharing, thereby limiting network throughput, in particular when the traffic matrix changes dynamically. E.g. if the relative traffic rate to any node increases, the wavelength used for reception at this node becomes oversubscribed, even though other wavelengths will be undersubscribed. Another drawback is the fairness problem of *node starvation*, i.e. the throughput between sources and destinations is lower than average for source nodes that are close to the destination node [132].

A flexible network has improved bandwidth sharing, which calls for accessibility of reading data from more wavelengths, as well as increased capacity of inserting data. The optimum flexibility is obtained for FT^{W}-FR^{W} architectures, in which any node can use any wavelength to communicate with any destination. Alternatively, a flexibility trade-off is obtained for $TT^{j<W}$-$TR^{n<W}$ architectures, possibly using waveband approaches [133]. To avoid TRs, i.e. tunable optical filters, $TT^{j<W}$-FR^{W} architectures are often preferred, such as [139, 142]. As illustrated in Fig. 6.9 b), each node receives a copy of all packets arriving on their upstream link, which makes it compatible with multicast [132]. To enable spatial wavelength reuse, the scheduler chooses which packets to forward to the downstream node, by using either W 2x2 switches [142], or W sets of 1:2 coupler with an optical gate in the forwarding path.

As described in Section B.3, our proposed *AIPSwIT* MAC protocol gives priority to packets under insertion. Contention resolution then requires wavelength conversion of the packet to be forwarded to a currently unused wavelength. To save components, we propose to combine the wavelength conversion with the optical gate functionality required for spatial wavelength reuse. The TWC then either shifts the wavelength of the packet to be forwarded, or remove the packet from the ring, as detailed in Section C.3.

Most slotted ring networks assume slot alignment, which requires either complex optical synchronisers at (at least in some) Ring Nodes, or careful dispersion compensation [138], depending on the slots' guard band. For multiple rings, synchronisers are at the very least needed at the ring interconnection, since different fibres may experience different environmental conditions, known to impact propagation delay. In

addition, when using a separate control channel, it should be synchronised to the corresponding data slots. This can be done electronically, simply by adjusting the timing of the control channel, provided that the data slots are aligned [136]. A benefit of the asynchronous operation in *AIPSwIT* is that it does not need any alignment between data, only between data and control. This enables electronic synchronisation when using a control channel, and even removes the problem altogether when using packet headers.

Table 6.5 compares these three node architectures. Increased flexibility comes at the expense of use of active components in the optical path, and an increased TT and FR count per node. Compared to both Slotted designs, our design eases synchronisation, and inherently supports VLP. However, the equally bandwidth-flexible slotted design uses SOA gates instead of TWCs. Hence, the attractiveness of *AIPSwIT* increases with decreased TWC/SOA cost ratio and with increased effect of the potential drawbacks of the VLP adaptations, summed up in Table 6.4

Table 6.5. Comparison of designs that allow spatial wavelength reuse. (Component count is given per Ring Node).

	VLP support	Bandwidth Flexibility	# TTs	# FRs	# Active Devices for Forwarding	Slot-by-Slot Synchr-onisation	Multicast Compatible?
Slotted Inflexible TT-FR	Not inherent	Low	1	1	N/A	Required	No
Slotted Flexible (TT^j-FR^W)	Not inherent	High	$1 \le j < W$	W	W SOAs	Required	Yes
AIPSwIT (TT^j-FR^W)	Inherent	High	$1 \le j < W$	W	W TWCs	Not Required	Yes

C.2. Ring Interchanger Design

The Ring Interchanger should switch inter-ring packets, in addition to forwarding of intra-ring packets, similar to the Ring Nodes. Hence, blocking may occur in the Ring Interchanger, when a packet should be forwarded to a ring which has no free wavelengths. However, the *AIPSwIT* wavelength insertion threshold on packets from the WAN increases the probability that all forwarded packets can find a free wavelength also at the Ring Interchanger output. Fig. 6.9 d) illustrates the generic ring interchanger design.

C.3. Realisation issues

Both the Ring Node and the Ring Interchanger architectures can be realised using conventional components, such as couplers and TWCs, as illustrated in Fig. 6.9 c) and d). TWCs support bitrates of 40 Gbit/s, as demonstrated in [p5], and TWCs are soon expected to be commercially available [131]. Combining TWC and optical gate functionality,

discussed in Section C.1, calls for possibility of erasing packets in the TWC. This is feasible by either: *i)* converting it to a drop-wavelength (subsequently filtered out using a fixed filter at the node output interface), *ii)* turning off the laser probe signal, or *iii)* modulating the SOA bias current (as in a SOA gate).

Alternatively, the optical functionalities required in the Ring Node and Ring Interchanger, can be realised in a more radical design, based on an All-Active Mach Zehnder Interferometer (AA-MZI) that combines switching and wavelength conversion, demonstrated in [p14]. This avoids one coupler in the Ring Node design, and enables use of a single TWC per channel in the Ring Interchanger design, potentially reducing overall component complexity. The choice of which design is beneficial does not impact the logical behaviour of the network, and is beyond the scope of this study.

D. Network Performance

D.1. Simulation Parameters

Network performance depends both on input traffic and node dimensioning. Regarding the input traffic, we vary both the traffic load and arrival statistics. Regarding the node dimensioning, we first evaluate the FT^W-FR^W architecture. We then evaluate the $TT^{j<W}$-FR^W architecture, including impact of limited buffer resources. To reduce buffer- and scheduler complexity, we model the ingress buffer as one large FIFO buffer storing up to *B* packets. When the ingress buffer is full, the first packet is discarded. Table 6.6 sums up main simulation parameters, further detailed below. The mean packet duration is 1 μs. Since the packet format is technology- and protocol dependent, we do not take packet overhead into consideration.

Table 6.6. Main MAC, performance and hardware parameters.

Symbol	Parameters
***N*=8**	Number of Ring Nodes in each of the two rings.
***W*=32**	WDM channel count, each channel is assumed to operate at 10 Gbit/s.
***H*=5.4**	Mean hop distance in the offered traffic matrix (expressed in number of links).
$B_{LINK,\ NW}$**=5.76 Tbit/s**	Total bandwidth of all links ($B_{LINK,\ NW}$=2 x (*N*+1) x *W* x 10 Gbit/s).
$Load_{ABSOLUTE}$	Total traffic (measured in bits) offered by ANs and WAN per s.
$Load_{NORMALISED}$	$LOAD_{ABSOLUTE}$ normalised to $BW_{LINK,NW}$, taking mean hop distance into account.
RT	Relative Throughput: Ratio of successfully received and offered bits per s.
τ	Mean ingress buffer delay of received packet.
$W_{FREE,AN}$	*AIPSwIT* MAC free wavelength insertion threshold, used in the Ring Nodes.
$W_{FREE,WAN}$	*AIPSwIT* MAC free wavelength insertion threshold, used in the Ring Interchanger.
j_{AN}	Tunable laser count of the AN-I in the Ring Node, (cf. the TT^j notation).
j_{WAN}	Tunable laser count per WAN-I in the Ring Interchanger (cf. the TT^j notation).
B	Max number of packets in the ring node ingress buffer.
D	Max delay of packets in ingress buffers. Corresponds to max packet inter-arrival jitter.

D.2. Traffic Matrix

Of the total input traffic, 20 % goes from the WAN to the 2N ANs, and an equal amount goes from the 2*N* ANs to the WAN. The remaining 60 % is AN-to-AN traffic, of which 2/3 is intra-ring traffic and 1/3 is inter-ring traffic. The source- and destination node is uniformly distributed for all these traffic types, and the traffic is balanced, i.e. evenly distributed over all links, with *H*=5.4.

We first model the ingress packet stream from the ANs and from the WAN by a Poisson arrival process, with exponentially distributed packet lengths. Whilst the Poisson arrival model may be representative for the highly aggregated traffic in the core network [130, 132], the traffic closer to the edge is expected to be more self-similar [132]. We assess this scenario, by an approximation of self-similar AN input traffic, termed "Bursty" in the following. It is modelled according to [143], using independent traffic generators, generating packet arrivals according to a Pareto distribution with Hurst parameter of 0.8.

Assuming 10 Gbit/s channel rates in our 18 link network with W=32, the total network link bandwidth, $B_{LINK,NW}$ is 5.76 Tbit/s. $Load_{ABSOLUTE}$ is the total amount of data that the ANs and WAN offer to the network per second, whilst $Load_{NORMALISED}$ represents a normalisation of the offered bandwidth to total link bandwidth, taking the mean hop distance, into account (6.1).

$$Load_{NORMALISED}=Load_{ABSOLUTE} \text{ x } H / B_{LINK,NW} = Load_{ABSOLUTE} \text{ x } 5.4/ 5760 \text{ Gbit/s} \qquad (6.1)$$

D.3. Performance parameters

The network operator is interested in operating at a high load, to increase revenues. The network users are primarily interested in the *PLR*, delay and jitter. The required level of each parameter depends on the application. For the purpose of this study, we focus on achieving a mean network *PLR* of 10^{-2}, which is similar to [115], deeming it sufficient for acceptable TCP throughput. However, we also show that a *PLR* of ~10^{-3} and below can be reached for relatively small load reductions. Similar to many network studies we evaluate throughput vs. load. We use the Relative Throughput, *RT*, which is a measure of the fraction of input traffic that is received at its destination. In the steady-state, it is related to the *PLR* by *PLR*=1-*RT*, hence *RT* should be above 0.99 to respect the *PLR* requirements.

τ is a measure of the ingress buffers' contribution to the mean end-to-end delay of successfully received packets. This comes in addition to the propagation delay, which depends on the hop distance between source and destination node. Each fibre link is modelled as 5 km long, thus giving a propagation delay of 25 µs per link, which gives a single ring RTT of 225 µs. Taking *H* into account gives a mean propagation delay of 135 µs. As discussed in Section B.2, to obtain "immediate access", interpreted as a delay significantly lower than that of protocols imposing an RTT access delay, we should have τ <<225 µs. The propagation delay will then be the dominant part of the total delay, but it is constant for packets belonging to the same stream. On the other hand, τ varies from

packet to packet. Many network applications and protocols are better served with a bounded packet interarrival jitter [p22]. To achieve a maximum jitter equal to *D*, we impose *D* as the maximum time in the buffer. Older packets are discarded.

D.4. Simulation Results

We use discrete event-driven simulations in OPNET to evaluate the network. Our simulations confirmed that the probability of blocking is independent of the packet length. Results are given with 95% confidence intervals, calculated using 10 independent simulations runs with different random generation seeds, ignoring the transitional period, and using results from the steady-state period. A parametric scan of different $W_{FREE,WAN}$ and $W_{FREE,AN}$ enables us to identify *optimum insertion thresholds* to maximise *RT*, respecting the steady-state requirement and obtaining a low τ. These insertion thresholds balance risk of blocking with under-utilisation of the ring bandwidth. The lower the load, the more conservative can the insertion threshold be without filling up the buffers, which in turn reduces blocking and increases *RT*.

Fig. 6.10 shows the *RT* (lid lines, left axis) and τ (stippled lines, right axis) vs. load, for both Poisson- and Bursty input traffic, using the optimised insertion thresholds. To achieve *RT* above 0.99, $Load_{NORMALISED}$ can be up to 0.78. The Poisson scenario obtains higher *RT* than bursty traffic. For both traffic types, τ =0-6 µs, which is orders of magnitude below the RTT, and comparable to e.g. [142].

Fig. 6.10 also shows *RT* when not applying an insertion threshold (dotted lines, left axis), i.e. when the MAC inserts packets on the ring as soon as there is a free wavelength. Compared to this case, the MAC protocol using optimum insertion thresholds enables roughly a 10-50% increase in acceptable load to obtain the same *RT*. The gain is highest for *RT* close to unity. For lower *RT*, significant blocking occurs per hop, which reduces mean hop distance, thus reducing the observable gain. Note that without insertion thresholds, we have τ <<1 µs, and this parameter is thus omitted from Fig. 6.10 in this case.

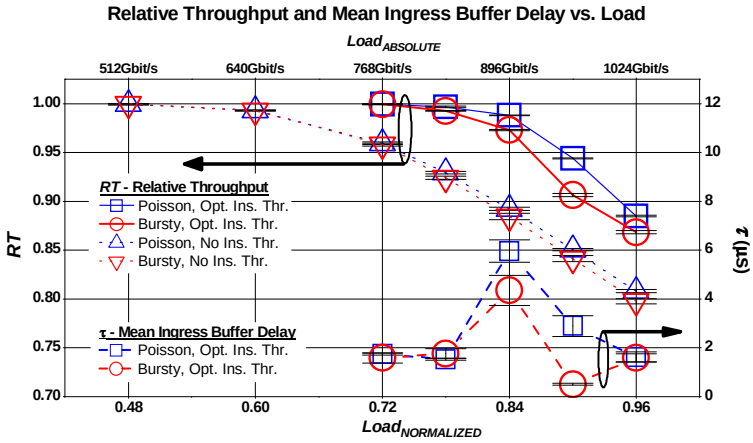

Fig. 6.10. *RT* and τ for optimised MAC insertion threshold ("Opt. Ins. Thr."), and w/o MAC insertion threshold ("No Ins. Thr."), for Poisson and Bursty traffic.

D.5. Hardware savings for TT^{j}-FR^{W} architectures

In this section we maintain $Load_{NORMALISED}$=0.78 and 0.84, since they enable *RT* around 0.99, thus *PLR* around 10^{-2}. When reducing TT count, as a minimum there must be enough TTs to insert the load from each node continuously. However, to enable a more flexible buffer insertion policy, we relax this minimum TT count by increasing it by 50% and round to the closest integer. These values are given in Table 6.7, both for j_{AN} and j_{WAN}, as a function of $Load_{NORMALISED}$. The overall transmitter count is then reduced from 2x(NxW+W) to 2x(Nxj_{AN}+j_{WAN}), which is almost an 80% reduction. Assuming that commercial TTs are less than twice as expensive as commercial FTs [136], this has the potential to decrease transmitter cost by ~60%.

Using these transmitter counts, we study *RT* versus *B*, shown in Fig. 6.11. Note that the buffer size of the Ring Interchanger is taken as *2B*, since it has higher load than the Ring Nodes. At the same time we impose D=16 μs and D=64 μs to bound jitter. For Poisson input traffic, D has only negligible influence for D≥16 μs, and is therefore maintained at D=16 μs. We observe that *RT* decreases when going from $Load_{NORMALISED}$ of 0.78 to 0.84. There is a performance-complexity trade-off between *RT* and *B*, giving the network designer some flexibility. For $Load_{NORMALISED}$=0.84, *RT* converges below 0.99. For $Load_{NORMALISED}$=0.78, *RT* increases from of 0.975 to 0.997, by increasing *B* from 16 to 64, where it reaches its asymptotic value.

RT of Bursty traffic was found to increase with *D*, and *D*=64 μs is therefore included. Again, there is a trade-off between *RT* and load, and between *RT* and *B*. However, we observe that Bursty traffic requires larger *B* to reach the same *RT* as Poisson traffic, and that increasing *D* makes the *RT* be closer to Poisson traffic, when sufficiently increasing *B*. This confirms [132], in that that self-similar traffic is more demanding than Poisson-like traffic, but it also indicates that the effects are not detrimental, since they can be overcome by moderately reducing the load, or by increasing buffer dimensions.

Table 6.7. Transmitter reduction enabled by the TT^j architecture.

$Load_{NORMALISED}$	j_{AN}	j_{WAN}	Overall TX count reduction
0.78	6	12	79.2%
0.84	7	13	76.0%

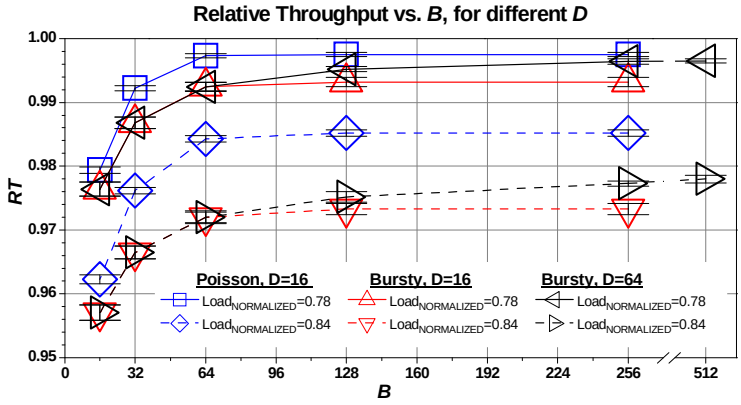

Fig. 6.11. *RT* vs. *B* for different $Load_{NORMALISED}$ and different *D*. Note the break along the x-axis.

E. Conclusion

We have proposed a distributed MAC protocol that is efficient, scalable and flexible. Similar to flexible slotted designs with spatial wavelength reuse, the associated node architecture requires active devices for forwarding. The added complexity of using a TWC instead of an SOA gate may be compensated by the inherent support of Variable Length Packets (VLP), since it does not have any other of the discussed drawbacks of the empty-slot VLP adaptations. Furthermore, being a distributed control, based on *a priori* scheduling, the scheduling speed requirement is relaxed compared to *a posteriori* scheduling.

In a dual ring network with 32 WDM channels operated at 10 Gbit/s, an absolute load of 832 Gbit/s can be supported with a *RT* above 0.99. This is obtained both for an FT^{W}–FR^{W} architecture, as well as for an $TT^{j<W}$–FR^{W} architecture with a transmitter reduction of almost 80 %, when using buffers supporting up to 32 packets per node with a bounded jitter of only 16 μs. It is hence an immediate access protocol. For more bursty traffic, the network obtains the same *RT* by doubling the buffer size. Even higher throughputs are reached when either increasing buffer size, up to a load- and jitter defined limit, or when decreasing load.

Given the above discussed properties, we believe this network architecture to be a promising candidate for an OPS MAN, and current work investigates fairness support for different traffic matrices.

ACKNOWLEDGMENTS

The authors would like to thank H. Øverby, Norwegian University of Science and Technology, and V. B. Iversen, Research Center COM, Technical University of Denmark, for fruitful discussions during this study. Also thanks to the reviewers, whose suggestions helped us improve this article.

6.4. Supporting Fairness in OPS Metro Rings

This chapter incorporates an article submitted to Elsevier Journal of Optical Switching and Networking [p26], (revised version resubmitted).

Fairness Support in Flexible Asynchronous Optical Packet Switched Ring Networks

M. Nord [1, 2]

Mail: ***mn@com.dtu.dk****, Tel: +45 45256604.*
(1) Research Center COM, Technical University of Denmark, DTU-B 345V, DK-2800 Lyngby, Denmark
(2) Telenor R&D, N-1331, Fornebu, Norway

Abstract. This article studies performance of an asynchronous Optical Packet Switched Ring Network (OPSRN) that uses the recently proposed *AIPSwIT* MAC protocol to efficiently support the Metropolitan Area Network. This study addresses relative throughput fairness between source-destination pairs, whilst guaranteeing maximum delay and delay jitter. Event-driven simulations are used to assess the OPSRN performance. First, its efficiency is highlighted, by comparing its throughput with that of a Static Wavelength Routed Optical Network (SWRON) for a uniform traffic scenario. Then, its flexibility is emphasised, by showing that it can also efficiently support non-uniform traffic scenarios in a fair manner. We conclude that this fairness technique for OPSRNs is particularly promising for non-uniform traffic, since it increases acceptable network load, without violating specific, source-destination throughput levels.

A. Introduction

Optical Packet Switching (OPS) is an excellent candidate for the future Metropolitan Area Network (MAN), which will be much more dynamic and demanding than today's networks [131]. Currently deployed Synchronous Digital Hierarchy (SDH) technologies are not suitable to accommodate the increasing amount of data traffic in the Metropolitan Area Network (MAN) [132]. The MAN bandwidth bottleneck prevents high-speed clients and service providers in ANs from tapping into the vast amounts of bandwidth available in backbone networks.

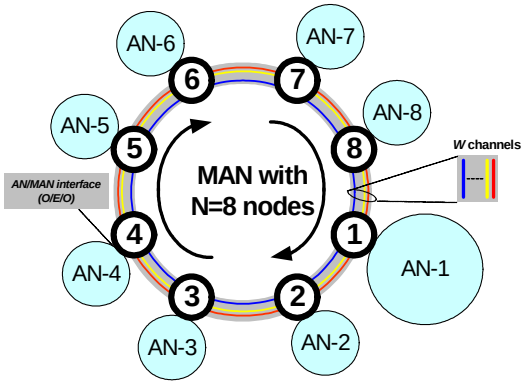

Fig. 6.12. A WDM-based OPSRN with 8 nodes (*N*=8), each serving an Access Network (AN).

Optical Packet Switched Ring Networks (OPSRNs) constitute a long-term solution to bridge this widening "metro-gap" [132]. They benefit from attractive properties of optical and electronic technologies, by combining flexible electrical buffers to store packets from the AN in ingress buffers until insertion on the ring is desirable, with cost-effective WDM transmission. Ideally, OPSRNs should have the following features:

- *Bandwidth-efficiency*: High loads can be supported by combining low MAC overhead with *space reuse*, i.e. that destination nodes remove packets from the ring [132].
- *Scalability*: Combine node architectures that decouple ring node- and WDM channel count, with a distributed MAC protocol that has less delay and complexity than centralised scheduling [137].
- *Flexibility*: Use a flexible node architecture to increase the range of supported protocols, applications and traffic scenarios, by supporting Variable Length Packets (VLPs), Quality of Service (QoS) and fairness [132].

It is at present not clear what the QoS requirements will be at the time of OPSRN deployment. In this study we address how to support the part of the traffic that tolerates a relaxed *statistical QoS,* with a certain PLR and delay-jitter. Very loss sensitive traffic and/or Constant Bit Rate (CBR) traffic, can be better supported by a circuit switched network, which enables *guaranteed QoS*. The potential for the statistical QoS paradigm is economic use of resources by high bandwidth efficiency, in particular for dynamic traffic. In order to increase the amount of traffic that can be supported by the OPSRN, we study an "immediate access" OPSRN with an ingress delay below the ~0.2 ms ring Round Trip Time, (RTT). This article extends earlier research on a distributed MAC protocol, termed *Asynchronous Insertion Priority Scheduling with Insertion Threshold*

(*AIPSwIT*) [p23], to include *fairness*. Fairness is a prerequisite to provide network-consistent QoS, but we do not address QoS differentiation. To highlight the fundamental properties of ring networks, we investigate a single unidirectional ring [133, 136, 139, 141, 144, 145]. This is an important step before expanding the study to include e.g. bidirectionality [138], increased connectivity, or interconnection of rings [p23, 146], which all may improve network resilience and scalability.

The remainder of this article is organised as follows. Section B details the node architecture and MAC protocol, and compares them with existing solutions featuring space reuse, VLP support and fairness mechanisms. Section C formalises the simulation parameters. Section D compares OPSRN performance with that of an SWRON for uniform traffic, and Section E studies OPSRN performance for non-uniform traffic scenarios. Section F concludes the article.

B. The AIPSwIT Node Architecture and MAC Protocol

B.1. Blocking in ring networks

To facilitate the analysis throughout this article, we list the different types of blocking that may occur in a ring network. As the list shows, blocking during contention between in-transit packets (packets on the ring) and packets under insertion depends on whether the MAC protocol gives priority to the former or the latter packet type, denoted *Tra-PRI* and *Ins-PRI*, respectively.

- "Buffer Discards" (*BD*), take place when the ingress buffer of an OPSRN node suffers from buffer overflow, or when the buffer delay timer of a packet expires.
- "Reception Blocking" (*RB*), takes place when a node does not have capacity to receive all time-overlapping packets on the ring, destined to this node.
- Contention between in-transit and inserted packets, results in either:
  - "Insertion Blocking" (*IB*), for *Tra-PRI* MAC protocols.
  - "Forwarding Blocking" (*FB*), for *Ins-PRI* MAC protocol.

B.2. Node Architecture

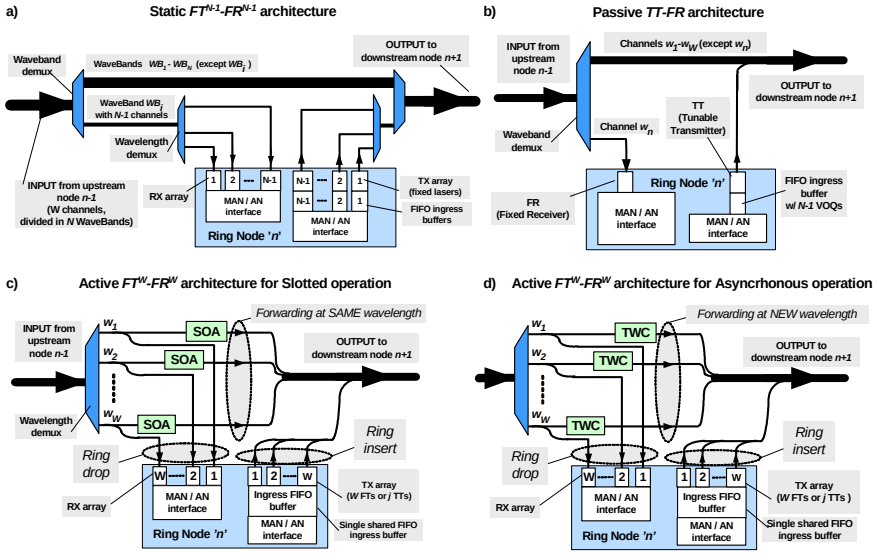

Fig. 6.13. Generic node architectures of: a) Static $FT^{i=N-1}$-$FR^{m=N-1}$, b) Passive Slotted *TT-FR*, c) Active Slotted $FT^{i=W}$-$FR^{m=W}$, and d) Active Asynchronous $FT^{i=W}$-$FR^{m=W}$.

Node architectures are described using the conventional FT^i-TT^j-FR^m-TR^n notation scheme [132], meaning Fixed Transmitter, Tunable Transmitter, Fixed Receiver, and Tunable Receiver. The indices denote the number of each type used in each node, omitted only when the index is '1'. We only include architectures that enable space reuse, illustrated in Fig. 6.13. Their features and complexity are summed up below, based on our study in [p23]:

- **Static $FT^{i<W}$-$FR^{m<W}$ architecture**:
  This architecture is suitable for a Static Wavelength Routed Optical Network (SWRON), which exploits WDM transmission technology, by only using passive multiplexers and demultiplexers to route the wavelengths to their destination node. Avoids *IB* and *RB*, by allocating separate channels for reception and transmission for each node. Benefits include destination stripping, enabling space reuse, and that the complexity with respect to packet insertion and reception is low. However, there is no spectral bandwidth sharing between packet transfers with different source-destination pairs. Moreover, N and W should be related: Full bidirectional connectivity requires a FT^{N-1}-FR^{N-1} architecture, with $W=W_{SWRON}=N(N\text{-}1)/2$ wavelengths per link. However, such a relationship between node count and WDM

channel count limits network flexibility, thus scalability, and prevents support of dynamic traffic.

- **Passive $FT^{i=W}$-$FR^{m<W}$ and $TT^{j<W}$-$FR^{m<W}$ architectures**:
  Again, all nodes use demultiplexers to terminate a (set of) wavelengths, enabling space reuse, which avoids *RB*. However, multiple sources share the same wavelength to address the same destination, which potentially causes *IB* or *FB*, depending on the MAC. Use of TT^j architectures, such as [136, 141], are motivated by reduced transmitter count, potentially enabling a cost reduction, since already commercial tunable semiconductors lasers are less than twice the cost of conventional semiconductor lasers [136]. Support for variations in node input load distribution without excessive *BD* is straightforward in the $FT^{i=W}$ architecture, whilst the TT^j architecture then requires equipping nodes with a high TT count, i.e. high j. However, neither is robust to destination distribution variations, since wavelengths leading to popular destinations may be oversubscribed.
- **Active $FT^{i=W}$-$FR^{m=W}$ and TT^{j}-$FR^{m=W}$ architectures**:
  These architectures use same insertion principle as above, but operating with full receiver accessibility maximises destination distribution flexibility and enables multicast. Both enable full bandwidth sharing. Similar to the Passive architectures, the $FT^{i=W}$-FR^{W} is fully flexible to node input load variation, whilst TT^{j}-FR^{W} requires a sufficiently high j. Space reuse is enabled by an active optical device in the transmission path to remove packets from the ring, similar to [133].

Table 6.8 compares these alternatives, highlighting that increased flexibility comes at the expense of use of active components in the optical path and an increased TT and FR count per node. The Active Slotted design assume an SOA for this purpose in *Tra-PRI* empty-slot MACs, and the Asynchronous Flexible design using the *Ins-PRI AIPSwIT* MAC, replaces the SOA by a TWC [p14, p23], to support VLPs and to minimise *FB*, as detailed in Section B.3.

Compared to the Passive Slotted design and the Active Slotted design, our Active Asynchronous design eases synchronisation, and inherently supports VLP [p23]. Assuming that an Active design is required, the attractiveness of the *AIPSwIT*-based Active Asynchronous design increases with decreased TWC/SOA cost ratio and with increased need for VLP support, without the drawbacks of VLP adaptations needed in slotted operation, as studied in Section B.3. Since many TWCs are SOA based, the main cost-difference may be the additional tunable laser required for the TWC.

Table 6.8. Comparison of node architectures and associated control protocol

Type	Node	Operation mode	MAC	*W* and *N* constraint	VLP support	Bandwidth Flexibility	# Active Devices	Synchronisation	Multicast Compatible?	Cause of Blocking
Static	**FT$^{i=N-1}$-FR$^{m=N-1}$**	Asynchronous	SWRON	$W_{SWRON}=N(N-1)/2$ (1)	Inherent	Low	N/A	Not Required	No	*BD*
Passive	**TT^{j}-FR^{m}**	Slotted	Empty-slot (*Tra-PRI*)	$W=cxN$ (2)	Not inherent	Medium	N/A	Required	No	*BD, IB*
Active	**TT^{j}-FR$^{m=W}$**	Slotted	Empty-slot (*Tra-PRI*)	N/A	Not inherent	High	*W* SOAs	Required	Yes	*BD, IB*
Active	**TT^{j}-FR$^{m=W}$**	Asynchronous	AIPSwIT (*Ins-PRI*)	N/A	Inherent	High	*W* TWCs	Not Required	Yes	*BD, FB*

(1)- Needed for full bidirectional connectivity in a unidirectional ring.

(2)- 'c' is a network-defined integer constant, typically c=1, for *TT-FR* architectures.

B.3. Network Control

B.3.1. Combining Centralised- and Distributed Network Control

As illustrated in Fig. 6.12, we assume in this study that each MAN node is directly connected to a single client entity, e.g. an Access Network (AN). A Service Level Agreement (SLA) covers the pricing and performance aspects between the MAN- and AN operator. The performance specifications describe AN output traffic, such as acceptable load range and traffic statistics, and the Quality of Service (QoS) values offered by the MAN operator.

[146] proposed to decouple the temporal dynamics of hub switching and node access, to simplify operation of a multi-ring network. Similarly, we propose to decouple network dynamics into a centrally-managed coarse MAC setting process, and a local high time-resolution packet scheduling. This is motivated by the time-scale difference between main categories of network events:

- Changes in average behaviour, such as foreseeable day-time load variation, or due to a popular broadcast of multimedia content on a specific server, happen on a large time scale, i.e. >> RTT.
- Packet arrivals from the AN or on the ring happen on small time scale, i.e. << RTT.

The first type of events can be responded to by a network *central controller*, which collects policy information from management interfaces, and user behaviour information, such as average input load at the nodes, by means of a control channel. It then calculates the MAC parameter settings that enable the desired network performance, before distributing them to the ring nodes. Consequently, the time to react to a change is as a minimum bounded by the ring RTT. In addition comes the time needed to perform measurement in the nodes and the processing time in the network controller. Hence, short periods between stable conditions, during which the network does not behave optimally, must be tolerated in this statistical QoS paradigm. The network central scheduler is also responsible for ensuring the relative throughput fairness, because the performance experienced by a user should depend as little as possible on the other users of this AN, and with which AN he wants to communicate [138]. There is no universally accepted fairness definition; we quantify fairness by the Fairness Index (*FI*) [147], which was applied in [145], in Section C.

The second type of events calls for fast distributed *access control* decisions, i.e. allocation of packet insertion and forwarding. This is achieved by an electronic scheduler in each node, which makes decisions

by combining information on packet arrivals and local wavelength availability with the MAC settings. The former can be obtained from either in-band packet headers, or from an out-of-band control channel. In this study we do not specify any particular method, since it is a design choice that does not intrinsically impact performance. In both cases, *processing FDL* must delay the data packets until the scheduler has made a decision.

B.3.2. Distributed MAC protocols

Asynchronous Insertion Priority Scheduling with Insertion Threshold MAC

Our distributed MAC protocol *AIPSwIT* optimises throughput by controlling under which conditions to insert packets from the node ingress buffer. The network operates asynchronously, and the scheduler avoids *IB*, by applying an *Ins-PRI* scheme to resolve contention between inserted and in-transit packets. It exploits the wavelength domain to minimise *FB*, by converting packets to be forwarded to other free wavelengths on the next hop. Simultaneously, the MAC protocol optimises the probability of finding such a free wavelength, by only inserting packets when the number of free wavelengths on the next-hop link is above an insertion threshold. This threshold is denoted $W_{FREE}(n)$, for each node, *n*. In summary, main node scheduler tasks consist of:

- Track number of free wavelengths on its downstream link, and of the transmitters.
- Forward packets by a first-fit selection of one of the idle wavelengths on its downstream link, and discard packets if none is found (prevent use of input wavelength, to avoid in-band conversion).
- Insert packet from the ingress buffer FIFO queue, by a first-fit selection of one of the idle wavelengths, if more than $W_{FREE}(n)$ are idle and a transmitter is free.
- Discard packets that have exceeded a maximum ingress buffer delay, *D*.
- Discard oldest packets if the buffer limit of *B* packets is exceeded.

Comparison with other MAC protocols supporting VLP

"Empty-slot" protocols constitute the main type of distributed MAC protocols allowing space reuse [p23]. They switch fixed-length packets synchronously, and typically apply *a posteriori* scheduling [132], combined with a *Tra-PRI* scheme to resolve contention, by only inserting packets from the ingress buffer when there are free slots on a suitable wavelength on the ring. Their fundamental problem with supporting VLP is that when the scheduler starts inserting a packet that is longer than the slot length, it cannot know whether it will be blocked by a packet in a

later slot on the same wavelength. Different techniques extend empty-slot protocols to support VLPs [136-139], which were compared with *AIPSwIT* in [p23]. Common drawbacks include a sub-optimal slot filling ratio, that synchronisation may be more complex than for asynchronous operation, and that the *a posteriori* scheduling tightens the scheduling time requirements. Other drawbacks, depending on the technique, include complexity and overhead increase associated with segmentation and reassembly, a hardware induced maximum VLP length, discrimination of long packets, and high optical complexity by use of switchable FDLs or SOAs.

In contrast, the asynchronous nature of *AIPSwIT* inherently supports VLPs, thereby avoiding these drawbacks, except the optical hardware complexity. However the drawback of using a TWC as an active optical device for forwarding, is mitigated by also making it carry out the optical gate functionality required to support space reuse in bandwidth flexible architectures, cf. Section B.2.

AIPSwIT applies *a priori* scheduling, in which the next packet to be inserted is chosen before knowing ring bandwidth availability. Still, since it has full bandwidth accessibility, it does not suffer from Head-Of Line (HOL) blocking, which is a drawback of empty-slot *a priori* schemes. Combined with moderate scheduling complexity, (first-fit searches in tables of deterministic size), this limits the per-packet processing time.

Providing Fairness in OPSRNS

Empty-slot MACs typically assume destination-stripping TT^{i}-FR^{m} architectures. A drawback is that nodes that are close to a channel's destination may be starved for bandwidth, since upstream nodes may have already filled up a large ratio of the slots on the wavelength [132]. A fairness technique is then needed.

In the Synchronous Round Robin (SRR**)** protocol, nodes keep a VOQ per destination, which the scheduler scans cyclically to find a packet to insert [144]. Only if the deterministically chosen VOQ is empty, a packet is chosen from the longest VOQ. This compensates starvation of VOQs with low posititional priority, which can be partly overcome with uniform traffic scenarios. However, even with only one oversubscribed channel, SRR is unfair, and some network global fairness control algorithm must be introduced [132, 144]. This calls for exchange of information between nodes, which increases complexity, and for regulation of their bandwidth access, introducing a fairness and channel utilisation trade-off [132].

Such bandwidth access regulation can be provided by credit-based techniques. No node sends more than their predefined credit allows, until all other nodes are satisfied, i.e. has spent their credit, or has nothing

more to send. This information is conveyed using either a token circulating on each wavelength, as in the Multi-Meta Ring Multiple SAT (MMR-MS) [144], or by using the slot headers, as in the Multiple-Asynchronous Transfer Mode Ring, (M-ATMR) [148]. The Distributed Queue Bidirectional Ring (DQBR) represents an alternative approach, and aims at obtaining one distributed First-Come, First-Served (FCFS) queue in the network [136]. Each node signals to their upstream nodes when new packets arrive in their VOQ, by use of a control channel. A counting system then ensures that packets are sent in the order in which they arrived in the network [148].

The Longest Queue First with Random Routing (LQF-RR) protocol aims at improving fairness without information exchange, in an asynchronous network supporting VLPs [145]**.** The node scheduler selects a packet from its longest VOQ (of those associated with a wavelength that contains empty slots), to counteract node starvation. This is combined with a random routing algorithm, in order to alleviate channel overloading, by routing a fraction of the packets through intermediate nodes. However, this increases average hop count, which decreases bandwidth efficiency, leads to additional O/E/O conversions of packets, and causes loss of packet sequence integrity.

This study shows how a network-global *AIPSwIT* parameter setting can provide fairness and high throughput for an asynchronous MAC protocol. Communication between nodes and the central scheduler is only required when significant changes in input load or network configuration occurs. Hence, the distributed scheduler complexity is the same as without fairness. On the other hand, the centralised setting of the insertion thresholds becomes more complex, since the criterion for choosing insertion thresholds now depends on both throughput and fairness. However, since this is a central control task, which is not so time-critical, and since relative coarse parameter granularity enables satisfying performance, we deem the scheduling complexity to be acceptable.

C. Network study formalisation

C.1. Input Traffic formalisation

The result of any network study depends on the network dimensioning and traffic. E.g., given an absolute network input data rate, $L_{NW,ABS}$, it is much more difficult to obtain a high throughput when links have a small bandwidth, or when the traffic has a high average hop-count, H. We capture this by introducing a load parameter, $L_{NW,NORM}$, which is normalised to the total ring bandwidth and the average hop count. We assume that the central controller prevents network oversubscription. However, even for $L_{NW,NORM} \leq 1$, network performance depends heavily on the traffic matrix. Uneven link load distributions leads to some links with high loads, and even link oversubscription, thus high loss probability. We introduce the *Link Load Uniformity Index (LLUI)* in (6.2) to quantify the uniformity of the link load distribution. It is calculated using the N different values for the *Link Network Traffic Share (LNTS)* parameter, applied in Section E. However, note that the network is much more tolerant to avoid *link* oversubscription, being the aggregate of many channels, than the *TT-FR* is towards *channel* oversubscription.

$$LLUI = \frac{\left(\sum_{n=1}^{n=N} LNTS(n)\right)^2}{N\left(\sum_{n=1}^{n=N} \left(LNTS(n)\right)^2\right)} \qquad (6.2)$$

Table 6.9. Performance Input Parameters of this study.

Input Parameter	Comment
N=8	Number of nodes in the network.
W=28	Number of WDM channels per link.
$C_{BITRATE}$=10 Gbit/s	Bitrate per WDM channel.
H	Average hop count of transfers. $H=N/2$ for *uniform* distributions.
$L_{NW,ABS}$	Average absolute amount of data offered to the network per second.
$L_{NW,\ NORM}$	Normalised load, describing the average (offered) utilisation of links for the traffic matrix. $L_{NW,\ NORM} = L_{NW,ABS} \times H / (W \times N \times C_{BITRATE})$
Node Load Distribution	*Server Relative Load Factor* (*SRLF*): Ratio of the server load and load of other nodes. **Uniform node load distribution**: All nodes have the same load
Destination Distribution	**Uniform:** The node's packet are uniformely destined to the other nodes. **Server-biased:** Node sends *SRLF* times more data to Server than to other nodes.
Traffic Distribution	In this study we cover both Poisson and Pareto packet inter-arrival distribution, with negative exponential packet length distribution.
Link Load Distribution	*Link Network Traffic Share, LNTS(n):* Ratio of load at node (n) downstream link and $L_{NW,ABS}$. *Link Load Uniformity Index* (*LLUI*): Quantifies the offered traffic's link load distribution.

We investigate performance for both a uniform traffic scenario, as well as two server scenarios, listed in Table 6.10. In the *UniformTS*, all nodes have same input load, and the link load distribution is balanced. The *SymSrvTS* corresponds to the server node serving an AN with *SRLF* times as many users, which nevertheless exhibits same behaviour as users in other ANs. However, symmetry can only be achieved for a server load smaller than what can be generated by the other nodes combined, hence $SRLF \leq (N\text{-}1)$. The *AsySrvTS* corresponds to the server node containing e.g. a content distribution centre, responding to requests by high data rate transfers.

Table 6.10. Traffic Scenarios studied.

Traffic Scenario	Input Load Distrib.	Destination Distribution	*LNTS* (*n*)	Link Load Distrib.
Uniform (*UniformTS*)	*Uniform* (SRLF=1)	Uniform	1/*N*	Balanced (*LLUI*=1).
Symmetric Server *(SymSrvTS)*	Server-type (SRLF>1)	Server-biased (Uniform in server)	1/N	Balanced (*LLUI*=1)
Asymmetric Server *(AsySrvTS)*	Server-type (SRLF>1)	Uniform	Cf. Section E.	Unbalanced (*LLUI*<1)

C.2. $TT^{j<W}$ architectures

To have enough transmitters for the node in $TT^{j<W}$ architectures, one must dimension j properly depending on the node's input load. In [p23], we defined both a minimum limit of, j_{MIN} (6.3), and a relaxed value, $j_{RELAXED}$ (6.4), for the *UniformTS*. The former corresponds to having just enough transmitters to let the node transmit its entire load, whilst the latter is more tolerant the stochastic packet arrival process and the varying link utilisation. For $L_{NW,NORM}$=0.8, j_{MIN}=6 and $j_{RELAXED}$=9.

$$j_{MIN} = \text{RoundUp}(Load_{NW,\,NORM} \times W / H) \quad (6.3)$$

$$j_{RELAXED} = \text{RoundDown}(1.5 \times j_{MIN}) \quad (6.4)$$

C.3. QoS parameters

Table 6.11 sums up the performance parameters. The *Relative Throughput* (*RT*) defines how much of the total input traffic is successfully transmitted to its destination node. Note that the overhead caused by packet headers and guard times is technology-dependent, and not accounted for. *RT* is related to *PLR*, by *PLR*=1-*RT*.

95 % confidence intervals are shown for main performance graphs, obtained by 10 independent simulation runs with different random seeds, omitting the transient.

This study assumes mean packet durations of 1 µs, link propagation delay of 25 µs (5 km link distance), and *D* between 1-64 µs. Hence, the ring propagation time will typically be the main delay contribution. However, it is constant for each source-destination pair (n_S,n_D), whilst the ingress buffer delay varies between packets. Hence, limiting the maximum time spent in the buffer, *D*, has the beneficial effect of limiting maximum delay jitter.

To quantify *relative* fairness, we adapt the "Fairness Index" (*FI*) proposal of Jain [147]. We use the $RT(n_S,n_D)$ of all $z=N(N-1)$ connections to calculate *FI*, which is bounded by 0 (completely unfair) and 1 (completely fair). As we will see in this study, *FI* should be very close to unity (>0.999), before we intuitively would describe it as "fair". E.g., [145] characterises *FI*=0.9995 as "good fairness".

$$FI = \frac{\left(\sum_{i=1}^{i=z} RT(n_S,n_D)_i\right)^2}{z\left(\sum_{i=1}^{i=z} RT(n_S,n_D)_i^2\right)} \tag{6.5}$$

Table 6.11. MAC protocol– and QoS Parameters

Symbol	Parameters
$W_{FREE}(n)$	*AIPSwIT* MAC protocol defined free wavelength insertion threshold at node *n*.
B	Max number of packets in the ring node ingress buffer.
D	Max delay of packets in ingress buffers (expressed in µs). Corresponds to maximum packet jitter.
RT	Relative Throughput: Ratio of successfully received and offered packets per s.
$RT(n_S,n_D)$	*RT* of packets going from source node n_S to destination node n_D.
FI	Quantifies variations in $RT(n_S,n_D)$, as defined in (6.5).

D. OPSRN performance in Uniform Traffic Scenarios

D.1. Network analysis

Recall from Section B.1, that in our *Ins-PRI* scheme with zero *Receiver Blocking Probability* (*RBP*), any reduction in *RT* is due to either a non-zero *Buffer Discards Probability* (*BDP*), or *Forwarding Blocking Probability* (*FBP*). Maximum *RT* is achieved when the sum of these contributions to packet loss is minimised. The parameters governing *BDP* and *FBP* are listed in Table 6.12, as illustrated throughout this section.

Table 6.12. Analysis of packet loss causes using the *Ins-PRI* based *AIPSwIT* MAC protocol in a TT^{j}-FT^{W} architecture

Parameter Category	*BDP* increases with:	*FBP* increases with:
Node Architecture	Decreased B and j	Increased B and j
QoS	Decreased D	Increased D
***AIPSwIT* MAC settings**	Increased W_{FREE}	Decreased W_{FREE}

Fig. 6.14 illustrates the impact of the *AIPSwIT* MAC protocol's use of insertion thresholds, by showing $RT(n_S,n_D)$ of the 56 different source-destination pairs. The projections on the Source-Destination plane, and of the *RT* values on the plane parallel with the *RT*-Source plane, facilitate reading the graph. Fig. 6.14 a) shows performance for the *SymSrvTS* without insertion thresholds. $RT(n_S,n_D)$=1.0 between a source node and its downstream node, which means that *BDP*=0. However, $RT(n_S,n_D)$ decreases with increasing hop count, meaning that the *FBP* is relatively high. The network is thus unfair, since $RT(n_S,n_D)$ depends on the hop count between source and destination. Fig. 6.14 b) shows performance using the insertion threshold that maximises overall *RT* (W_{FREE}=6 in all nodes). $RT(n_S,n_D)$ is close to unity for all sources and destinations. Hence both overall *RT* and fairness is improved.

Fig. 6.15 and Fig. 6.16 give further insight into the effect of *B*, *D*, *j* and W_{FREE} on *RT* and *FI*. For the FT^{W} architecture, the value of W_{FREE} that maximises *RT* increases with increasing *D*. *FI* converges with W_{FREE} towards unity. Interestingly, *FI* has almost converged at the same W_{FREE} that maximises *RT*, for sufficiently high *D*. Hence, the network is fair without throughput penalty. This shows that when buffer resources are sufficient to enable the MAC to postpone packet insertion until local wavelength utilisation is low without increasing *BDP*, decreases the hop-count dependent *FPB*. For higher W_{FREE}, *RT* decreases, but *FI* remains

high. This represents a “collective misery” phenomenon, in which *FBP* is low, but all nodes suffer equally from high *BDP*.

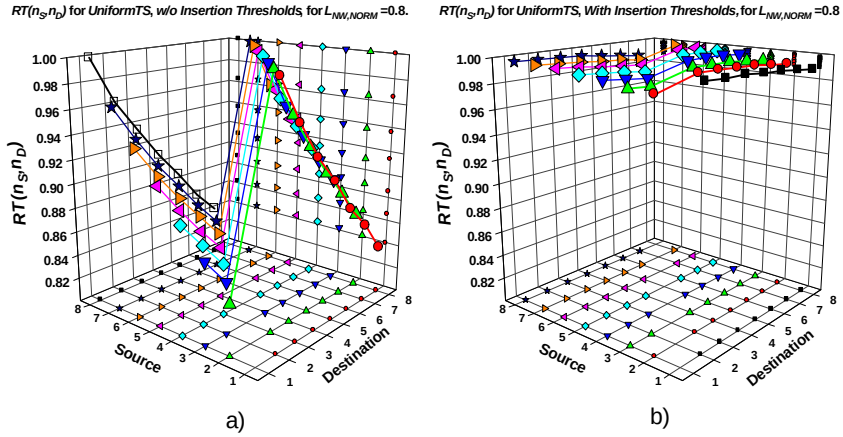

Fig. 6.14. *RT(n_S,n_D)* for FT^W–FR^W architecture with *B*=64 and *D*=16 μs. a) without- and b) with insertion threshold (W_{FREE}=6).

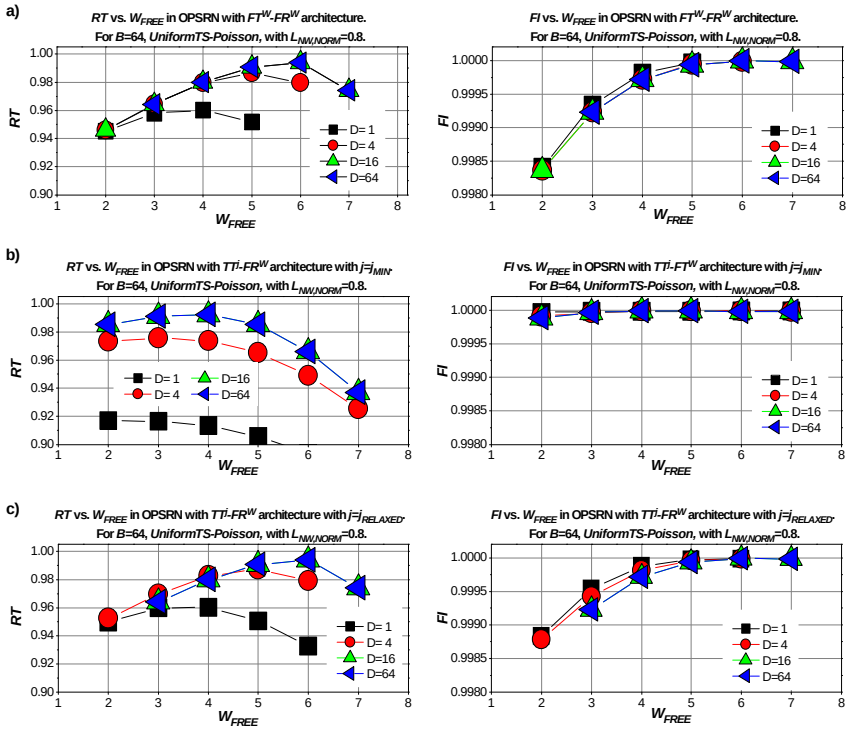

Fig. 6.15. *RT* and *FI* vs. W_{FREE}, for *B*=64. a) FT^W-FR^W, b) TT^j-FR^W with $j=j_{MIN}$, and c) TT^j-FR^W with $j=j_{RELAXED}$.

Similar results are obtained for the TT^j architecture, using both $j=j_{MIN}$=6 and $j=j_{RELAXED}$=9. Of these two, the latter case obtains a slightly higher maximum *RT*, for a higher W_{FREE}. This can be expected, since a higher TT count enables higher parallel insertion capability, thus waiting longer in buffer for beneficial insertion moments, without being penalised, similar to above. However, the former case obtains higher *RT* and *FI* for low W_{FREE} (for sufficient buffer resources), since the limited transmitter count in fact acts as an insertion limitation.

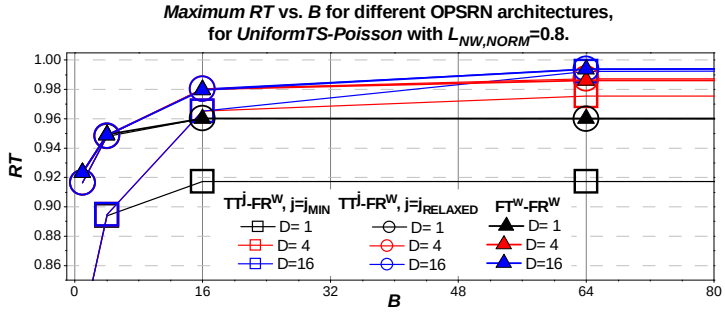

Fig. 6.16. Maximum *RT* vs. *B*, for different node architectures and *D*.

Fig. 6.16 includes the impact of *B*, and shows how maximum *RT* (*RT* for the optimum W_{FREE}) converges towards an asymptotic value governed by the node architecture and *D*, with increasing *B*. For sufficiently high *B*, using $j_{RELAXED}$ in the TT^j architecture has the same performance as the FT^W architecture, whilst using j_{MIN} requires a larger *D* to reach the same *RT*. Conversely, for *D*=16 µs, we see that using j_{MIN} requires *B*=64, instead of *B*=16 for $j_{RELAXED}$, to reach the same *RT* as the FT^W architecture. The importance of having enough transmitters is hence greater for small *B* and *D*. In the remainder of this Section, $j=j_{RELAXED}$.

Fig. 6.17 quantifies the *RT* and *FI* improvement achieved by the *AIPSwIT* MAC insertion threshold, for a large range of $L_{NW,NORM}$ values. We assume *D*=64 µs, and study four values of *B*. Without the insertion threshold, performance does not increase with *B*, i.e. ingress buffers are not exploited. On the other hand, with the insertion threshold, a high performance depends on sufficient buffer resources, to be able to postpone insertion to a period with low link utilisation. For $Load_{NW,NORM}$>0.6, the insertion thresholds improve *RT*, and *FI* in particular, since it reduces *FBP*.

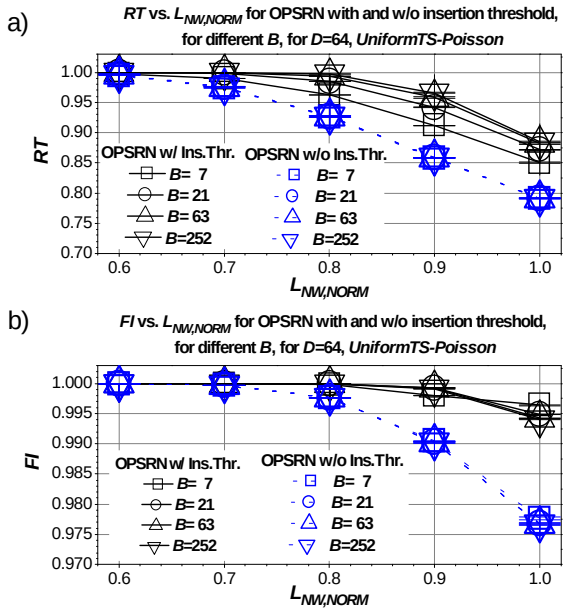

Fig. 6.17. Performance for D=64 μs and different B, with and w/o insertion thresholds. a) Max RT vs. $L_{NW,NORM}$, and b) FI vs. $L_{NW,\,NORM}$.

D.2. Comparing OPRSN and SWRON performance

D.2.1. Complexity

We start by noting that the choice of W=W_{SWRON}=N(N-1)/2=28 for N=8, is beneficial for the SWRON, cf. Section B.2. Regarding the node design, we study four values of B between 7 and 252. In the SWRON, each node has a FIFO ingress buffer per FT, which stores packets from the AN, before packets are inserted on the ring, when the FT is free.

Fig. 6.18 compares transmitter count in the SWRON $FT^{i=N-1}$ and OPSRN TT^{j} architecture, with j=$j_{RELAXED}$, vs. N, for different $L_{NW,NORM}$. We observe that for N=8, OPSRN enables a transmitter count reduction for $L_{NW,\,NORM}$ <0.7.

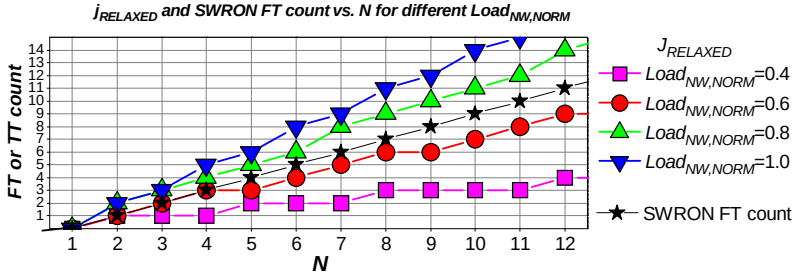

Fig. 6.18. SWRON FT count and OPSRN TT count, using j=$j_{RELAXED}$.

D.2.2. Performance

The SWRON has complete isolation between all source-destination pairs, and it is therefore completely fair in this *UniformTS*. Moreover, this enables assessing network performance by studying a single unidirectional communication. It also enables verifying our simulation results for the SWRON in the extreme cases, with 'infinite" buffers, and with no buffers. In the former case, we find that *RT* equals 1 for $L_{NW,NORM} \leq 1$. This confirms that in a non-oversubscribed FIFO buffered system, no packets are lost for sufficiently large buffers and without any limit on storage time. In the latter case, i.e. $B=0$, *RT* was confirmed to be equal to the value provided by the Erlang-B formula.

Poisson input traffic

Fig. 6.19 shows that the OPSRN has a higher *RT* than the SWRON for $L_{NW,\ NORM} \leq 0.9$ and $B \leq 63$. Recall from Fig. 6.17 that the OPSRN has a $FI > 0.9999$ for $L_{NW,\ NORM} \leq 0.8$, for $B \geq 21$. Hence, the OPRSN is very fair in this range. On the other hand, with a sufficiently high *B*, the SWRON has higher *RT*.

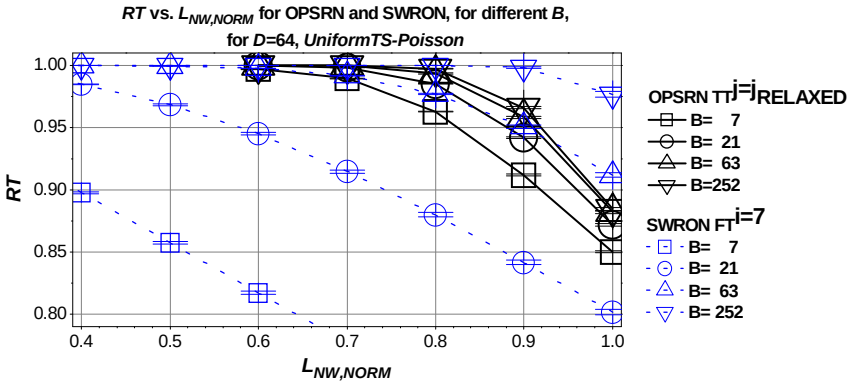

Fig. 6.19. Comparing OPSRN and SWRON *RT* vs. $L_{NW,NORM}$, for *D*=64 µs, for different *B*.

Fig. 6.20 details the impact of buffer dimensioning, including the impact of *D*, for $L_{NW,NORM}=0.8$. It confirms that for low *B*, the OPSRN clearly outperforms the SWRON, and the difference increases with decreasing *D*. E.g. for *D* of 1 and 4 µs, the *RT* of the SWRON converges towards asymptotic values of 0.72 and 0.9, respectively, whereas the OPSRN can reach 0.96 and 0.98, for *B* as low as 16. We attribute this advantage to the increased bandwidth sharing of the OPSRN and the economy of scale of operating a large buffer, instead of a buffer per channel. Both factors are most beneficial when buffer capacity is scarce. On the other hand, for

B>96, the SWRON is able to reach higher *RT* for *D*>16 µs, since *FB* does not take place at all.

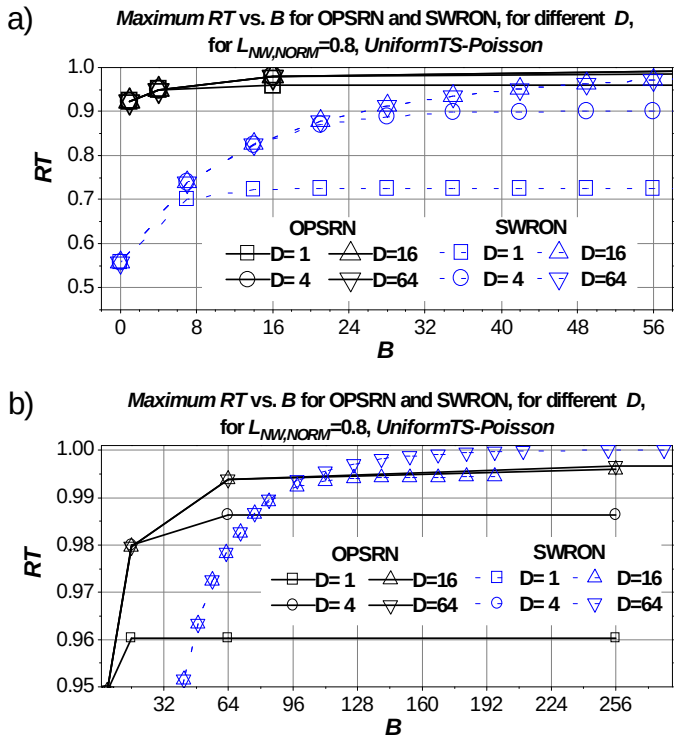

Fig. 6.20. *RT* vs. *B*, for Poisson arrival with $L_{NW,\ NORM}$=0.8, for different *D*. a) overview, b) zoom-in on high *RT* region.

Bursty input traffic

Whilst the Poisson arrival model may be representative for the highly aggregated traffic in the core network [132, 130], the traffic closer to the edge is expected to be more self-similar [132]. We assess this scenario, by an approximation of self-similar AN input traffic, termed *bursty* in the following. It is modelled using independent traffic generators, generating packet arrivals according to a Pareto distribution [143], with Hurst parameter of 0.8. Compared to the Poisson arrival of Fig. 6.19, Fig. 6.21 shows that *RT* decreases moderately both for the OPSRN and the SWRON, using the same node architecture, with $j=j_{RELAXED}$. The OPSRN can support $L_{NW,NORM}$ up to 0.8 with *FI*>0.999. Hence, the network is rather robust to self-similar traffic, but note that higher Hurst values further detoriates performance. However, this can be solved by reducing the network load.

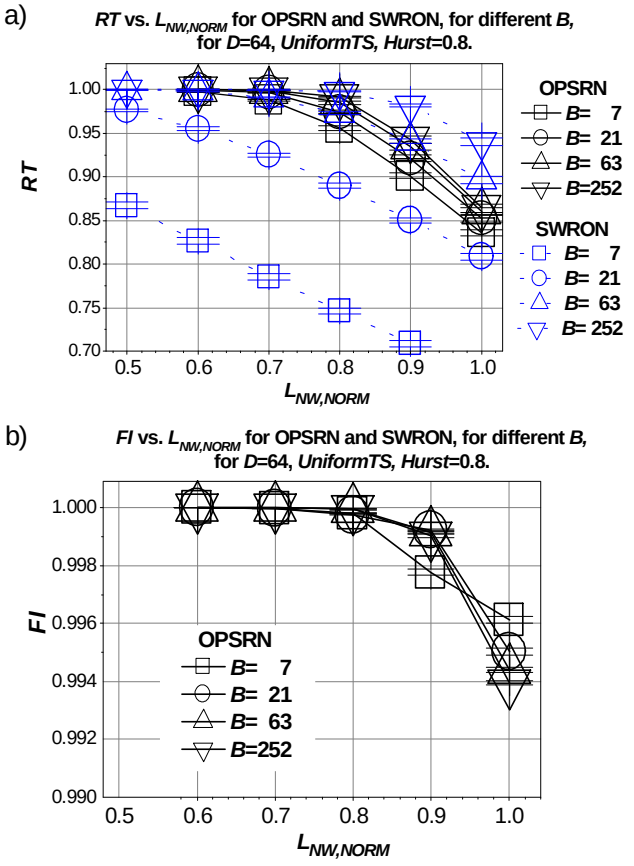

Fig. 6.21. a) *RT* for OPSRN and SWRON for bursty input traffic, b) *FI* for OPSRN for bursty input traffic.

D.2.3. Summary of OPSRN-SWRON comparison for uniform traffic

We observed that for the *UniformTS*, the *AIPSwIT* MAC protocol and associated active node architecture enabled a comparable fairness and a higher throughput than the SWRON for a wide load range, for either: *i)* low-to-moderate *B*, or *ii)* low-to-moderate *D*. However, the transmitter reduction is not significant, and the optical complexity is significantly higher for the OPSRN, since it uses a TWC per wavelength per node. Hence, unless *B* is a very important cost factor, or *D* is a critical QoS requirement, the SWRON is probably the more cost-effective solution. However, as discussed in 2.2, the SWRON node architecture is static, and cannot easily adapt to changing traffic matrices without channel oversubscription, thus unfairness. Moreover, its constraint between *N* and *W* makes it cumbersome to insert nodes on the ring.

We conclude that OPSRN is very bandwidth efficient, but that its opportunity lies in non-uniform traffic, which is dynamically changing. The remainder of this study assesses the suitability of the OPSRN for these scenarios.

E. Performance for Non-uniform traffic

We aim to show that the network can be efficient and fair both for *SymSrvTS* and *AsySrvTS*, with Poisson input traffic. For maximum flexibility, we assume a FT^W–FR^W architecture, which means that any of the nodes can be the server node, without hardware modifications. Moreover we maintain buffer resources of B=64 and D=16 μs, to limit hardware usage, and to be compatible with stringent QoS delay-jitter requirements.

E.1. Symmetric Server Traffic Scenarios

Recall from Section C that this scenario is limited to $SRLF \leq 7$. Fig. 6.22 shows $RT(n_S,n_D)$ for the *SymSrvTS* with the server in node '1', without thresholds, and with *AIPSwIT* insertion thresholds that maximises overall *RT*. In the former case, similar to the *UniformTS*, *BDP*=0. However, $RT(n_S,n_D)$ decreases with increasing hop count, meaning that the *FBP* is relatively high. There is a steeper decrease in $RT(n_S,n_D)$ when a transfer passes node 1, which means that the risk for *FB* is particularly high at the server, due to its high input load. This decrease can be easily observed by studying the "gap" in $RT(n_S,n_D)$ projections. The network is thus unfair, since $RT(n_S,n_D)$ depends on the hop count between source and destination, and also whether the transfer passes the server or not. Fig. 6.22 b) shows that that the *AIPSwIT* MAC protocol significantly improves both throughput and fairness. The insertion thresholds which maximise overall *RT* also correspond to a very high fairness: $RT(n_S,n_D)$ is close to unity for all sources and destinations.

Fig. 6.23 shows that with *AIPSwIT*, one achieves RT>0.999 and FI>0.99999, for $L_{NW,NORM}$ up to 0.7, for $SRLF$=6. For comparison, without insertion thresholds, one needs $L_{NW,NORM}$< 0.6 to have RT>0.99, even for SRLF=1. Note that a $L_{NW,NORM}$ =0.8 can be supported for RT>0.99 and FI>0.99997, for $SRLF$=6.

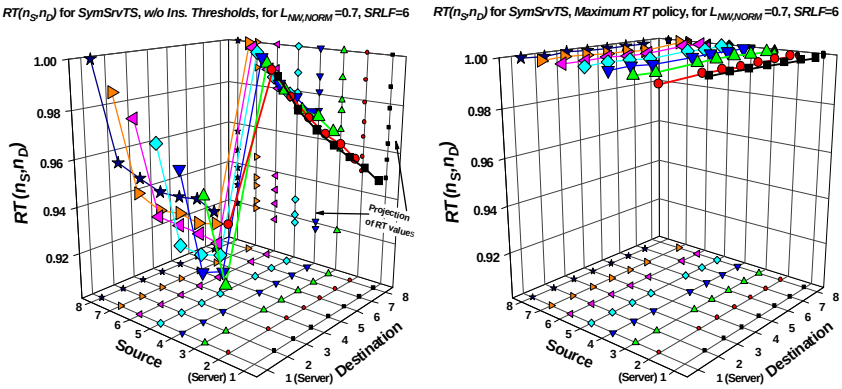

Fig. 6.22. *RT*(n_S,n_D) for *SymSrvTS* with $L_{NW,NORM}$=0.7 and *SRLF*=6. a) w/o Ins. Thresholds, and b) *Maximum RT* policy.

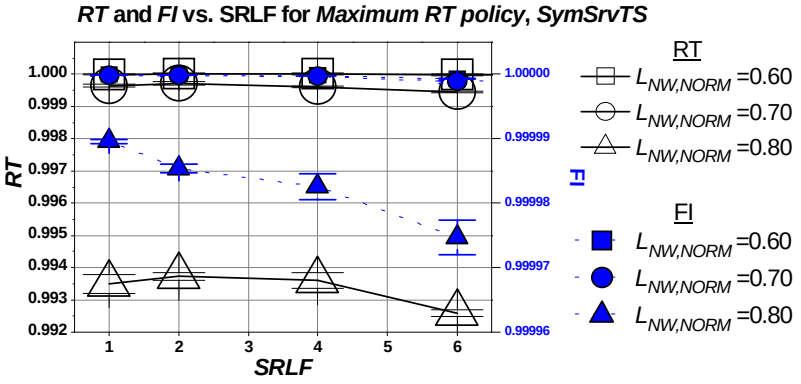

Fig. 6.23. *RT* and *FI* for *SymSrvTS* vs. *SRLF*, for different $L_{NW,NORM}$, in the *Maximum RT* policy.

E.2. Asymmetric Server Traffic Scenarios

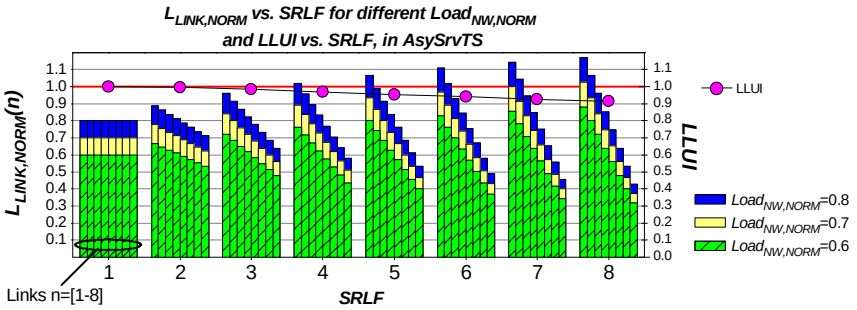

Fig. 6.24. Normalised link load vs. *SRLF* for different $L_{NW,NORM}$ (primary axis), and *LLUI* (secondary axis) vs. *SRLF*, for *AsySrvTS*.

Whilst link oversubscription is avoided for *UniformTS* and *SymSrvTS* when $Load_{NW,NORM} \leq 1$, this is not the case for *AsySrvTS*, where we have an unbalanced link load distribution. Fig. 6.24 illustrates the link bandwidth normalised load of all 8 links, $L_{LINK,NORM}(n)$, as a function of *SRLF*, for the three studied $L_{NW,NORM}$. We observe that link '1', which is the downstream link of the server, has the highest load. Note that $L_{LINK,NORM}(1)>1$ inevitably results in a sub-unity *RT*. Hence, the network suffers when both *SRLF* and $L_{NW,NORM}$ are high. Fig. 6.24 also quantifies this unbalanced distribution of offered traffic on the links by the *Link Load Uniformity Index*, *LLUI*. As stated in [132], challenging traffic scenarios require a network-wide control, to limit insertion, which generates a throughput-fairness trade-off. We observe the same phenomenon for asymmetric traffic with high load. Hence, the network operator has flexibility between prioritising fairness, or high overall throughput, by varying the MAC parameter setting. We propose two scheduling policies to govern the MAC protocol parameter setting, namely *Maximum RT* and *Balanced RT-FI*. The former aims at maximising *RT*, and the latter aims at achieving a high *RT*, whilst maintaining a high *FI*.

We consider the case of $L_{NW,NORM}$=0.7 and *SRLF*=8, which gives oversubscription of link 1. Fig. 6.25 a) shows that without insertion thresholds, $RT(n_S,n_D)$ suffers dramatically when a packet transfer passes the server node. Fig. 6.25 b) shows performance of the *Maximum RT* policy, which maximises overall *RT* and to a certain degree improves fairness. Fig. 6.25 c) shows that the *Balanced RT-FI* policy further improves fairness, by increasing the lower *RT* "floor", at the expense of decreasing the *RT* of transfers from the server. This slightly lowers overall *RT*, since these connections constitute more than half of the network load. This difference in performance is due to a slight increase in

insertion threshold for the server, compared to that of the *Maximum RT* policy.

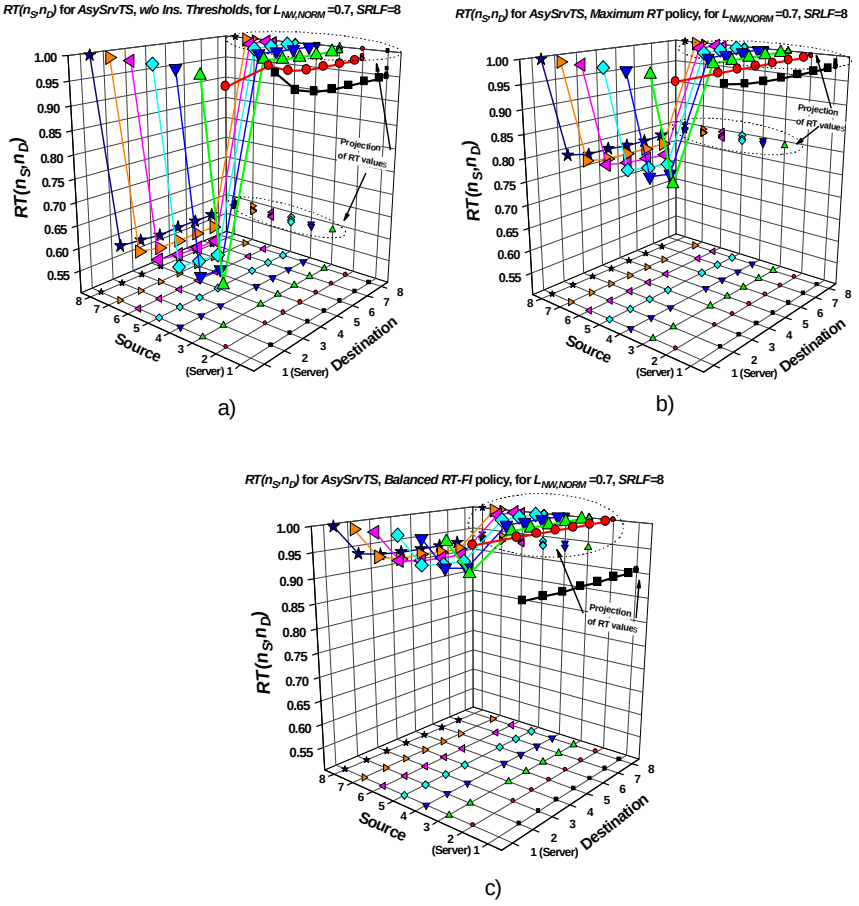

Fig. 6.25. MAC protocol's impact on fairness in an oversubscribed link scenario, for $L_{NW,NORM}$ =0.7 and *SRLF*=8, a) *w/o Insertion Threshold*, b) *Maximum RT* policy, and c) *Balanced RT-FI* policy.

Fig. 6.26 quantifies the performance for $0.6 \leq L_{NW,NORM} \leq 0.8$ and $1 \leq SRLF \leq 8$. In particular, for the above example of $L_{NW,NORM}=0.7$ and $SRLF=8$ it confirms that the *Balanced RT-FI* achieves a significant *FI* improvement (from 0.987 to 0.998), for a minor *RT* decrease (from 0.946 to 0.935), compared to the *Maximum RT* policy.

The two policies yield the same parameter setting at low loads, i.e. maximum *RT* coincides with a high *FI*, just as for the *SymSrvTS*. These graphs illustrates that the performance detoriates with increasing load and increasing *SRLF*, as listed below for the *Balanced RT-FI* policy. There is hence a trade-off between acceptable network load and flexibility with respect to traffic scenarios.

- $L_{NW,NORM} \leq 0.6$ and $SRLF \leq 8$ enables $RT \geq 0.995$ and $FI \geq 0.99997$.
- $L_{NW,NORM} = 0.7$ and $SRLF = 3$ enables $RT \geq 0.995$ and $FI \geq 0.99999$.
- $L_{NW,NORM} = 0.8$ and $SRLF > 1$ yields $RT < 0.99$.

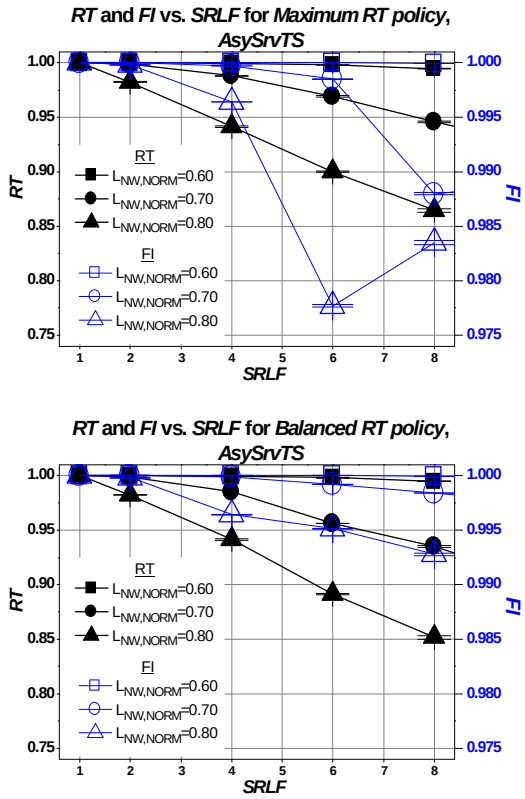

Fig. 6.26. *RT* and *FI* for asymmetric server scenarios, both for *Maximum RT* policy and for *Balanced RT-FI* policy.

F. Conclusion

Fairness is a crucial, but often overlooked issue in statistically multiplexed network studies. Envisaging a scenario where the MAN operator offers minimum relative throughput levels between all source-destination pairs, fairness techniques enable a MAN operator to increase the network load, which potentially increases revenues.

The high flexibility and bandwidth efficiency of the *AIPSwIT* MAC protocol and associated node architecture enable an OPSRN with a higher throughput than a SWRON, for moderate loads and limited buffer resources. However, its use of more sophisticated optical technology will probably only be justified when traffic is non-uniform and dynamically varying. In this case, maximum overall throughput and fairness can be achieved simultaneously for symmetric server scenarios. The asymmetric server scenarios result in a more demanding unbalanced link load distribution. However, even in this scenario, controlling the insertion thresholds enables controlling the throughput-fairness trade-off. Considering its use of distributed access control and moderate complexity centralised threshold settings, we deem the *AIPSwIT* MAC protocol a promising fairness technique for OPSRNs.

7. Conclusion

This thesis has reviewed the rationale for OPS and its main design options. We conclude that OPS is a promising switching paradigm for a future optical network. Its main benefit is the combination of high capacity optical switching and the efficiency of statistical multiplexing. Additional benefits include compatibility with packet based traffic engineering and capability for advanced network features, such as Quality of Service differentiation and Fairness. However, further research and development efforts are required to make OPS a practical and attractive candidate for the optical layer. This Ph.D. project has worked towards this target, mainly through the OPS research contributions summed up in this chapter.

It was highlighted that the OPS and OBS concepts are approaching each other. Both may use the same packet handling scheme, and recent works on OBS employ buffering for contention resolution, as opposed to the assumptions in early works on OBS. However, some important differences remain: The relatively large OBS payload relaxes its timing constraints on both scheduling and switching operations. Hence, OBS may be more suitable in the short term. This project has focused on OPS, which has the higher flexibility. However, due to the assumed asynchronous operation and the abstraction of the control plane communication, most of the achieved results can be applied also in an OBS context.

The overview of optical switching technologies identified Array Waveguide Grating Routers (AWGR) and Broadcast-and-Select (B&S) as the two most promising types of switches. Their potential is clear by having demonstrated Tbit/s capacity. However, each architecture has different complexity challenges, e.g. with respect to component count, maximum parameters and the number of interconnections in a switch. Progress in integration is therefore a prerequisite to enable realistic switch designs. The technological maturity and cost determine which architecture is the most suitable at a given point in time.

Regarding other aspects of node design, it was highlighted that a number of the proposed OPS switching matrices, header processing schemes and regenerator designs apply wavelength converters. Wavelength converters also enable contention resolution and adaptation between switch internal

and WDM channel wavelengths. It was stressed that combining several functionalities in the wavelength converters can reduce overall component count, and a proposal for such a design was made.

To obtain a sufficiently low PLR for a reasonably high load, it is necessary to efficiently resolve contention. This thesis has focused on contention resolution pools, where TWCs and/or FDLs are shared among the packets at different input ports. This enables e. g. ~30-40 % reduction in the TWC count, for a load of 0.7. Moreover, the *SoftRSV* algorithm was proposed and shown to increase the efficiency of using FDLs in a port-constrained SPN pool. Compared to a pure TWC pool, a mixed pool containing TWCs and FDLs enables either reducing PLR by a decade, or replacing half of the TWCs by FDLs without any PLR penalty. This gives the network designer increased flexibility to find the most attractive mix of FDLs and TWCs for SPN pools.

Optical packet switches face several scalability constraints. The proposed Shared Per Waveband Plane (SPWP) switch design can overcome scalability constraints, by using parallel switching planes, with passive separation and recombination. Additional benefits include modular upgrade scenarios, QoS differentiation, hybrid OCS and OPS networks, as well as OCS to OPS migration scenarios. Dimensioning flexibility was further improved by introducing WP-internal QoS differentiation, in the SPWP+ design.

Hybrid OCS/OPS networks may constitute an attractive solution to offer transfer guarantees to a fraction of the traffic, whilst benefiting from statistical multiplexing gains. This project has contributed to development of the hybrid network concept. In particular, the use of a Polarisation Beam Splitter and an AA-MZI based TWC for input interface processing was proposed. This design combines the segregation of two CoS, the separation of packet header and payload, as well as the wavelength conversion needed in the first stage of an AWGR switch matrix. The very high functionality level obtained by relatively few active components makes it a potentially very attractive input interface design.

QoS differentiation enables support of a wide range of services, whilst maintaining low overall resource usage. However, buffer based Active Queue Management (AQM) algorithms are not suitable for QoS differentiation in OPS. Therefore, bufferless QoS differentiation schemes were evaluated with respect to performance and complexity. The Pre-emptive Drop Policy (PDP) is the most efficient, i.e. giving the least throughput penalty for the same isolation, especially important at high system loads. The Intentional Packet Dropping (IPD) scheme suffered from very poor performance, whilst the Wavelength Allocation (WA)

based scheme represents a performance-complexity compromise. Note that WA is a special case of Access Restriction (AR).

The proposed two-dimensional AR-based QoS differentiation for SPN designs, decreases the PLR of the high-priority CoS by roughly an order of magnitude (for the same PLR of the low-priority CoS), compared to when applying AR on TWCs or wavelengths, exclusively. The algorithm was shown to scale well with increasing fibre- and wavelength count, and to be robust towards overload situations. In another AR QoS differentiation study, the overall PLR was reduced by adding FDLs in the SPN pool. However, FDLs induce jitter, which may not be tolerated by a number of applications. This motivates including jitter tolerance as a CoS aspect, in addition to PLR. A number of schemes were proposed, depending on the desired CoS granularity. The penalty paid to obtain a given isolation increased with increasing CoS granularity, with decreasing jitter tolerance of traffic, and when the CoS with the lowest PLR threshold should be jitter-free. These results highlight the need for careful analysis of the input traffic requirements to both jitter and loss, when dimensioning QoS differentiated networks.

Optical packet switched ring networks are promising candidates to bridge the metro gap. This project proposed ring node- and ring interchanger designs, using AA-MZI based TWCs for dropping and forwarding of packets. The designs represent a new approach to support Variable Length Packets (VLP) in Optical Packet Switched Ring Networks. The flexibility of the TWC enables using it as an optical gate to enable space reuse, as well as a wavelength converter to resolve contention when VLP are inserted. The scheduling is controlled by the proposed *Asynchronous Insertion Priority Scheduling with Insertion Thresholds* (*AIPSwIT*) distributed MAC protocol. Its use of insertion thresholds improved the load supported by the network by roughly 10 - 50 %, for low PLR values. Extensions to *AIPSwIT* enable combining high throughput with high fairness, both for balanced and unbalanced traffic matrices.

In addition to these studies, this project has contributed to lab-demonstrations of optical wavelength conversion, bitrate conversion and optical logic processing, not described in detail in this thesis. Progress in this field paves the way for increased use of optics in OPS networks.

This thesis may be used as a fundament for a number of future OPS research studies. Regarding hardware aspects, the proposed optical designs would benefit from scalability/cascadability analysis, investigation of integration potential, and assessment of power consumption. Of particular interest are new designs that combine data-

and control plane functionalities. Regarding performance studies, interesting topics for future work include network level QoS differentiation performance for the WAN. Moreover, a waveband concept can be applied to the proposed OPS ring networks for the MAN to reduce the TWC count. To optimise performance during dynamically varying traffic for the proposed QoS differentiation schemes, as well as for the MAC protocols, algorithms for optimally setting scheduling parameters should be devised. Finally, to identify when OPS is more attractive than OCS, quantification of statistical multiplexing gains of OPS, including the benefit of packet-level traffic engineering, should be combined with techno-economic studies, when prices can be more accurately assessed.

The results from this Ph.D. project have been disseminated through publications in international journals and conferences, combined with participation in European projects. Hence, this project has contributed to the progress of OPS research through analysis of existing solutions, and proposals for new designs and methods for advanced network features, such as QoS differentiation, support of VLP and Fairness. The continuing effort from both the academic and the industrial world will increase the maturity and viability of the OPS networking paradigm, which may eventually unleash the power of packet switching also in the optical layer.

REFERENCES

REFERENCE LIST

1 Michael Lahanas. “Ancient Greek Communication Methods” (URL:www.mlahanas.de/Greeks/Communication.htm), Jan. 14 2005.

2 R. V. Jones. “Optical Telegraphy-The Chappe Telegraph Systems“ (URL:http://people.deas.harvard.edu/~jones/cscie129/images/history/chappe.html), Feb. 2 2004.

3 “Nobel Lectures, Physics 1963-1970”, Elsevier Publishing Company, (Amsterdam, The Netherlands), 1972.

4 R. Ramaswami, and K. N. Sivarajan. “Optical Networks: A Practical Perspective”, Second Edition, Morgan Kaufmann, 2002.

5 S. Yin. “Declines in Internet traffic growth won’t derail capex”, Lightwave, (URL:www.rhk.com/rhk/press/Lightwave%208.04.pdf), Aug. 19 2004.

6 W. Leland, M. Taqqu, and D. Wilson. “On the self-similar nature of Ethernet traffic”, IEEE/ACM Trans. Networking, vol. 2, pp. 1-15, Feb 1994.

7 A. M. Odlyzko. “Internet traffic growth: Sources and implications”. Optical Transmission Systems and Equipment for WDM Networking II, (ed: B. B. Dingel, W. Weiershausen, A. K. Dutta, and K.-I. Sato), Proc. SPIE, vol. 5247, pp. 1-15, 2003.

8 DSL Forum. “DSL Subscribers to 31 December 2004 - FACTS”, (URL:www.dslforum.org/pressroom.htm), March 11th 2005.

9 F. Neri, and G. N. Rouskas. “Editorial”, Elsevier Journal on Optical Switching and Networking, vol. 1, pp. 1-2, 2005.

10 N. McKeown, “How scalable is the capacity of (electronic) IP routers”, Techn. Digest OFC 2001, (Anaheim, CA, USA), 2001.

11 F. M. Chiussi, and A. Francini. “Scalable Electronic Packet Switches”, IEEE Journ. Selected Areas in Communications, 21(4), pp. 486-500, 2003.

12 D. Saha. “Control Plane for Optical Networks: The State of the Standards”, Proc. ECOC 2002, vol. 3, Tutorial 6, (Copenhagen, Denmark), 2002.

13 Y. Lee, and B. Mukherjee. “Traffic Engineering in Next-Generation Optical Networks”, IEEE Communications Surveys, 6(3), pp. 16-33, 2004.

14 R. H. Zakon. ”Hobbes' Internet Timeline v8.0”, (URL:www.zakon.org/robert/internet/timeline/), Jan. 1st 2005.

15 B. Feng, N. Stol, E. Zouganeli, R. Ø. Andreassen, A. Solem, R. B. Haugen, A. Sudbø, and B. E. Helvik. “Direct comparison between optical circuit and optical packet/ burst switching using simulations”, Proc. 9th IEICE OptoElectronics and Communications Conference/ 3rd Conference on Optical Internet (OECC/COIN 2004), paper 14A2-2, (Yokohama, Japan), 2004.

16 E. Zouganeli, R. Ø. Andreassen, B. Feng, N. Stol, A. Solem, B. E. Helvik, A. Sudbø, and R. B. Haugen. “Optical packets, circuits, or packets and circuits: Viability of optical packet/burst switching”, To be published in Proc. of 10th OptoElectronics and Communications (OECC 2005), (invited paper), (Seoul, South Korea), 2005.

17 S. J. B. Yoo. “Optical-Label Switching, MPLS, MPLambdaS and GMPLS”, SPIE/Kluwer Optical Networks Magazine, 4(3), pp. 17-31, 2003.

18 M. J. O’ Mahony, D. Simeonidou, D. K. Hunter, and A. Tzanakaki. ”The application of optical packet switching in future communication networks”, IEEE Comm. Mag., (39)3, pp. 128-135, 2001.

19 T. S. El-Bawab, and J.D. Shin. “Optical Packet Switching in Core Networks: Between Vision and Reality”, IEEE Comm. Mag., 40(9), pp. 60-65, 2002.

20 G. E. Moore. “Cramming more components onto integrated circuits “, Electronics, 38(8), 1965.

21 Converge Network Digest, “Intel Develops Silicon Laser”, (URL:www.convergedigest.com/DWDM/dwdmarticle.asp?ID=13810), Feb. 16 2005.

22 Converge Network Digest., “Infinera Goes Live with Photonic IC Technology”, (URL:www.convergedigest.com/DWDM/DWDMarticle.asp?ID=13616), Jan. 31 2005.

23 C. Guillemot, M. Renaud, P. Gambini, C. Janz, I. Andonovic, R. Bauknecht, B. Bostica, M. Burzio, F. Callegati, M. Casoni, D. Chiaroni, F. Clerot, S. L. Danielsen, F. Dorgeuille, A. Dupas, A. Franzen, P. B. Hansen, D.K. Hunter, A. Kloch, R. Krahenbuhl, B. Lavigne, A. Le Corre, C. Raffaelli, M. Schilling, J.-C. Simon, and L. Zucchelli. “Transparent Optical Packet Switching: The European ACTS KEOPS Project Approach”, IEEE Journ. Lightwave Techn., 16(12), pp. 2117-2133, 1998.

24 P. Gambini, M. Renaud, C. Guillemot, F. Callegati, I. Andonovic, B. Bostica, D. Chiaroni, G. Corazza, S. L. Danielsen, P. Gravey, P. B. Hansen, M. Henry, C. Janz, A. Kloch, R. Krahenbuhl, C. Raffaelli, M. Schilling, A. Talneau, and L. Zucchelli. “Transparent optical packet switching: network architecture and demonstrators in the KEOPS project”, IEEE Journ. Selected Areas in Communications, 16 (7), pp. 1245-1259, 1998.

25 L. Dittmann, C. Develder, D. Chiaroni, F. Neri, F. Callegati, W. Koerber, A. Stavdas, M. Renaud, A. Rafel, J. Solé-Pareta, W. Cerroni, N. Leligou, L. Dembeck, B. Mortensen, M. Pickavet, N. Le Sauze, M. Mahony, B. Berde, and G. Eilenberger. ”The European IST Project DAVID: A Viable Approach Toward Optical Packet Switching”, IEEE Journ. Selected Areas in Communications, 21(7), pp. 1026-1040, 2003.

26 L. Dittmann. “Optical Packet Networks– Conclusions from the IST DAVID project”, Techn. Digest OFC 2004, vol. 1, pp. 560-562, paper TuQ1, (Los Angeles, CA, USA), 2004.

27 D. Chiaroni, A. Dupas, E. Dutisseuil, B. Lavigne, H. Weissing, B. Mortensen, M. Berger, L. Dittmann, H. Linardakis, A. Salis, A.Stavdas, W. Lautenschlaeger, J. Karstaedt, L. Dembeck, and G. Eilenberger. “Optical packet switching solutions for the metro and the backbone: Main conclusions from the DAVID project demonstration”, Proc. ECOC 2004, vol. 4, pp. 886-887, (Stockholm, Sweden), 2004.

28 M. J. O'Mahony, K.M. Guild, D. K. S. Hunter, I. Andenovic, I.H. White, R. V. Penty, and L. Zucchelli. “An optical packet switched network (WASPNET)- concept and realization”, SPIE/Kluwer Optical Networks Magazine, 2(6), pp. 46-53, Nov/Dec 2001.

29 “OPSnet Optical Packet Switching for the Internet”. (URL:http://www.comms.eee.strath.ac.uk/OPSnet/main.html), Feb. 28 2005.

30 D. Klonidis, R. Nejabati, C. Politi, M. O'Mahony, and D. Simeonidou. “Demonstration of a fully functional and controlled asynchronous optical packet switch at 40Gb/s”, Proc. ECOC 2004, Post-Deadline Paper Th4.5.5, pp. 60-61, (Stockholm, Sweden), 2004.

31 N. Wada. “NICT activities on optical packet switched networking and JGN2: Advanced Network Testbed for R&D”, e-Photon/ONe Workshop on International Collaborations, (URL:www2.nict.go.jp/jt/a115/member/JGN2_wada_rev.pdf), September 7th, 2004.

32 N. Wada, H. Harai, and F. Kubota. “40Gbit/s, multi-hop optical packet routing using optical code label processing based packet switch prototype”, Techn. Digest OFC 2004, paper FO7, vol. 2, pp. 640-642, (Los Angeles, CA, USA), 2004.

33 M. Yoo, M. Jeong, and C. Qiao. ”A High Speed Protocol for Bursty Traffic in Optical Networks”, SPIE All-Optical Communications Systems, vol. 3230, pp. 79-90, November 1997.

34 D. J. Blumenthal, P.R. Sauer, and J.R. Sauer. “Photonic Packet Switches: Architectures and Experimental Implementations”, Proc. of the IEEE, 82(11), pp. 1650-1667, 1994.

35 Lightreading. “The 100-Tbit Router Race”, (URL:www.lightreading.com/document.asp?doc_id=53116), May 21st 2004.

36 T. Durhuus. “Semiconductor optical amplifiers: amplification and signal processing“, Ph.D. Thesis, Dept. of Electromagnetic Systems, Technical University of Denmark, LD 114, 1995.

37 B. Mikkelsen. Ph.D. Thesis, Dept. of Electromagnetic Systems, Technical University of Denmark, 1994.

38 S. L. Danielsen. “Traffic analysis and signal processing in high-capacity optical networks”, Ph.D. Thesis, Dept. of Electromagnetic Systems, Technical University of Denmark, LD 133, 1998.

39 P. B. Hansen. “Optical packet switched networks”, Ph.D. Thesis, Dept. of Electromagnetic Systems, Technical University of Denmark, ISBN 87-90974-13-1, 1999.

40 T. Fjelde. “Traffic analysis and signal processing in optical packet switched networks”, Ph.D. Thesis, Research Center COM, ISBN 87-90974-21-2, Feb. 2002.

41 M. L. Nielsen. “Experimental and Theoretical investigation of Semiconductor Optical Amplifier (SOA) based All-Optical Switches”, Ph.D. Thesis, ISBN 87-90974-53-0, June 2004.

42 S. Bjørnstad. “Packet switching in optical networks”, Doctoral Thesis at Norwegian University of Science and Technology (NTNU), ISBN 82-471-6420-5, 2004.

43 H. Øverby. “Quality of Service Differentiation, Teletraffic Analysis and Network Layer Packet Redundancy in Optical Packet Switched Networks”, submission for Doctoral Thesis at Norwegian University of Science and Technology (NTNU), 2005.

44 V. B. Iversen. “Teletraffic Engineering and Network Planning”, COM course 34340, pp. 296-298, Technical University of Denmark, 2003.

45 A. Sudbø, and S. Bjørnstad. “Compensating the Penalty from Internal Blocking in an Asynchronous Optical Packet Switch with a Fibre Delay Line”, Proc. ECOC 2004, vol. 3, pp. 774-775, paper We4.P.153, (Stockholm, Sweden), 2004.

46 S. Bigo. “Multi-terabit/s optical transmission: where do we stand?”, 16th Norwegian Electro-Optics Meeting, (Tønsberg, Norway), (URL:www.unik.no/~eloptikk/submissions/sbigoNorway.pdf), 2004.

47 G. Charlet, E. Corbel, J. Lazaro, A. Klekamp, R. Dischler, P. Tran, W. Idler, H. Mardoyan, A. Konczykowska, F. Jorge, and S. Bigo. “WDM transmission at 6 Tbit/s capacity over transatlantic distance, using 42.7Gb/s differential phase-shift keying without pulse carver”, Techn. Digest OFC 2004, vol. 2, 764-766, paper PDP 36, (Los Angeles, CA, USA), 2004.

48 D. Z. Chen, G. Wellbrock, S. J. Penticost, D. Patel, C. Rasmussen, M. C. Childers, X. Yang, and M.Y. Frankel. “World’s first 40 Gbps overlay on a field-deployed, 10 Gbps, mixed-fiber, 1200 km, ultra long-haul system”, Techn. Digest OFC 2005, paper OTuH4, 2005.

49 LightReading. “Cisco launches HFR”, (URL:www.lightreading.com/document.asp?doc_id=53321), May 25th 2004.

50 C. Hullin, C. Gueritch, E. Grand, D. Lesterlin, S. Ruggeri, M. Adlerfligel, J. P. Blondel, V. Bouder, C. Trecasser, L. Curincks, E. Brandon, O. Courtois, and D. Filet. “Ultra long-haul 2500 km terrestrial transmission of 320 channels at 10 Gbit/s over C+L bands with 25 GHz wavelength spacing”, Proc. ECOC 2002, vol. 1, paper 1.1.3, (Copenhagen, Denmark), 2002.

51 D. J. Blumenthal. “Photonic packet switching and optical label swapping”, SPIE/Kluwer Optical Networks Magazine, pp. 54-64, Nov/Dec 2001.

52 N. Wada, K. Fujinuma, T. Wada, F. Kubota. “40 Gbit/s Packet Bit Error Ratio and Loss Real-time Measurement for Ultra-high Speed Packet Switched Network”, Proc. 9th IFIP/IEEE 2005 conference on Optical Network Design and Modelling (ONDM 2005) pp. 41-48, (Milan, Italy), 2005.

53 R. Walsworth, S. Yelin, and M. Lukin. “The Story Behind ‘Stopped Light’”, OSA Optics&Photonics News (OPN), 13(5), pp. 51-54, May 2002.

54 F. Callegati, G. Corazza, and C. Raffaelli. “Exploitation of DWDM for Optical Packet Switching With Quality of Service Guarantees”, IEEE Journ. Selected Areas in Communications, 20(1), pp. 190-201, 2002.

55 I. White, R. Penty, M. Webster, Y. J. Chai, and Adrian Wonfor, and S. Shahkooh. “Wavelength Switching Components for Future Photonic Networks”, IEEE Comm. Mag., 40(9), pp. 74-81, September 2002.

56 B. Lavigne, E. Balmefrezol, P. Brindel, B. Dagens, R. Brenot, L. Pierre, J. –L. Moncelet, D. de la Grandiere, J. –C. Remy, J.-C. Bouley, B. Thedrez, and O. B. Leclerc. “Low input power all-optical 3R regenerator based on SOA devices for 42.66 Gbit/s ULH WDM RZ transmissions with 23dB span loss and all-EDFA amplification”, Techn. Digest OFC 2003, PD15, (Atlanta, GA, USA), 2003.

57 M. Yoo, C. Qiao, and S. Dixit. “QoS performance in IP over WDM networks”, IEEE Journ. Selected Areas in Communications, 18(10), pp. 2062-2071, 2000.

58 G. Hu, K. Dolzer, and C.M. Gauger. “Does burst assembly really reduce the self-similarity?”, Techn. Digest OFC 2003, vol. 1, pp. 124-125, MF100, (Atlanta, GA, USA), 2003.

59 Z. Haas, “The 'staggering switch': an electronically controlled optical packet switch”, IEEE Journ. Lightwave Techn., 11(5/6), pp. 925-936, 1993.

60 F. Masetti, J. Benoit, F. Brillouet, J.M. Gabriagues, A. Jourdan, M. Renaud, D. Bottle, G. Eilenberger, K. Wunstel, M. Schilling, D. Chiaroni, P. Gavignet, J.B. Jacob, G. Bendelli, P. Cinato, P. Gambini, M. Puleo, T. Martinson, P. Vogel, T. Durhuus, C. Joergensen, K. Stubkjaer, R. Baets, P. Van Daele, J. C. Bouley, R. Lefevre, M. Bachmann, W. Hunziker, H. Melchior, A. McGuire, F. Ratovelomanana, and N. Vodjdani. ”High speed, high capacity ATM optical switches for future telecommunication transport networks”, IEEE Journ. Selected Areas in Communications, 14(5), pp. 979-998, 1996.

61 I. Chlamtac, A. Fumagalli, L. G. Kazovsky, P. Melman, W. H. Nelson, P. Poggiolini, M. Cerisola, A. N. M. Masum Choudhury, T. K. Fong, R. T. Hofmeister, C. Lu, A. Mekkittikul, D. J. M. S. Ix, C. Suh, and E. W. M. Wong. “CORD: contention resolution by delay lines”, IEEE Journ. Selected Areas in Communications, 14(5), pp. 1014-1029, 1996.

62 F. Callegati, M. Casoni, G. Corazza, C. Raffaelli, D. Chiaroni, F. Masetti, and M. Sotom. “Architecture and performance of a broadcast and select photonic switch”, Optical Fiber Technology, 4(3), pp. 266-284, 1998.

63 F. Callegati. “Optical buffers for variable length packets”, IEEE Communications Letters, 4(9), pp. 292-294, 2000.

64 L. Tančevski, S. Yegnanarayanan, G. Castanon, L. Tamil, F. Masetti, and T. McDermott. "Optical routing of asynchronous, variable length packets", IEEE Journ. Selected Areas in Communications, 18(10), pp. 2084-2093, 2000.

65 S. Bjørnstad, D. R. Hjelme, and N. Stol. "An optical packet switch design with shared electronic buffering and low bit rate add/drop inputs", Proc. IEEE International Conference on Transparent Optical Networks (ICTON 2002), vol.1, pp. 69-72, (Warzaw, Poland), 2002.

66 S. Bjørnstad, D.R. Hjelme, and N. Stol. "A scalable optical packet switch for variable length packets employing shared electronic buffering", Proc. ECOC 2002, vol. 3, P4.07, (Copenhagen, Denmark), 2002.

67 C. Develder, J. Cheyns, E. Van Breusegem, E. Baert, A. Ackaert, M. Pickavet, and P. Demeester. "Node architectures for optical packet and burst switching", Proc. Int. Topical Meeting on Photonics in Switching (PS2002), (invited) paper PS.WeA1, pp. 104–106, (Cheju Island, Korea), 2002.

68 F. Callegati and W. Cerroni, "Time-wavelength exploitation in optical feedback buffer with trains of packets", Proc. SPIE Optical Networking and Communications (OptiComm 2002), Vol. 4874, pp. 274-285, (Boston, MA, USA), 2002.

69 C. Develder, J. Cheyns, M. Pickavet, and P. Demeester. "Multistage Architectures for Optical Packet Switching Using SOA-Based Broadcast-and-Select Switches", Techn. Digest OFC 2003, vol. 2, pp. 794-795, paper FS3, (Atlanta, Ga, USA), 2003.

70 J. Cheyns, C. Develder, E. Van Breusegem, A. Ackaert, M. Pickavet, and P. Demeester. "Routing in an AWG-Based Optical Packet Switch", Kluwer Photonic Network Communications, 5(1), pp. 69-80, 2003.

71 J. Cheyns, E. Van Breusegem, A. Ackaert, M. Pickavet, and P. Demeester. "Scheduling window for AWG-based blocking optical switches", IEE Electronics Letters 39(6), pp. 546-547, 2003.

72 P. B. Hansen, S.L. Danielsen, and K.E. Stubkjaer, "Optical packet switching without packet alignment", Proc. ECOC 1998, vol. 1, pp. 591-592, pp. 591-592, (Madrid, Spain), 1998.

73 K. Claffy, G. Miller, and K. Thompson, "The nature of the beast: Recent traffic measurements from an Internet backbone", Proc. ISOC INET'98, 1998.

74 S. Keshav, and R. Sharma. "Issues and Trends in Router Design", IEEE Comm. Mag., 36(5), pp. 144-151, 1998.

75 B. Meagher, G. K. Chang, G. Ellinas, Y.M. Lin, W. Xin, T. F. Chen, X. Yang, A. Chowdhury, J. Young, S. J. Yoo, C. Lee, M. Z. Iqbal, T. Robe, H. Dai, Y.J. Chen, and W. I. Way. "Design and Implementation of Ultra-Low Latency Optical Label Switching for Packet-Switched WDM", IEEE Journ. Lightwave Techn., (18)12, pp. 1978-1987, 2000.

76 Sulur, T. Koonen, H. de Waardt, and I. Monroy. "IM/FSK Format for Payload/Orthogonal Labeling IP Packets in IP over WDM Networks Supported by GMPLS Based LOBS", Proc. 7th IFIP Working Conference on Optical Networking Design and Modelling (ONDM 2003), vol. 2, pp. 703-716, (Budapest, Hungary), 2003.

77 N. Wada, H. Harai, and W. Chujo. "Multi-hop, 40Gbit/s variable length photonic packet routing based on multi-wavelength label switching, waveband routing, and label swapping", Techn. Digest OFC 2002vol. 1, pp. 216-217, paper WG3, (Anaheim, CA, USA), 2002.

78 K. Kitayama, N. Wada, and H. Sotobayashi. "Architectural considerations for Photonic IP Router Based upon Optical Code Correlation", IEEE Journ. Lightwave Techn., 18(12), pp. 1834-1844, 2000.

79 K. Dolzer, C. Gauger, J. Späth, and S. Bodamer. "Evaluation of reservation mechanisms for optical burst switching", AEÜ Int. Journal of Electronics and Communications. 55(1), 2001.

80 E. Van Breusegem, J. Cheyns, B. Lannoo, A. Ackaert, M. Pickavet, and P. Demeester. "Implications of Using Offsets in All-Optical Packet Switched Networks", Proc. 7th IFIP Working Conference on Optical Networking Design and Modelling (ONDM 2003), vol. 2, pp. 1053-1072, (Budapest, Hungary), 2003.

81 N. Sahri, D. Prieto, S. Silvestre, D. Keller, F. Pommerau, M. Renaud, O. Rofidal, A. Dupas, F. Dorgeuille, and D. Chiaroni. "A highly integrated 32-SOA gates optoelectronic module suitable for IP multi-terabit packet routers", Techn. Digest OFC 2002, vol. 4, PD32 (Anaheim, CA, USA), 2002.

82 F. Masetti, D. Zriny, D. Verchère, J. Blanton, T. Kim, J. Talley, D. Chiaroni, A. Jourdan, J.-C. Jacquinot, C. Coeurjolly, P. Poignant, M. Renaud, G. Eilenberger, S. Bunse, W. Latenschleager, J. Wolde, and U. Bilgak. "Design and Implementation of a Multi-Terabit Optical Burst/Packet Router prototype", Techn. Digest OFC 2002, vol.1, FD11 - FD13, (Anaheim, CA, USA), 2002.

83 H. Buchta, E. Patzak, J. Saniter, and C. Gauger. “Limits of Effective Throughput of Optical Burst Switches Based on Semiconductor Optical Amplifiers”, Techn. Digest OFC 2003, vol. 1, pp. 215-217, paper TuJ3, (Atlanta, GA, USA), 2003.

84 S. Kamei, M. Ishii, M. Itoh, T. Shibita, and T. Kitagawa.“64 x 64-channel Uniform-loss and Cyclic-frequency Arrayed-waveguide Grating Router Module”, Proc. ECOC 2002, vol. 3, paper 6.2.6, (Copenhagen, Denmark), 2002.

85 D. J. Blumenthal, J. E. Bowers, L. Rau, L. Hsu-Feng Chou, S. Rangarajan, Wei Wang, and K. N. Poulsen. “Optical Signal Processing for Optical Packet Switching Networks”, IEEE Comm. Mag., 41(2), pp. S23-S30, 2003.

86 M. C. Chia, D. Hunter, I. Andonovic, P. Ball, I. Wright, S. Ferguson, K. Guild, and M. O'Mahony. “Packet Loss and Delay Performance of Feedback and Feed-Forward Arrayed-Waveguide Gratings-Based Optical Packet Switches With WDM Inputs-Outputs”, IEEE Journ. Lightwave Techn., 19(9), pp. 1241-1254, 2001.

87 C. M. Gauger, “Performance of Converter Pools for Contention Resolution in Optical Burst Switching”, Proc. SPIE Optical Networking and Communications (Opticomm), vol. 4874, pp. 109-117, (Boston, MA, USA), 2002.

88 A. Pattavina. “Multi-wavelength switching in IP optical nodes adopting different buffering strategies”, Elsevier Journal on Optical Switching and Networking, 1(1), pp. 65-75, (URL:www.sciencedirect.com/science/journal/15734277), 2005.

89 C. Develder, M. Pickavet, and P. Demeester. “Strategies for an FDL based feed-back buffer for an optical packet switch with QoS differentiation”, Proc. International Conference on Optical Internet (COIN 2002), paper COIN.TuD1, (Cheju Island, Korea), 2002,

90 M. Listanti, V. Eramo, and R. Sabella. “Architectural and Technological Issues for Future Optical Internet Networks”, IEEE Comm. Mag., (38)9, pp. 82-92, 2000.

91 L. Xu, H.G. Perros, and G. Rouskas. “Techniques for Optical Packet Switching and Optical Burst Switching”, IEEE Comm. Mag., (39)1, pp. 136-142, 2001.

92 C. Fenger, and V.B. Iversen. “Wavelength Conversion by using Multiple Fibres”, Proc. ECOC 2002, vol. 1, paper 2.4.2, (Copenhagen, Denmark), 2002.

93 F. Borgonova, L. Fratta, and J. Bannister. “Unslotted Deflection Routing in All-Optical Networks”, Proc. IEEE Global Telecommunications Conference 1993, vol 1, pp. 119-125, (Houston, TX, USA), 1993.

94 S. L. Danielsen, B. Mikkelsen, C. Joergensen, and T. Durhuus. “WDM packet switch architectures and analysis of the influence of tuneable wavelength converters on the performance”, IEEE Journ. Lightwave Techn., 15(2), pp. 219–227, 1997.

95 S. L. Danielsen, C. Jørgensen, B. Mikkelsen, and K.E. Stubkjær. ”Optical Packet Switched Layer without Optical Buffers”, IEEE PTL, (10)6, pp. 896-898, 1998.

96 S.B. Yoo, Y. Bansal, Z. Pan, J. Cao, V. K. Tsui, S. K. H. Fong, Y. Zhang, J. Taylor, H. J. Lee, M. Jeon, and V. Akella. “Optical-label based packet routing system with contention resolution in wavelength, time, and space domains”, Techn. Digest OFC 2002, pp. 280-282, paper WO2, (Anaheim, CA, USA), 2002.

97 J. S. Turner. “WDM Burst Switching for Petabit Data Networks”, Techn. Digest OFC 2000, vol. 2, pp. 47-49, (Baltimore, MD, USA), 2000.

98 V. Eramo, and M. Listanti. “Packet Loss in a Bufferless Optical WDM Switch Employing Shared Tunable Wavelength Converters”, IEEE Journ. Lightwave Techn., 18(12), pp. 1818-1833, December 2000.

99 S.Yao, B. Mukherjee, and S. Dixit. “Advances in Photonic Packet Switching; An Overview”, IEEE Comm. Mag., 38(2), pp. 84-94, 2000.

100 S. Bjørnstad, D.R. Hjelme, and N. Stol. ”A highly efficient optical packet switching node design supporting guaranteed service”, Proc. ECOC 2003, pp. 110-111, paper Mo 4.4.5, (Rimini, Italy), 2003.

101 M.C. Chia, D. K. Hunter, I. Andonovic, P. Ball, P, and I. Wright. ”Optical Packet Switches: A Comparison of Designs”, Proc. IEEE 8th International Conference On Networks (ICON 2000), pp. 365-369, (Singapore), 2000.

102 M. L. Nielsen, T. Fjelde, J. D. Buron, and B. Dagens. “All-optical bit-pattern recognition in data segments using logic AND and XOR in a single all-active MZI wavelength converter”, Proc. ECOC 2002, vol. 1, paper 1.4.6, (Copenhagen, Denmark), 2002.

103 F. Callegati, G. Muretto, C. Raffealli, P. Zaffoni, and W. Cerroni. “A framework for performance evaluation of OPS congestion resolution”, Proc. 9th IFIP/IEEE 2005 conference on Optical Network Design and Modelling (ONDM 2005), pp. 243-250, (Milan, Italy), 2005.

104 International Telecommunication Union. “ICT – Free Statistics”, (URL:www.itu.int/ITU-D/ict/statistics/), April 2004.

105 DSL forum.“2003 Global DSL Subscriber Chart”, (URL:www.dslforum.org/pressroom.htm), March 2nd, 2004.

106 J.S. Turner. “Terabit burst switching”, Journal of High Speed Networks, 8(1), pp. 3-16, 1999.

107 S. Rangarajan, Z. Hu, L. Rau, and D. J. Blumenthal. ”All-Optical Contention Resolution with Wavelength Conversion for Asynchronous Variable-Length 40 Gb/s Optical Packets”, IEEE PTL, 16(2), pp. 689-691, 2004.

108 X. Xiao, and L. M. Ni. “Internet QoS: A Big Picture”, IEEE Network, 13(2), 8-18, 1999.

109 M. J. O’Mahony, D. Simeonidou, D. K. Hunter, and A. Tzanakaki. “The Application of Optical Packet Switching in Future Communication Networks”, IEEE Comm. Mag., 39(3), pp. 128-135, 2001.

110 B. Wydrowski, and M. Zukerman. “QoS in Best-Effort Networks”, IEEE Comm. Mag., 40(12), pp. 44-49, 2002.

111 H. Øverby, and N. Stol, ”Quality of Service in Asynchronous Bufferless Optical Packet Switched Networks”, Kluwer Telecommunication Systems, 27(2-4), pp. 151-179, 2004.

112 H. Øverby, and N. Stol, “A Teletraffic Model for Service Differentiation in OPS networks”, Proc. 8th Optoelectronic and Communications Conference (OECC), vol. 2, pp. 677-678, (Shanghai, China), 2003.

113 Y. Chen, M. Hamdi, D. H. K. Tsang, and C. Qiao, “Providing Proportionally Differentiated Services over Optical Burst Switching Networks”, Proc. IEEE Global Telecommunications Conference, (San Antonio, TX, USA), 2001.

114 S. Bjørnstad, N. Stol, and D.R. Hjelme. “Quality of Service in optical packet switched DWDM transport networks”, Proc. Asia-Pacific Optical and Wireless Communications Conference (APOC) 2002, Proc. SPIE 4910, pp. 63-74, (Shanghai, China), 2002.

115 S. Yao, B. Mukherjee, S. J. Ben Yoo, and S. Dixit, “A Unified Study of Contention-Resolution Schemes in Optical Packet-Switched Networks”, IEEE/OSA Journ. Lightwave Techn., 21(3), pp. 672-683, 2003.

116 N. Christin. “Quantifiable Service Differentiation for Packet Networks”, Ph.D. Thesis, Faculty of the School of Engineering and Applied Science, University of Virginia, August 2003.

117 H. Øverby, and N. Stol. ”Providing QoS in Asynchronous Bufferless Optical Packet Switching Networks”, Kluwer Wireless Networks (to appear 2005).

118 T.V. Lakshman, and U. Madhow. “The Performance of TCP/IP for Networks with High Bandwidth-Delay Products and Random Loss”, IEEE/ACM Transactions on Networking, 5(3), June 1997.

119 S. Yao, J. B. Yoo, and B. Mukherjee. “A Comparison Study between slotted and unslotted all-optical packet-switched network with priority-based routing”, Techn. Digest OFC 2001, vol. 2, paper TuK2, (Anaheim, CA, USA), 2001.

120 H. Øverby, and N. Stol. “A Teletraffic Model for Service Differentiation in OPS Networks”, Proc. 8th Optoelectronic and Communications Conference (OECC 2003), vol. 2, pp. 677-678, (Shanghai, China), 2003.

121 H. Øverby, and N. Stol. “Effects of bursty traffic in service differentiated Optical Packet Switched networks”, OSA Optics Express, 12(3), pp. 410-415, 2004.

122 H. Øverby. “An Adaptive Service Differentiation Algorithm for Optical Packet Switched Networks”, Proc. 5th International Conference on Transparent Optical Networks (ICTON 2003), vol. 1, pp. 158-161, (Warsaw, Poland), 2003.

123 S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss et al. “Architecture for Differentiated Services”, IETF RFC 2475, (URL:http://www.ietf.org/rfc/rfc2475.txt), December 1998.

124 “Quality of service and performance: End-user multimedia QoS categories”, ITU-T Recommendation G.1010, 2001.

125 T. Jensen ”Network Planning- Introductory Issues”, Telektronikk (ed. O. Espvik), 99(3/4), pp. 9-46, ISSN 0085-7130, (URL:www.telektronikk.com), 2003.

126 F. Callegati, and C. Rafaelli, “End-to-End Delay Evaluation for an Optical Transparent Packet Network”, Kluwer Photonic Network Communication, 1(2), pp. 147-160, 1999.

127 T. N. Quynh, H. Karl, A. Wolisz, and K. Rebensburg, “Relative Jitter Packet Scheduling for Differentiated Services”, Proc. 9th IFIP Conference on Performance Modelling and Evaluation of ATM & IP Networks, pp. 139-151, (Budapest, Hungary), 2001.

128 Tiernan white paper, “Measuring and Controlling Jitter in Digital Video Transmission Systems”, (URL: www.tiernan.com/pdf/videocelljitter.pdf), May 3rd, 2001.

129 E. Blanton, and B. Allmann, “On Making TCP More Robust to Packet Reordering”, ACM SIGCOMM Computer Communication Review archive, 32(1), pp. 20-30, 2002.

130 T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido, ”A Nonstationary Poisson View of Internet Traffic”, Proc. IEEE INFOCOM, (Hong Kong, China), 2004.

131 S. Yao, S.J.B. Yoo, B. Mukherjee, and S. Dixit. “All-optical packet switching for metropolitan area networks: opportunities and challenges”, IEEE Comm. Mag., 39(3), pp. 142 -148, 2001.

132 M. Herzog, M. Maier and M. Reisslein, "Metropolitan area packet-switched WDM networks: a survey on ring systems”, IEEE Communications Surveys, 6(2), pp. 2-20, 2004.

133 C. Develder, R. V. Caenegem, E. Baert, M. Pickavet, and P. Demeester."Active versus passive OPS architectures for metro rings: network dimensioning point of view”, Proc. ECOC 2003, paper We 1.4.4., (Rimini, Italy), 2003.

134 K. Bengi. “Access Protocols for an Efficient Optical Packet-Switched Metropolitan Area Ring Network Supporting IP Datagrams”, Proc. 11th International Conference on Computer Communications And Networks (ICCCN’02), pp. 284-289, (Miami, FL, USA), 2002

135 J. Cai, A. Fumagalli, and I. Chlamtac. “The Multitoken Interarrival Time (MTIT) Access Protocol for Supporting Variable Size Packets Over WDM Ring Network”, IEEE Journ. Selected Areas in Communications, 18(10), pp. 2094-2104, 2000.

136 K. Shrikande, A. Srivatsa, I. M. White, M.S. Rogge, D. Wonglumsom, S. M. Gemelos, and L.G. Kazovsky. “CSMA/CA MAC Protocols for IP-HORNET: An IP over WDM Metropolitan Area Ring Network”, Proc. IEEE Global Telecommunications Conference, vol. 2, pp. 1303–07, (San Francisco, CA, USA), 2000.

137 A. Bianco, M. Bonsignori, E. Leonardi, and F. Neri. “Variable-size Packets in Slotted WDM Ring Networks”, Proc. 6th IFIP Working Conference on Optical Networking Design and Modelling (ONDM 2002), pp. 179-198, (Torino, Italy), 2002.

138 I. M. White, M.S. Rogge, K. Shrikande, and L.G. Kazovsky. “A Summary of the HORNET Project: A Next Generation Metropolitan Area Network”, IEEE Journ. Selected Areas in Communications, 21(9), pp. 1478-1494, November 2003.

139 W.-P. Chen, and W.-S. Whang. “A packet pre-classification CSMA/CA MAC protocol for IP over WDM ring networks”, Proc. 8th IEEE International Conference on Communication Systems (ICCS), vol. 2, pp. 1217-1222, (Singapore), 2002.

140 S. Bjørnstad, D. R. Hjelme, and N.Stol. ”Asynchronous feedback buffer with reduced self-induced contention for optical packet switches”, Proc. 9th IEICE OptoElectronics and Communications Conference/ 3rd Conference on Optical Internet (OECC/COIN 2004), pp. 16-17, (Yokohama, Japan), 2004.

141 R. Gaudino, A. Carena, V. Ferrero, A. Pozzi, V. De Feo, P. Gigante, F. Neri, and P. Poggiolini. “RINGO: a WDM ring optical packet network demonstrator”, Proc. ECOC 2001, vol. 4, pp. 620-621, (Amsterdam, The Netherlands), 2001.

142 D. Dey, A. van Bochove, A. Koonen, D. Geuzebroek, and M. Salvador. “FLAMINGO: A Packet-switched IP-over-WDM All optical MAN”, Proc. ECOC 2001, vol. 3, pp. 480-481, (Amsterdam, The Netherlands), 2001.

143 A. Ge, F. Callegati, L. and S. Tamil. “On Optical Burst Switching and Self-Similar Traffic”, IEEE Comm. Lett., 4(3), pp. 98- 100, March 2000.

144 M. A. Marsan, A. Bianco, E. Leonardi, M. Meo, and F. Neri, MAC protocols and fairness control in WDM multirings with tunable transmitters and fixed receivers, IEEE Journ. Lightwave Techn., 14 (6), pp. 1230–1244, June 1996.

145 S. K. Kim, H. Okagawa, K. Shrikande, and L. G. Kazovsky. “Unslotted Optical CSMA/CA MAC Protocol with Fairness Control in Metro WDM Ring Networks”, Proc. IEEE Global Telecommunications Conference 2002, (Taipei, Taiwan), 2002.

146 A. Bianco, J. M. Finochietto, G. Galante, F. Neri, and V. Sarra. “Scheduling Variable-Size Packets in the DAVID Metropolitan Area Network”, Proc. IEEE International Conference On Communications (ICON 2004), pp. 1750-1754, (Paris, France), 2004.

147 R. Jain, D. Chiu, and W. Hawe. “A quantitative measure of fairness and discrimination for resource allocation in shared computer systems”, Technical Report DEC-TR-301, Digital Equipment Corporation, Sept. 1984.

148 K. Bengi, and H. R. van As. “QoS Support and Fairness Control in a Slotted Packet-Switched WDM Metro Network”, Proc. IEEE Global Telecommunications Conference 2001, pp. 1494-1499, (San Antonio, TX, USA), 2001.