Optical packet switching with distributed control for high performance data center networks

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus, prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties in het openbaar te verdedigen op woensdag 18 december 2013 om 16.00 uur

door

Stefano Di Lucente

geboren te Rome, Italië
Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter: prof.dr.ir. A.C.P.M. Backx
1e promotor: prof.dr. H.J.S. Dorren
copromotor: dr. N. Calabretta
leden: prof. D. Simeonidou PhD (University of Bristol)
        prof.dr. G. Cincotti (Roma Tre University)
        prof.dr.ir. O.J. Boxma
        prof.dr.ir. P. Demeester (Universiteit Gent)
        prof.dr. K.A. Williams
A catalogue record is available from the Eindhoven University of Technology Library.

Optical packet switching with distributed control for high performance data center networks

Author: Stefano Di Lucente  
Eindhoven University of Technology, 2013. 
ISBN: 978-90-386-3523-1 
NUR: 959

Keywords: Optical packet switching / data center / distributed control / large port-count

The work described in this thesis was performed in the Faculty of Electrical Engineering of the Eindhoven University of Technology and was financially supported by the Netherlands Organization of Scientific research (NWO) and the Netherlands Technology Foundation (STW) through the VI and NRC Photonics programs.

Copyright © 2013 by Stefano Di Lucente

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the prior written consent of the author.

Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress

Typeset using \LaTeX
...to my niece Sara and my nephew Marco...
Summary

Optical Packet Switching with Distributed Control for High Performance Data Center Networks

Data Centers (DCs) are facing the rapid growth of data traffic due to the increasing demand of emerging bandwidth-hungry internet services. Typically, the intra-DC network is based on a fat-tree architecture. This network topology is easy scalable and failure tolerant due to the redundant interconnections between the tree layers. However, it is affected by communication bottlenecks in the server interaction process caused by the bandwidth subscription between each layer and the upper one. Moreover, in a common DC the employment of optical technology is limited to point-to-point links. Small-factor pluggable transceivers (SFPs) are employed for server-to-switch and switch-to-switch interconnections, while the switching process is performed by electronic packet switches. As a consequence, optical-electrical-optical (O-E-O) conversions and the intense use of electronic buffers are unavoidable and limit the power efficiency and the minimum latency within these systems. Each packet travelling in the DC network experiences multiple queueing delays due to the store-and-forward processing required at each switching node.

The employment of photonic technology in the switching process of a DC environment is attractive for solving these issues. An easy controllable, scalable, large port-count optical packet switch (OPS) would allow flattening the intra data center topology. It would improve the DCs performance in terms of end-to-end latencies and power consumption, allowing larger bandwidths, more efficient resources sharing among the servers and avoiding the costly and power-hungry O-E-O conversions. Despite the intense
research on this topic, an OPS capable of handling high data-rate, scaling to a large number of ports, while being easily controllable and adding negligible latency to the system does not exist. The main work presented in this thesis aims at investigating and implementing an optical packet switching system suitable for high performance data center networks.

Scaling OPSs to a large number of input/output ports while providing high bandwidths and low latency is an open question. Besides technology impairments such as splitting losses, that limit the number of possible interconnections or the lack of practical photonic buffers, that makes the store and forward process impossible in the optical domain, controlling a large number of ports while providing low configuration times is not trivial. In fact, the commonly employed OPS architectures, like Beneš or Banyan, require a centralized controller. As a consequence, the switch configuration time scales at least linearly with the number of ports. Therefore, building an OPS with a large port-count based on these architectures will not meet the system end-to-end latency requirements. A simple analytic model that allows computing end-to-end latency and throughput as function of the employed OPS architectures is presented. It is shown that optical interconnects that employ a centralized controller cannot scale over thousands of ports while providing the sub-microsecond end-to-end latencies required in some DC applications, due to their high reconfiguration time. As a consequence, a novel modular WDM OPS architecture is presented. It is a strictly non-blocking architecture with contention resolution based on wavelength conversion. WDM is employed to scale the number of logical interconnections of the architecture that results to be equal to the number of input fibers multiplied by the number of WDM channels carried by each fiber. The OPS operates transparently in the optical domain. However, electronic buffering is required at the edges of the switching matrix. Highly distributed control is enabled by the modular structure of the proposed architecture and by the optical in-band packet labeling and parallel label processing techniques employed. As a consequence, the switch reconfiguration time is port-count independent. This results in nanosecond scale switch reconfiguration time, that aims at reducing the total end-to-end latencies in a DC environment.

The employment of the architecture in a computer communication network is first investigated by means of simulations. System performance in terms of latency, packet loss and throughput are reported. It is shown that
sub-microsecond latencies, low packet loss and high throughputs are achievable under various traffic patterns in both synchronous and asynchronous operations and considering fixed and variable packet sizes. The effects of the electronic buffer capacity on the system performance are studied. The results reveal that the use of larger buffers does not improve the system performance under heavy traffic load conditions. The simulations results in terms of throughput are in agreement with the results obtained employing the analytic model, while some discrepancies are visible in the latency results. The difference in latency is caused by the fact that the model does not consider the buffer capacity, a factor that strongly affects the latency of the system.

The implementation of a single photonic module of the OPS architecture in combination with the packet labeling process, the label processor and the switch controller is then investigated in an experimental set-up. A $1 \times 8$ optical switch based on discrete optical components is employed as main building block of the OPS module. Label processor and switch controller are implemented programming an FPGA board. Error free operation is shown for 8 WDM channels at 40 Gb/s with 1.6 dB penalty, while only 25 ns are required for the overall switching process. Subsequently, an InP integrated $1 \times 4$ optical cross-connect is employed to investigate the possible integration of the OPS. Error free operation is shown for 4 WDM channels at 40 Gb/s with 4 dB penalty.

Two distinct flow control implementations are reported. This functionality, required in the case that packet contentions cannot be solved by wavelength conversion, enables the packet retransmission. An FPGA-based flow control implementation is successfully tested in an experimental set-up. Packet flow control and packet retransmission are experimentally investigated showing that a limited input (electronic) buffer capacity ensures packet loss lower than $10^{-5}$ for loads up to 0.5. However, this method requires a dedicated electrical network for being implemented. To avoid the issue related to the need of an extra interconnection network in a large-scale system such as a DC, an all-optical solution is presented. The all-optical flow control implementation is based on the re-utilization of a portion of the optical power of the extracted in-band label to generate an acknowledge message. The optical acknowledgement, after being modulated by a semiconductor optical amplifier (SOA) driven by the switch controller, is sent back to the transmitting node on the same optical link used for the packet
transmission. Error free operation is shown with only 0.5 $dB$ penalty. The all-optical solution is robust to possible power fluctuations.

Further integration of the OPS architecture is shown by the employment of a Silicon-on-Insulator (SOI) integrated label extractor and by the employment of a SOI integrated WDM channels demultiplexer and label extractor. The first device is based on cascade microring resonators. Error free operation with 1 $dB$ penalty is achieved after the extraction of three in-band optical labels from a 160 $Gb/s$ packet payload. Extracted labels are successfully recovered, processed by the FPGA-based controller that correctly drives the OPS module. The SOI integrated WDM channels demultiplexer and label extractor features a $1 \times 32$ array waveguide grating (AWG) and a microring resonator at its input. Three WDM channels at 40 $Gb/s$ are successfully demultiplexed and error free operation with 0.5 $dB$ penalty is shown for both packet payload and label.

The simulations results reported in this thesis reveal that the novel OPS architecture could provide promising results in terms of latency, throughput and packet loss if employed in a DC environment. According to the simulations, sub-microsecond latency can be achieved in the simulated systems under relatively high traffic conditions. This is possible due to the modular structure of the architecture, that enables high distributed control.

Nanosecond switch configuration time is shown in the experiments carried out in this work. The $1 \times 4$ subsystem and the $1 \times 8$ subsystem, which correspond to a $16 \times 16$ and to a $64 \times 64$ OPS based on the novel architecture, require the same configuration time, confirming the switch architecture allows port-count independent configuration time.

The integrated $1 \times 4$ cross-connect, the SOI integrated label extractor and the integrated channel demultiplexer and label extractor employed in this work partially demonstrate the possible integration of the architecture. The scalability and the total integration of the novel OPS architecture are discussed in this thesis.

Further research is required to realize a prototype that can be tested in a real DC scenario. Nonetheless, the results provided in this dissertation are promising for the realization of an integrated WDM OPS that can be successfully employed in an intra-DC network.
# Contents

Summary  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>1.1</td>
<td>Datacenters</td>
</tr>
<tr>
<td>1.1.1</td>
<td>Datacenter architecture</td>
</tr>
<tr>
<td>1.1.2</td>
<td>A future scenario</td>
</tr>
<tr>
<td>1.1.3</td>
<td>Costs of a datacenter</td>
</tr>
<tr>
<td>1.2</td>
<td>Optics in datacenters</td>
</tr>
<tr>
<td>1.3</td>
<td>Optical datacenter: requirements and challenges</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Datacenter traffic features</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Lack of optical memories</td>
</tr>
<tr>
<td>1.3.3</td>
<td>Latency</td>
</tr>
<tr>
<td>1.3.4</td>
<td>Photonic integration</td>
</tr>
<tr>
<td>1.4</td>
<td>Optical packet switches: state of the art</td>
</tr>
<tr>
<td>1.5</td>
<td>Preview of the thesis</td>
</tr>
<tr>
<td>1.6</td>
<td>Novel contributions</td>
</tr>
<tr>
<td>2</td>
<td>Novel OPS architecture</td>
</tr>
<tr>
<td>2.1</td>
<td>System under investigation</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Architectures control issue</td>
</tr>
<tr>
<td>2.2</td>
<td>Single-chain model</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Throughput</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Latency</td>
</tr>
<tr>
<td>2.3</td>
<td>Novel architecture</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Ideal CRB implementation</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Practical CRB implementation</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Packet loss performance in a buffer-less system</td>
</tr>
</tbody>
</table>
# CONTENTS

2.4 Summary ........................................... 38

3 Performance investigation .......................... 39
3.1 Simulation environment ............................. 39
3.2 Synchronous operation ............................ 43
  3.2.1 Simulations vs Model .......................... 47
  3.2.2 Buffer sizing ................................... 47
3.3 Asynchronous operation ............................ 49
  3.3.1 Self-similar traffic .............................. 49
  3.3.2 Latency, throughput and packet loss ........... 52
  3.3.3 Buffer sizing ................................. 53
3.4 Discussion ......................................... 55
3.5 Summary ............................................ 57

4 Dynamic switching .................................... 59
4.1 Labeling technique ................................. 60
4.2 8-port optical module subsystem ................. 62
  4.2.1 Experimental results ......................... 64
4.3 Integrated 4-port optical module subsystem .... 67
  4.3.1 Integrated optical module ...................... 68
  4.3.2 Experimental results ......................... 69
4.4 Discussion ......................................... 73
4.5 Summary ............................................ 76

5 Flow control implementation ......................... 77
5.1 FPGA-based flow control ......................... 78
  5.1.1 Experimental results .......................... 80
5.2 All-optical flow control .......................... 84
  5.2.1 Experimental results .......................... 85
5.3 Discussion ......................................... 89
5.4 Summary ............................................ 90

6 Towards the total OPS integration .................. 93
6.1 Introduction ........................................ 94
6.2 Integrated optical switch ......................... 95
6.3 Integrated label extractor ......................... 96
  6.3.1 Dynamic label extraction ...................... 97
  6.3.2 Experimental results ......................... 99
CONTENTS

6.4 Integrated demultiplexer and label extractor 102
   6.4.1 Dynamic demultiplexing and label extraction 103
   6.4.2 Experimental results 104
6.5 Discussion 106
   6.5.1 Label detection 107
   6.5.2 Label processor and switch controller 107
   6.5.3 Fixed wavelength converter 108
   6.5.4 Conclusions 108
6.6 Summary 109

7 Conclusions 111

Bibliography 119

Acronyms 131

List of Publications 135

Acknowledgments 139

Curriculum Vitæ 141
Chapter 1

Introduction

1.1 Datacenters

Over the last few years, emerging services like high-definition video streaming [1], social networking [2, 3] and cloud computing [4] in combination with the large diffusion of mobile devices such as smart phones and tablets [5] have contributed to the rapid growth of the Internet traffic.

On the one hand, optical technology, capable of providing large communication bandwidth, is deeply employed in nowadays global communication system to face the continuously increasing data exchange. The entire world is connected by optical cables across oceans and continents and optical links with aggregate capacities up to 1 $Tb/s$ represent the backbone of the Internet network. Moreover, the increasing deployment of fiber-to-the-home (FTTH) technology [6] in the access network, replacing the existing copper cables, extends the access data-rate up to 1 $Gb/s$, answering to the users high-speed connection demand.

On the other hand, optical technology plays a marginal role in the systems that make the emerging services available to the constantly growing user population: the datacenters.

A datacenter (DC) physically houses miscellaneous equipment, such as servers (web servers, application servers and database servers), switches and data storage devices, interconnected by a complex wired network. The enormous hardware and software resources available in this facility allows providing numerous services. In general, it can be said that DCs store,
manage, process and exchange digital data and information.

The huge volumes of packetized data travelling to and from the DCs, generated to satisfy users requests, represent only a small fraction of the total traffic handled by these systems. Cisco Global Cloud Index study [7] reveals that more than 75% of the traffic handled by a DC is internal traffic. Huge amount of flows are generated to spread the workload among the servers for parallel processing and to read/write, replicate and back up data to and from the storage facilities. The DC-to-users traffic contributes only for 7% of the total data exchange. This results in a growing factor of 1000 every 10 years in the DCs bandwidth requirements [8]. As a consequence, DCs built by the major service providers, have reached massive proportions to sustain the users service demand.

1.1.1 Datacenter architecture

A large-scale DC, addressed also as a warehouse-scale computer (WSC) [9], hosts over 100000 servers. The typical intra-DC network is a canonical fat-tree architecture, as depicted in Figure 1.1.

![Figure 1.1: Typical DC fat-tree topology network](image)

Servers are arranged in racks and several racks are grouped together into clusters. Servers of the same rack are interconnected by so-called top-of-the-rack (TOR) switches, while inter-racks and inter-clusters communications
Datacenters are handled by cluster switches at the second hierarchical level of the DC architecture. Depending on the size of the DC, an extra aggregation layer, which employs high data-rates switches, may be present on top of the cluster level.

This architecture has two main advantages: it can be easily scaled and it is fault tolerant (a TOR switch is usually connected to two or more cluster switches). However, this network topology presents also some drawbacks. A high number of interconnection links is required. Moreover, the switching process is performed by electronic packet switches, while server-to-switch and switch-to-switch links rely on optical technology. This translates in high power consumption of the TOR and cluster switches, mainly caused by the optical-to-electrical (O-E) and electrical-to-optical (E-O) transceivers and by the power consumed by the electronic switch fabrics like crossbar switches, buffers and high speed serializer/deserializer (SerDes). For the same reason, the system is affected by high latency. Servers placed in different racks are affected by communications lags. Packets travelling across the DC network experience multiple O-E and E-O conversions and the latency is affected by the implicit queueing time due to the store-and-forward processing at each switching node. Moreover, there is a communication bottleneck between the different levels of the DCs fat-tree architecture. Typically, a rack hosts up to 40 servers (in the form of blades), while a commonly employed TOR switch features 48 ports. It is clear that only a portion of the TOR switches bandwidth is available for inter-rack communications (the typical subscription ratio is 1/5), limiting the possibility of communication bandwidth between different racks or clusters. A similar limitation exists between the cluster and the aggregation layers. This issue is more evident during the communications between servers placed at the edges of the bottom layer of the current DC topology since the transmitted packets have to climb up and down the fat-tree architecture to reach their destination.

1.1.2 A future scenario

As the DCs continue to increase to sustain the emerging services, which require intense interactions among the servers and increasing computational power, more efficient interconnections are required.

Theoretically, flattening the intra-DC network, providing full bidirec-
tional bandwidth between the servers, will solve all the aforementioned issues of the current employed fat-tree topology and will maximize the DC resources sharing and performance. Being able to connect each server to any other will also minimize the communication latency and drastically reduce the number of the needed wired interconnections and networking devices. In this scenario, a centralized switch with hundreds of thousands of ports will be required. As will be explained later in this chapter, this is not a practical solution since the bandwidth requirements and the links utilization vary significantly within the actual DC network.

Nonetheless, the employment of high port-count optical packet switches (OPSs) in the switching process may represent a valuable solution, providing higher bandwidth, lower latency and reduced energy consumption.

In the vision of the author, a high data-rate thousand-port optical packet switch could be used to flatten the intra-DC network at the cluster layer, as shown in Figure 1.2.

![Figure 1.2: A possible future scenario of the intra-DC network employing a centralized OPS at the cluster layer.](image)

Such architecture would lead to significant improvements in the DC performance in terms of available communication bandwidth, end-to-end latency and would contribute to reduce the energy-related costs of the DC network. In fact, switching the DC traffic in the optical domain would allow avoiding the power-hungry E-O and O-E conversions required by the current employment of electronic packet switches. Flattening the DC network, even only at the cluster level, would improve the resource sharing between DC hardware, increasing the number of devices reachable in very
Datacenters

low latency through an extended connection map. Moreover, employing such a centralized cluster switch, all racks, and thus servers, would result at the same logical distance from each other and could be equally reached with any difference in the transmission latency, improving the DC resources and workload distribution.

1.1.3 Costs of a datacenter

In the recent years major cloud service companies, such as Facebook, Google and Apple, have made large investments for the realization of massive DCs. Besides the building costs and the initial costs for hardware and software purchasing and installation, running a large scale DC is mainly a power consumption matter.

In [10], the US DCs have been estimated to contribute as much as 1.5% of the 2006 energy consumption in the country with a total of 61 billion kWh consumed, corresponding to a cost of over 4.5 billions dollars. DCs have been reported by Greenpeace [11] to have demanded as much as 330 billions kWh in 2007 and this value was predicted to be over 1000 billions kWh by the 2020. Although, this estimation is most likely not accurate, since DC power consumption growth rate is slowing down, as explained in [12], the total power consumed in these systems represents a significant portion of the world energy demand (between 1.1% and 1.5%).

In [13] the total power consumption of a Google DC in 2007 is divided as follows. The massive number of servers consumes 45% of the total power. 25% of the power is dissipated by the power distribution and the cooling system, while 15% is due to utility losses. The last 15% is associated with the energy consumed by the network.

Reducing thermal management costs is of increasing interest [14]. About this topic, it is interesting to see how the service providers chose their DC location taking into account the local climate. Facebook datacenter in Sweden [15] and Google datacenter in Finland [16] are valid examples.

Even if the network does not have the largest contribution on the power consumption, more efficient networking and innovation in this sector are the key to reduce the energy-related costs of a DC. The employment of optical technology in the switching process could provide higher communication bandwidths, lower latency, while reducing considerably the power
consumptions [17–20]. According to the study in [21], the employment of optical interconnects within a DC would provide, over a period of 10 years, energy savings for over 150 million dollars.

A clear requirement for the employment of optical switches in a DC network is that such devices should provide limited power consumption, limiting the networking costs.

### 1.2 Optics in datacenters

In current DCs, optical technology is employed only for point-to-point interconnections between servers and TOR switches and between TOR switches and cluster switches. In particular, short optical links are based on cost-effective multi-mode fibers (MMFs). Servers and TOR switches are interconnected by Small Form-factor Pluggable transceivers (SFPs) at 1 $Gb/s$, while TOR-cluster switches communications relies on SFPs+ at 10 $Gb/s$. Due to the short-reach distances required, low-power and inexpensive vertical cavity surface emitting lasers (VCSELs) at 850 $nm$ are employed. Higher bandwidth transceivers are already available, such as $4 \times 10 \ Gb/s$ QSFPs (Quad-SFPs). Due to modal dispersion, VCSEL-based transceivers have limited reach-bandwidth product. The maximum reach at 10 $Gb/s$ is around 300 $m$. The maximum reach decreases if higher data-rate is considered. The choice on the employed optical technology is mainly driven by the low costs and power consumption of the devices, due to the large number of transceivers required. In a DC, distances beyond 300 $m$ at 10 $Gb/s$ are covered by more expensive and higher-power transceivers based on distributed feedback (DFB) lasers, in combination with single-mode fibers (SMF). DFB-based transceivers can be employed to increase interconnection bandwidth and reach at the expense of higher power and costs.

As the DC bandwidth requirements continue to increase, the employment of wavelength division multiplexing (WDM) represents a valid approach to sustain the DC traffic volumes, in particular for long links as the ones at the clusters layer [22]. WDM is widely employed in metro and long-haul transmissions. In a WDM system, multiple high data-rate wavelength channels co-propagate within the same optical SMF. Experimental results achieved transmission of over 100 $Tb/s$ in a single SMF [23]. Basically, a WDM transceiver operates with multiple transmitter-receiver
pairs on different WDM channels contained in the same optical fiber. In a DC environment this will result in an enormous advantage in scaling the bandwidth of the network. Once the SMF is deployed, as the number of wavelength channels or the data-rate per channel is increased the network bandwidth is greatly enhanced. To increase the network bandwidth in a MMF-based system, more parallel fibers have to be rolled out, increasing the number of interconnections and the installation costs.

WDM is considered too costly for the data-rate currently in use in DCs. However, it may represent the only possible solution when scaling further the bandwidth (e.g. 400 Gb/s) in particular for the aggregated traffic at the clusters layer, and for distances beyond 300 m to be bridged in larger DCs.

It may be expected that the intra-DC network follows the same evolution of the telecommunication networks. Long-haul networks and metro area networks (MAN’s) have evolved from traditional opaque networks into all-optical networks. This evolution was driven by the issues caused by the rapid growth of those networks. In an opaque network, the optical signals undergo optical-to-electrical-to-optical (O-E-O) conversions at each routing node. This translates, for continuously increasing bandwidths and distances, in higher costs, heat dissipation, power consumption and high operation and maintenance costs. On the contrary, all-optical networks, in which optical cross-connects and reconfigurable optical add/drop multiplexers (ROADM) are employed, provide higher bandwidth, reduced power consumption and reduced operation costs [24].

1.3 Optical datacenter: requirements and challenges

To realize a more efficient intra-DC network (in terms of bandwidth, power consumption, latency and interconnection capabilities between the DC resources) optical packet switches (OPSs) have captured the attention of many research groups.

The advantage of using a transparent optical network and the issues caused by the currently employed fat-tree architecture have already been discussed. Flattening the DC architecture employing an OPS capable of optically interconnecting the massive number of servers of a WSC would
solve the fat-tree architecture issues and provide the benefits of a transparent optical network.

In principle, an all-to-all interconnection network between the servers of a DC is the best solution to maximize the resources sharing and minimize the latency of the systems. Interconnecting all the servers of a DC will require a centralized OPS with hundreds of thousands of ports. However, this is not a practical solution. In fact, the bandwidth requirements of the actual DC architecture differ as function of the considered hierarchical level. A clear understanding of DC traffic characteristics is a crucial point for designing an efficient and cost-effective intra-DC network.

1.3.1 Datacenter traffic features

DCs offer a large variety of services and in general three distinct classes of DC can be identified: private DCs, university campus DCs and cloud computing DCs. Traffic within all these systems shows some common features (average packet size, for example), while other characteristics (as applications or traffic flows) are quite different as function of the environment.

There are few publications available on this topic based on measurements of real DCs [25–27]. The most interesting findings for the network design are the following. A traffic flow is defined as an established connection between two servers.

- Traffic flow locality: describes the direction of the traffic. Packets generated by a server may be directed to servers of the same rack (intra-rack flow) or to servers placed in a different rack (inter-rack or inter-cluster flow). This traffic feature is environment-dependent. In educational DCs intra-rack traffic ranges between 10% to 40% of the total traffic, while it’s up to 80% in cloud computing DCs. However, inter-rack and inter-cluster traffic (also known as east-west traffic) is continuously growing [28]. In WSCs, it is at the cluster level where the most traffic is confined [9]. In case of high inter-rack traffic, high bandwidth is required between racks while cost-effective commodity switches can be employed for intra-rack communications.

- Link utilization: in all DC classes, the link utilization inside the rack and between racks is quite low. For 95% of the operations the bandwidth utilization is below 10% and it is below 30% for 99% of the
time. It sounds reasonable that the 1 Gb/s links employed to interconnect the servers of the same rack would be able to sustain also the future traffic demand. Thus, intra-rack communications are not the main issue in a DC environment. On the contrary, the utilization of inter-cluster links and at the aggregation layer is high and higher bandwidth is a near requirement.

- Traffic flow size and duration: the majority of the traffic flows are small (≤ 10 kB) and last for few hundreds of milliseconds. This traffic feature affects the choice of the network architecture and the choice of the switch fabric. If most of the traffic has a long duration (seconds or tens of seconds), then an optical device with high reconfiguration time could be employed in the switching process. This is not the case for a DC environment.

- Packet size: DC traffic shows a bimodal distribution around the minimum and the maximum Ethernet frame size (40 and 1500 bytes). This phenomenon is due to the fact the packets are generally small control packets or portions of large files that are exchanged between the DC storage facilities.

These empirical results reveal that it is not necessary to interconnect all the servers of a DC in a flat network. However, even interconnecting only the clusters of a WSC by means of an OPS, a large number of switch input/output ports is required. In addition, the traffic characteristics make the employment of a fast reconfigurable OPS a necessity. Those are not the only challenges in implementing an optical DC network.

1.3.2 Lack of optical memories

The employment of an OPS that transparently switches packets in the optical domain avoiding O-E-O conversion is very challenging. Electronic packet switches employ input and output buffering to perform the store-and-forward processing needed at each switching node. This is impossible to implement in the optical domain, for the simple reason that an optical random access memory (RAM) does not exist.

Two approaches are commonly used to overcome the lack of optical RAM. Several studies [29, 30] rely on the employment of fiber delay line based buffers to store signals in the optical domain. Basically, in case of
contention at the switch, one of the packet contending the output port is switched to a fiber delay line (FDL). After a fixed time delay, dependent on the length of the used FDL, the packet is fed again at the input of the switch. However, this approach increases significantly the complexity of the switch architecture and its control. As a consequence, extra-latency is introduced in the system. Moreover, this method affects the signal power and the fact that the packet can be delayed only for fixed delay times does not solve the problem completely.

The second approach consists in confining the buffering process at the edges of the switching network [31]. In this case the buffering is performed in the electrical domain by employing commonly used electronic buffers, while the switching process is kept transparent in the optical domain. The employment of an OPS with fast reconfiguration times becomes important to reduce the needed buffer capacity. Moreover, this method can be applied only in systems that feature short host-switch distances. In case of contentions, in fact, the packets stored at the edge of the switch need to be retransmitted. This process affects the system latency, that is basically dependent on the number of contentions and on the link length. Fortunately, intra-DC networks feature links that range from few to hundreds meters.

The employment of WDM technology would benefit the DC network also in this context. In fact, it would enable the opportunity to reduce the packet contentions exploiting the optical domain. Wavelength converters (WCs) could be used to solve contentions “on-the-fly” in the optical domain. In other words, since many wavelength channels can co-propagate in a single optical fiber, converting the contended packet to a different and available wavelength would reduce the number of contentions and increase the system throughput. High data-rate WCs have already been demonstrated [32].

1.3.3 Latency

The actual DC network is affected by high latency due to the store-and-forward processing needed at each routing node of the network. These delays are dependent on the physical position of the packets destination target. Inter-rack communications delays can be in the 100 µs order while few more hundreds µs have to be added for inter-cluster communication delays [9].

Besides the intrinsic advantage in speed provided by the use of optical
signals over electronics, the choice of the switch architecture and technology plays an important role in this context. For example, switching fabrics that features high configuration times, as optical micro electro-mechanical systems (MEMS) based switch, are not suited for optical packet switching in DC networks.

Summarizing the previous considerations, the desired optical packet switch has not only to handle a large number of high data-rate input/output ports, but needs to add negligible latency to the intra-DC communication network.

1.3.4 Photonic integration

Limited power consumption is one of the requirements to employ an OPS in a intra-DC network. Reducing the optical switch footprint will result in diminished power consumption and reduced operation and fabrication costs. The only approach to reach this goal is by means of photonic integration.

Photonic integration technology enables the opportunity to incorporate multiple photonic functions on a single chip. The main advantages of fabricating photonic integrated circuits (PICs) over the use of discrete optical components are: reduced system footprint, reduced power consumption, reduced fiber coupling losses and reduced packaging costs, since many components can be co-fabricated on a single device.

Despite the multi-decennial research in this field, few market-oriented PICs are available. The reason of this low diffusion can be found in the lack of a fabrication standard [33]. To overcome this issue, fabrication platforms, that employ a small set of standardized technologies, have been developed [34, 35].

Unlike electronic integrated circuits, where silicon is the dominant material, PICs are being fabricated employing heterogeneous materials. III-V semiconductors and in particular Indium Phosphide (InP) [36], silica on silicon, namely photonic lightwave circuit (PLC) [37] and silicon on insulator (SOI) [38] are the most used material systems.

Each material can provide different advantages and limitations depending on the desired photonic functionality to integrate. Low losses and low thermal sensitivity make PLC suitable for the integration of passive photonic circuits. For instance, PLC has been widely employed for the fabrica-
tion of integrated array waveguide gratings (AWGs) \[39, 40\]. However, this platform does not allow the small footprint achievable by the other materials. InP is perhaps the most versatile material platform since it allows the integration of both passive and active functionalities on the same chip. Waveguides, light sources, modulators and amplifiers can be co-integrated on the same device allowing the fabrication of high-complexity PICs \[41, 42\]. SOI chips are promising for their limited footprint size and the CMOS compatibility of their fabrication process \[43\], while active functionalities on this material are still an open challenge.

To exploit the advantages and overcome the limitations of the different material platforms, hybrid solutions have been considered \[44–48\].

The requirements and the challenges described in this section represent the motivations of this thesis. This work aims to investigate, study and implement an OPS architecture for intra-DC networks. The scenario depicted in Figure 1.2 has to be considered as the starting point of this research. The desired OPS has not only to be capable of handling a large number \((\geq 1000)\) of high data-rate input/output ports but it has to add extremely low latency to the system \((\leq 1\mu s)\) and its integration should be feasible.

### 1.4 Optical packet switches: state of the art

In the last decade, many research groups have focused on the implementation of multi-port optical switches. In this work, the interest is not only on the switch port-count, but it is also on the switching speed. Switches based on optical MEMS \[49\] or thermo-optic \[50\], although capable of switching high data-rates and of scaling to relatively high port-count, are not taken into account in this section due to their limited switching speed.

The most important attempt to realize a large port-count OPS is perhaps the OSMOSIS project \[51\]. A broadcast and select switching architecture for high performance computing (HPC) applications was implemented by using optical switches based on semiconductor optical amplifiers (SOAs). While the optical switches used in the system could be switched between the on and off state in few nanoseconds, the overall response time was much slower due to the controller complexity. Moreover, the costs of the switch implementation were too high due to the high number of components
required.

Several integrated multi-port switches have been presented. In [52] a monolithically integrated 16 × 16 InP switch is presented. The switch relies on a Beneš architecture based on SOAs and can provide a maximum throughput of 640 Gb/s. The MOTOR device [53] is an 8 × 8 InP monolithically integrated optical switch based on tunable wavelength converters. In [54, 55] an InP integrated 1 × 16 optical switch has been presented. It can simultaneously handle both on off keying (OOK) and differential phase shift keying (DPSK) modulated data packets. In [56], this concept is scaled to a 1 × 100 optical switch. The IRIS project [57] demonstrated a multi-stage architecture based on a combination of space switches and wavelength routing switches that utilize fast wavelength converters and 40 × 40 array waveguide gratings (AWGs). Another interesting approach to switching is the one adopted in the vortex switch [58]. Multi-wavelength packets are delivered to the destination through concentric loops of optical nodes. In the SPINet network [59] packets are routed through an Omega architecture based on 2 × 2 switching nodes. In [60] a qualitative categorization and comparison of optical interconnects for DC applications is reported.

Although the referred works show a great progress in technology development in this field, these devices still feature a relatively low number of ports. Moreover, many of them have not yet been demonstrated in a system environment.

1.5 Preview of the thesis

The goal of this work is to identify, study and implement an optical packet switch architecture for high performance data center networks. Large port-count, high data-rate per port and low reconfiguration time are the main requirements taken into account to reach this goal. Such a switch will allow flattening the data center network enhancing the resources sharing capabilities within these systems, providing higher bandwidth and reducing the system latency and costs.

In Chapter 2, a novel architecture is presented. The choice on the architecture is motivated by the study of the effects of switch architectures on the intra-DC network performances. A novel single-chain analytic model is used as a tool to compute system latency and throughput as function of
the employed switch architecture and as function of the switch port-count. An ideal and a practical implementation of the proposed OPS architecture are presented in this chapter.

The novel architecture is based on a strictly non-blocking architecture. The switch operates transparently in the optical domain while being electronically controlled. The lack of optical random access memory is overcome by confining the buffering process at the edges of the switching matrix, where it can be performed electronically. The proposed architecture features a modular structure. The modules operate in parallel and autonomously by each other. Wavelength division multiplexing is employed to scale the architecture to a large port-count, thus the number of effective interconnections is given by the number of input fiber $F$ multiplied by the number of wavelength channels carried by each fiber $M$. Wavelength conversion is employed to avoid contentions between packets coming from different input fibers and directed to the same output port, enabling the autonomous and parallel operations of the modules.

Each module requires several functional blocks:

- Channel demultiplexer: to demultiplex the $M$ channels carried by each input fiber.
- Label extractor: to extract the packet label of each WDM channel.
- Label processor: to identify the packet destination.
- $1 \times F$ optical switch: to forward the packet payload of each WDM channel to the appropriate output port.
- Switch controller: to drive the optical switch according to the packet destination, to solve contentions and to generate acknowledge message for the transmitting nodes.
- $M \times 1$ wavelength selector: to select one out of a maximum of $M$ packets in case multiple packets coming from the same input fiber (and thus handled by the same module) have the same destination. Controlled by the switch controller.
- Fixed wavelength converter: to convert the packets to a fixed wavelength in order to avoid contentions between different modules.

Flow control enables packet retransmission of packets dropped by the wavelength selector. The modular structure of the architecture allows employing highly distributed control. Each module can be autonomously con-
trolled. This, in combination with in-band parallel labelling and label processing techniques, results in nanosecond reconfiguration time, minimizing the buffering needed at the edges of the switch and reducing the added latency associated with the packet retransmissions. Also, the modular structure allows investigating the architecture performances studying a single module subsystem.

In Chapter 3, the performances of the novel architecture, in terms of latency, packet loss and throughput are investigated by means of simulations of the OPS in an intra-DC environment. Systems employing switches with different port-count are considered as well as synchronous and asynchronous operations. The effects of the buffer capacity on the network performances are also investigated in this chapter.

In Chapter 4, the first implementation of the proposed OPS architecture (without the flow control functionality) and its controller is presented. The dynamic switching performances are investigated employing two distinct experimental set-up. In the first experiment, a single $1 \times 8$ switching module is implemented by employing discrete optical components. In the second experiment, an integrated $1 \times 4$ optical cross-connect is used to investigate the possible integration of the OPS. In both cases, an in-band labeling technique, which allows “on-the-fly” parallel packet labels detection is employed. The in-band packets labels, after being extracted from the payload, are O-E converted and detected by a dedicated electronic circuitry and finally processed by the switch controller. The switch controller is realized by programming a field-programmable gate array (FPGA) board, that is employed to solve possible contentions and to drive the OPS module.

In Chapter 5, two distinct flow control implementations are presented. The OPS switch architecture with a FPGA-based flow control performances are investigated in an experimental set-up. This implementation requires a dedicated electronic network to enable the flow control functionality. To overcome the impairments due to the employment of a distinct network for the flow control process in a large scale system as a DC, the implementation of an all-optical flow control is studied with promising results. This implementation relies on the re-utilization of part of the extracted label power to generate an optical acknowledge message by appropriately driving a SOA. The acknowledgement is sent back to the transmitting node on the same optical link employed for the packet transmission. Thus, using this method, there is no need of a dedicated network for the flow control process.
Novel contributions

In Chapter 6, further steps towards the total integration of the OPS architecture are taken. The employment of a microring resonators array as in-band label extractor for the proposed architecture is investigated. The employed device is able to successfully extract multiple in-band label wavelengths with limited penalty on the packet payload. Moreover, the employment of a channel demultiplexer and a microring-based label extractor integrated on a single device is exploited. Also, the integration of the other functional blocks required by each architecture module is discussed in this chapter.

Finally, Chapter 7 presents the conclusions of this work. The main results are summarized. Considerations about port-count scalability, latency, data-rate, throughput, integration and power consumption of the proposed switch architecture are given in this chapter. Also, alternative approaches and recommendations regarding future research related to the work presented in this thesis are reported.

1.6 Novel contributions

1. A simple analytic model that allows computing the latency and the throughput of a system as function of the employed switch architecture is presented. The model is used as a tool to identify a suitable OPS architecture for intra-DC networks.

2. A novel WDM optical packet switch architecture for intra-DC networks is presented. The main advantage of the selected architecture is its modular structure that enables the employment of a highly distributed control system. The optical switching modules operate in parallel and independently, due to the employment of wavelength converters that avoid inter-module contentions. This allows to reduce the overall switch reconfiguration time to the time to configure a single module.

3. The proposed architecture performances in a DC environment are investigated for the first time by means of simulations. Promising performances are provided under all traffic conditions simulated. Systems with port-count ranging from 4 to 1024 ports are simulated to investigate the scalability of the proposed architecture. Interesting results on the effects of the buffer capacity on the performances are
Novel contributions

reported.

4. The feasibility of a $64 \times 64$ OPS based on the proposed architecture employing discrete optical components is demonstrated. Although not full-functional (the flow control mechanism is not implemented), the $1 \times 8$ switching module and its controller are tested successfully in an experimental set-up injecting in the module $8 \times 40$ Gb/s WDM channels.

5. The feasibility of a $16 \times 16$ OPS based on $1 \times 4$ optical integrated cross-connect modules is demonstrated by injecting in the photonic chip $4 \times 40$ Gb/s WDM channels. Dynamic switching capabilities and error free operation are shown.

6. Two possible implementations of the controller flow control functionality are presented. An FPGA-based and an all-optical solution for the flow control system are tested in experimental set-up with promising results.

7. Further steps towards the total integration of the OPS architecture are taken investigating the employment of an integrated label extractor and an integrated WDM channels demultiplexer and label extractor suitable for the studied OPS architecture.
Chapter 2

Novel OPS architecture

In this chapter\(^1\), a novel optical packet switch (OPS) architecture suitable for intra-DC network is presented. Motivations on the choice of the proposed OPS architecture are presented in combination with an analytic study of the effects of the switch architectures on the intra-DC network performances.

2.1 System under investigation

This chapter aims at identifying an OPS architecture that is capable of optically interconnecting over thousand ports while being controllable in nanosecond time scales. The scenario described in Section 1.1.2 represents the starting point of this research. The future intra-DC network is assumed to feature a centralized thousand-port OPS which is employed to flatten the DC cluster layer. In this scenario, intra-rack and intra-cluster communications are managed by the pre-existing reliable and cost-efficient electronic switches, while inter-cluster communications are handled by the OPS. The system under investigation is depicted in Figure 2.1.

The OPS interconnects a number of ingress clusters with a number of egress clusters by means of high bit-rate input/output ports. The switch

is assumed to operate transparently in the optical domain, thus no optical-to-electrical (O-E) and electrical-to-optical (E-O) conversions at the inputs and outputs of the switch are required. The buffering is confined at the edges of the switching matrix, where it is electronically performed.

2.1.1 Architectures control issue

Many architectures can be used to implement an OPS. Despite many research groups are focusing on this topic since many years, scaling an OPS architecture to a large port-count is an open challenge, as discussed in Chapter 1. Many technological impairments can limit the number of ports of such architectures. However, what is often overlooked is the impact of the adopted switch architecture on the switch control time. In this respect, this section focus on the effects of the employed switch architecture on the switch reconfiguration time.

Realizing an optical packet switch that takes too long to be reconfigured will limit the system performance, requiring large buffers at the edges of the switching network and affecting the system operations in terms of latency and throughput. The following considerations are valid for the studied switch architectures regardless the technologies used for their implementations.

To facilitate the discussion, the focus is on synchronous switching opera-
tions. In a synchronous system, the operation is time slotted and the system operates in discrete time. At every time step, the switch controller handles all requests for routing simultaneously. It is possible to choose between many different switch architectures, and this choice has a deep impact on the switch control and on the network performance. For example, consider the switching matrix in Figure 2.1 implemented as a Beneš architecture. A Beneš switch is a rearrangeable non-blocking architecture that allows any input to be connected to any output. However, at any time slot in which the input state of the switch changes, the entire switch matrix needs to be reconfigured to establish a new connection map. The computation time on a single processing core of the best-known algorithm for reconfiguring a switch based on this architecture scales at least with $N \log_2 N$, where $N$ equals the number of inputs/outputs of the switch (this is the well-known “Looping Algorithm”, [62]). $N \log_2 N$ implies that, in principle, scaling up the number of ports by a factor 1000 will increase the computational time by a factor of 10,000. While small improvements and multi-core implementations have been considered [63–65], all the implementations scale at least linearly with $N$. Figure 2.2 shows the configuration times of a switch based on Beneš architecture (different rearrangeable non-blocking architectures such as Banyan or Omega will lead to equivalent results) for several control algorithms [66].

![Figure 2.2: Reconfiguration time of a Beneš switch matrix expressed in clock cycles employing a looping algorithm, a trial partition machine (TPM), and an algorithm that scales linearly with the switch port-count.](image)
The horizontal axis shows the number of inputs and outputs of the switching matrix, while the vertical axis shows the number of clock cycles needed to set the connection map. Assuming that a single clock cycle takes 1 $\text{ns}$, Figure 2.2 indicates that a 1024-port Beneš switch architecture with a scheduler employing the looping algorithm requires more than 10 $\mu\text{s}$ to be configured. During the configuration time, the switch is unable to handle data; thus incoming packets are either lost or need to be stored in electronic buffers, increasing the link latency and limiting the throughput of the system. Assuming that some services in a DC environment allow a maximum latency of 1 $\mu\text{s}$, it can be concluded that, if the number of switch ports exceeds a critical threshold, the switch is no longer capable of handling the incoming switching requests within the maximum latency allowed in the system.

Consider now the switching matrix based on the architecture depicted in Figure 2.3. It is a Spanke architecture with $N$ input/output ports. This switch is based on a strictly non-blocking architecture in which the main components are the $1 \times N$ and $N \times 1$ space switches \cite{67}. This architecture has a modular structure: the number of $1 \times N$ and $N \times 1$ switches scales linearly with the number of ports, and each of these switches can be autonomously controlled. Thus, the control complexity and the configuration time of the entire multi-port switch can be drastically reduced. In this case, the overall time to configure the switch matrix is the time to configure a single $1 \times N$ switch and a single $N \times 1$ switch. In fact, packet headers at each of the $N$ inputs can be locally processed by independent controllers.


2.2 Single-chain model

The author wishes to model the system depicted in Figure 2.1 into a simple single-chain model in order to compute the average throughput and latency of the system as function of the switch architecture employed. Beneš and Spanke architectures are employed to highlight the effects of a central or distributed control on the switch reconfiguration time. Figure 2.4 shows the simple model adopted to investigate the switch architectures performances.

![Figure 2.4: Schematic of the simple model used to study latency and throughput of the system under investigation](image)

Consider a short host-switch distance of 56 m. The round trip time (RTT) results in being 560 ns, considering that a packet takes 5 ns to cross 1 m of optical fiber. The packet duration is set to 40 ns. A binomial arrival pattern with fixed probability (hereafter called load) is assumed at each input port. Obviously there is a probability of contentions (i.e. $P_{cont}$, the probability of having simultaneously two or more packets destined to the same output port). Uniform destinations are also assumed. Thus, the probability of each input to have a packet destined to a specific output port is $\text{load}/N$. As a consequence of the Bernoulli arrival pattern considered, the contention probability at the output of a switch with $N$ inputs and $N$ outputs is given by:

$$P_{cont} = \sum_{i=2}^{N} \binom{N}{i} \left( \frac{\text{load}}{N} \right)^i \left( 1 - \frac{\text{load}}{N} \right)^{N-i}$$  \hspace{1cm} (2.1)

where $i$ represents the number of packets destined to a single output in a single time slot.

To compute the latency and the throughput of the system, two timescales are considered, hereafter denoted as service times.
- The round trip time (RTT): this is the minimum time that a single packet spends in the system. The packet takes $RTT/2$ to reach the output, but it can be considered successfully delivered only after an acknowledgment message is received back by the input (taking again $RTT/2$).

- The configuration time of the switch ($t_{\text{switch}}$): this time delay depends on the architecture employed to realize the switch.

In addition, it has to be considered that the packets in the chain could contend the output port with the other $N - 1$ inputs. Consider a packet coming from input 1 and destined to output $N$ of the switch of Figure 2.1. There is a contention whenever a packet is present at input 1 and there is at least one packet coming from the other $N - 1$ inputs addressed to the same output port. When the packet experiences a contention and it is dropped (for output contentions in the Beneš architecture and in case of contentions in the $N \times 1$ switches for the Spanke architecture), it has to be retransmitted.

The packet takes $RTT/2$ to reach the output, and when it is successfully received an acknowledge message is generated and sent back to appropriate sender, again taking $RTT/2$. If the sender does not receive the acknowledgment after $RTT$, the packet is automatically retransmitted. Retransmitting implies serving the packet again; the packet is thus delayed by the service times associated with $RTT$ and $t_{\text{switch}}$. The retransmission probability depends on the specific arrival pattern considered. The probability of contention for a given packet ($P_{\text{cont, gp}}$) at one of the $N$ inputs of the system, considering a Bernoulli arrival pattern and uniform destinations, can be expressed as follows:

$$P_{\text{cont, gp}} = \sum_{k=1}^{N-1} \binom{N-1}{k} \left( \frac{\text{load}}{N} \right)^k \left( 1 - \frac{\text{load}}{N} \right)^{N-1-k}$$

(2.2)

where $k$ represents the number of packets coming from the others $N - 1$ inputs and destined to the same output port. Note that when $k = 0$ there is no contention. When there are $k$ packets arriving from the other $N - 1$ inputs, our given packet loses the contention with a probability $k/(k + 1)$. Equal priority is assumed for the packets here, so each packet has the same probability of being the contention winner. Thus, the probability of retransmission ($P_{\text{Retr}}$), i.e. the probability that a packet is present at one
input, experiences a contention, and loses it is given by:

\[ P_{\text{Retr}} = \text{load} \cdot \sum_{k=1}^{N-1} \binom{N-1}{k} \left( \frac{\text{load}}{N} \right)^k \left( 1 - \frac{\text{load}}{N} \right)^{N-1-k} \frac{k}{k+1} \] (2.3)

where \( \text{load} \) represents the probability of having a packet at the considered input.

The probability of retransmission expressed in Equation 2.3 affects \( \text{load} \) at each of the input ports. After each RTT, the probability of having a packet at the input of the system is increased by \( P_{\text{Retr}} \), as expressed in Equation 2.4. After a RTT, the input \( \text{load} \) is increased by the packets that lost the contention at the output. This effect is taken into account by the definition of real load (\( \text{load}_R \)). It is important to know that this is an approximation. In fact, once again uniform destinations for the packets that lose contentions are considered. In other words, \( \text{load} \) represents the traffic injected in the system, while \( \text{load}_R \) represents the real traffic in the system, considering also the retransmitted packets.

\[ \text{load}_R(i) = \text{load} + P_{\text{Retr}}(i-1) \] (2.4)

This increment on the probability of having a packet at the input port affects in turn the probability of retransmission, which becomes:

\[ P_{\text{Retr}}(i) = \text{load}_R(i) \cdot \sum_{k=1}^{N-1} \binom{N-1}{k} \left( \frac{\text{load}_R(i)}{N} \right)^k \left( 1 - \frac{\text{load}_R(i)}{N} \right)^{N-1-k} \frac{k}{k+1} \] (2.5)

\( \text{load}_R \) can be recursively computed by using Equation 2.5 in Equation 2.4 with the conditions \((i)\) \( \text{load}_R(0) = \text{load} \) and \((ii)\) \( 0 \leq \text{load}_R(i) \leq 1 \). Figure 2.5 shows the behavior of \( \text{load}_R \) as function of \( \text{load} \), considering different number of iterations \((I)\) in Equation 2.4 for a system with four input/output ports.

It is visible that, for a number of iterations larger than three, the changes are negligible. Figure 2.5 shows also that \( \text{load}_R \) is equal to 1, due to condition \((ii)\), for \( \text{load} \) values around 0.8. Under these traffic conditions, the buffers of a real system will be rapidly filled up and buffer overflowing will become unavoidable. However, it is important to note that the model does not take into account the buffer size, which is supposed to be infinite.
2.2.1 Throughput

Computing the average throughput (defined as the utilized bandwidth over the total bandwidth) of the system is relatively straightforward, once the real load ($load_R$) has been computed. The throughput of the system is equal to 1 every time slot in which there is at least one packet destined to the output port of the model of Figure 2.4. However, the throughput of the system is limited by the configuration time of the switch ($t_{switch}$); no packets are sent to the output during this delay. The configuration time of the switch, as already mentioned, is dependent on the architecture used to implement the switch.

$$\text{throughput} = \frac{RTT}{RTT + t_{switch}} \sum_{k=1}^{N} \binom{N}{k} \cdot \left( \frac{load_R}{N} \right)^k \left( 1 - \frac{load_R}{N} \right)^{N-k}$$

(2.6)

The summation in Equation 2.6 represents the probability that there is at least one packet destined to the output of the system, while the first term takes into account the delay introduced for configuring the switch matrix.
Figure 2.6 shows the average throughput computed using Equation 2.6 for the two architectures considered in this study with $load_R = 1$. In the case of the Beneš architecture both the looping algorithm and an algorithm in which the configuration time scales linearly with $N$ have been considered, setting a clock cycle equal to 1 ns. A constant configuration time of 6 ns is considered to configure the switch based on the Spanke architecture. This number is assumed to be independent of the number of ports of the switch. This assumption is justified by experimental results in which it is shown that integrated $1 \times N$ switches can be fabricated with control time independent of the number of ports [68, 69]. The configuration time for the Beneš switch is dependent on the number of ports, as shown in Figure 2.2.

![Figure 2.6: Throughput of the system under investigation computed according to Equation 2.6 for a switch based on Beneš (considering the looping algorithm and a linear algorithm) and Spanke architectures. The results are computed considering $load_R = 1$.](image)

It is visible from Figure 2.6 that the throughput of the switch based on the Beneš architecture is strongly port-count dependent and decreases for larger switching matrices. In the case of the Spanke architecture, the throughput remains almost constant, even for a switch matrix with more than a thousand ports, due to the port-count independent switch control time. The small decrements visible for larger switching matrices are associated with the fact that the contention probability for a given packet slightly increases as function of the number of input/output ports. This
can be explained by plotting the contention probability for a given packet as function of the port-count $N$, as described by Equation 2.2.

### 2.2.2 Latency

Computing the latency is more complicated. This is due to the fact that the analytic model described in this section does not consider the size of the input buffer of the system, while the latency of the system strongly depends on the buffer capacity. However, the aim of the model presented in this section is to provide a simple and fast tool to compute the latency and throughput of the studied system in order to obtain a general understanding of the node performance as function of the different architectures employed to realize the OPS.

The probability of retransmission is maximized ($P_{RetrMax}$), considering $\text{load} = 1$ in the summation of Equation 2.3 to compute the upper bound of the average system latency ($\text{latency}_{UB}$). Hence, the mean $\text{latency}_{UB}$ can be computed considering that:

\[
\text{latency}_{UB} = RTT + t_{\text{switch}} + P_{RetrMax} \text{latency}_{UB} \tag{2.7}
\]

or

\[
\text{latency}_{UB} = \frac{RTT + t_{\text{switch}}}{1 - P_{RetrMax}} \tag{2.8}
\]

Figure 2.7 shows the upper bound of the average system latency as function of the port-count, considering a switch based on Beneš and Spanke architecture according to Equation 2.8 and for $\text{load} = 1$.

Figure 2.7 clearly shows that for a large port-count switching matrix the latency of the system increases for the Beneš architecture. This degradation of the performance reflects the fact that, if the switch port number $N$ increases, the time to reconfigure the switch scales as $N \log_2 N$ for the looping algorithm case. In the figure, it is also visible that the latency is below 1 μs regardless of the switch port-count and the input load when the Spanke architecture is employed. For these reasons Spanke-type switch architectures with highly distributed control are investigated in the next sections of this chapter. In high-performance computing applications the link latency is a critical requirement. End-to-end latencies up to 1 μs are tolerated [70]. Some applications in DCs require a similar performance.
Figure 2.7: Latency as a function of the port-count for a switching matrix based on Beneš architecture employing the looping algorithm, and with configuration time that scales linearly with the port-count, and for a switching matrix based on the Spanke architecture as a function of the number of ports. The results are computed considering \( load = 1 \).

Thus, 1 \( \mu s \) will be considered as the maximum allowed latency in the system under investigation. Figure 2.7 shows also that, at full \( load \), a Beneš architecture cannot meet these requirements for port-count larger than 16, regardless of the switching speed. It is the switch control that limits the performance of the system. Even if one is able to improve the control time of the switch by making it scale linearly with the port-count, it is still not possible to scale these switches to more than 64 ports while meeting the latency requirements. For this reason, it is important to investigate switch architectures that have a configuration time port-count independent.

### 2.3 Novel architecture

The goal of this section is to individuate a switch architecture capable of optically interconnecting thousands of input and output ports and able to provide very small end-to-end link latency. The strategy is to find a scalable switch architecture with a control time that is port-count independent. The form of distributed control that will be described allows for fast (few \( ns \))
reconfiguration of the switch matrix and on timescales that are independent of the number of ports. This approach will allow realizing switching matrices that can meet the latency requirements for DC applications.

The starting point of the investigation is the results for the Spanke architecture shown in Section 2.4. The schematic implementation of the proposed Spanke-like architecture is depicted in Figure 2.8.

![Figure 2.8: WDM Spanke architecture with highly distributed control and contention resolution block (CRBs) at its outputs.](image)

It is based on a strictly non-blocking architecture which employs WDM. WDM has been widely employed and investigated in the contents of multi-bit-rate data communication networks [71, 72]. However, it has never been employed in a data center environment. The OPS of Figure 2.8 has $F$ input (output) fibers and $F$ WDM demultiplexers (multiplexers) at the input (output) to separate (combine) the wavelength channels, indicated by $\lambda_1$ to $\lambda_M$. Each of the wavelength channels is fed into a $1 \times F$ optical switch. The logical $F \cdot M \times F \cdot M$ switch ($F$ fibers, each of them carrying $M$ wavelengths) operates in parallel for each wavelength channel. The switch matrix does not guarantee any-to-any connection between the logical $F \cdot M$ inputs/outputs. However, the packets need only to be directed to the right output fiber and thus to the appropriate cluster of the DC. At that point, traffic is forwarded to the right server by the top-of-the-rack switches. The OPS is strictly non-blocking; any input can be connected to any output at any time. However, the CRB that is placed at the output introduces
Novel architecture

connectivity, and thus a contention probability, between different channels. The CRB implementation is discussed later in this section.

It is important to mention that monolithically integrated $1 \times 16$ and $1 \times 100$ optical space switches that potentially support such schematic architecture have already been presented and tested [54, 56]. The $1 \times 16$ ($1 \times 100$) switch has been shown to operate almost penalty free for different data formats at bit-rate up to 160 ($10$) Gb/s per channel, whereas the time to configure the switch ($\sim 5$ ns) is independent of the number of output ports. The results reported indicate that, for example, a $256 \times 256$ ($F = 16, M = 16$) OPS, implemented by using $1 \times 16$ switches as the one in [54] and considering 16 WDM channels at 160 Gb/s, could have an aggregate throughput which is larger than $40$ Tb/s ($16 \times 16 \times 160$ Gb/s).

Introducing the employment of wavelength division multiplexing in the OPS architecture is a necessity. Consider for example to realize a 1024-port optical switch based on a single-wavelength architecture. This will require the employment of 1024 $1 \times 1024$ optical switches, which is not feasible with current technologies. Alternatively, a 1024-port optical switch based on 32 wavelength channels architecture will similarly require 1024 $1 \times F$ switches, but now with $F$ equal to 32. It has been already mentioned that $1 \times F$ optical switches where $F$ is larger than 32 have been presented [56].

2.3.1 Ideal CRB implementation

The switch architecture investigated in this work features a contention resolution block (CRB) at each of its output ports. The purpose of this block is to perform packet contention resolution by using the wavelength domain and employing wavelength conversion. A schematic of a CRB subsystem is shown in Figure 2.9. This example sub-system has $V$ input channels at $Z$ different wavelengths, schematically indicated by $in_1$ to $in_V$. The sub-system has $Z$ outputs (indicated by $out_1$ to $out_Z$).

An ideal CRB implementation implies that, where there are $Z$ or fewer packets at the input of each CRB, it is possible to forward all of them to the CRB output. However, it is difficult to implement such a contention resolution block, since a large number of tunable converters and interconnections are required. This can be easily understood, since the scheme of Figure 2.8 requires that each of the $F \cdot M$ inputs of the CRB is provided with a tunable wavelength converter. Thus for the case that $F = M = 32$, each
output needs over 1000 tunable wavelength converters, requiring a large footprint and large costs. Moreover, all the 1024 inputs of a CRB have to be controlled at the same time and this will lead to the same control issue discussed in Section 2.1.1.

2.3.2 Practical CRB implementation

Figure 2.10 presents the Spanke-type architecture, in which the CRB is implemented by using wavelength selectors (WSs) and fixed wavelength converters (FWCs). Comparing with Figure 2.8, note that each CRB is implemented in a “decoupled” fashion. This architecture is more practical to implement: the number of WSs and FWCs scales linearly with the number of outputs. In other words, if the switch has 1024 logical output ports \( F = M = 32 \), 32 WSs and 32 FWCs are needed at each CRB. Further, the WSs at the input of each CRB limit the number of fiber connections and therefore the costs. A WS consists of an \( M \times 1 \) AWG with SOA gates at each of its inputs; devices with this functionality have already been demonstrated [73]. In addition, the CRBs allow for simple local control; each WS can be autonomously controlled, enabling highly distributed control of the overall architecture. The architecture of Figure 2.10 can be regarded as a classical Spanke architecture replacing the WSs with \( F \times 1 \) switches.

To investigate the performance of the architecture with distributed control in a system with electronic input buffers, the model of Section 2.2 is now applied. Figure 2.11 shows the mean latency upper bound and the av-
Figure 2.10: Spanke-type architecture for a WDM node with CRBs consisting of AWGs, fast wavelength selectors (WSs), and fixed wavelength converters (FWCs).

Average normalised throughput obtained by applying Equations 2.6 and 2.8 and considering the practical implementation of the CRB.

![Figure 2.11](image)

(a) Latency upper bound and (b) average throughput obtained by using the single-chain model of Section 2.2 on a system with input buffer and thus retransmission, and where the CRB is implemented according to the practical implementation.

According to the results in Figure 2.11 (a), the latency of the system is always below the threshold of 1 $\mu s$ regardless of the port-count and the input load value. The latency is still slightly port-count dependent, even if this effect becomes negligible for switching matrices with more than 64...
ports. This is due to the saturation of the retransmission probability for high port-count values.

In Figure 2.11 (b), it is visible that the throughput of the system is completely port-count independent for load up to 0.7, and increases linearly with load. Once again, there is no sensible port-count dependency when the system has more than 64 input/output ports. Both the latency and the throughput saturate for load values higher than 0.7, due to condition (ii) relative to load$_R$ definition of Section 2.2. Moreover, the employed simple analytic model does not take into account the length of the buffer queues: thus the latency does not increase rapidly for high loads. The throughput is instead limited by the number of contentions that in turn are dependent on the traffic destination pattern. Larger switching matrices saturate to higher latency and to lower throughput values, due to the higher contention probability.

2.3.3 Packet loss performance in a buffer-less system

The simple analytic model adopted does not allow computing the packet loss of the system, since it does not take into account the capacity of the input buffers, which is assumed to be infinite. It is a relatively simple exercise to calculate the packet loss probabilities at the CRB in a buffer-less system in which retransmission is not provided. To find the packet loss, the system is modelled in a similar fashion as in the single-chain model of Section 2.2, but without the input buffers. This implies that retransmission is not possible, and the chain is as shown in Figure 2.12.

![Figure 2.12: Single-chain analytic model of the system in Figure 2.1 in which there are no input buffers. In this model, the probability of retransmission is replaced by the packet loss probability.](image)

For the packet loss calculation in the case of the ideal CRB, denote the number of inputs of the CRB (see Figure 2.9) by $V$ and the number of its outputs by $Z$. Once again, a Bernoulli arrival pattern is assumed. At
each of the $V$ input ports, there is a probability $p$ of having an arriving packet, independent of the previous time slots and neighboring ports. This is a rather natural assumption. Analysis in this case is simple: given that there are $k \in \{0, \ldots, V\}$ input packets in each time slot, the number of lost packets is $\max(k - Z, 0)$. The probability of having $k$ inputs follows a binomial distribution: choosing $k$ elements out of $V$ independent trials, each with probability $p$; thus the loss probability ($P_{\text{loss}}$) can be obtained with the following formula:

$$P_{\text{loss}} = \frac{\sum_{k=Z+1}^{V} \binom{V}{k} p^k (1-p)^{V-k} (k - Z)}{Vp}$$

Equation 2.9 is now applied to the Spanke-type architecture of Figure 2.8. Denote the number of input fibers of the switch by $F$ and the number of wavelength channels of each fiber by $M$. Hence, the number of inputs of each CRB is $V = F \cdot M$, and the number of outputs is $Z = M$. Assume furthermore that the probability of having a packet at each input port of the CRB is then $\text{load}/F$, under the assumption of uniform destinations. Solving Equation 2.9 under these assumptions, it is found that the packet loss depends on $\text{load}$, as shown in Figure 2.13. The packet loss is presented for switches with $F = 8, 16, 32$ and assuming $M = F$, implying that the switches have 64, 256, and 1024 logical input/output ports, respectively. It follows from Figure 2.13 that this architecture allows for lower packet loss, increasing the switch port-count.

However, as already mentioned, the ideal CRB is not easily implementable. Thus the packet loss in a buffer-less system of a switch based on the architecture with the practical CRB implementation is studied. In this case, the packet loss can be computed as follows. Suppose that each CRB has $F \cdot M$ inputs and $M$ outputs, and a probability of $\text{load}/F$ for a packet to arrive at each channel of the WSs. In this configuration, packets are lost if two or more packets arrive simultaneously at the input of each WSs, since the selector can only handle one packet at a time. Thus the packet loss is given by:
Figure 2.13: Packet loss probabilities versus load for a Spanke-type architecture with \( F = 8, 16, 32 \), \( F \) being the number of optical fibers and the number of wavelength channels at each input fiber, considering the ideal CRB.

\[
P_{loss} = \frac{\sum_{k=2}^{M} \binom{M}{k} \left( \frac{\text{load}}{F} \right)^k (1 - \frac{\text{load}}{F})^{M-k} (k - 1)}{M^{\frac{\text{load}}{F}}} \tag{2.10}
\]

The numerator represents the probability of having two or more packets at the inputs of a single WS and the number of packets that are lost if \( k \) packets are simultaneously present \((k - 1)\). The denominator represents the expected value of arrivals at the inputs of a single WS. To evaluate the model, simulations of the buffer-less system are run employing the OMNet++ simulation package [74]. Using Equation 2.10, the packet loss of the system according to the model of Figure 2.12 is computed, considering different input loads and switch sizes. Figure 2.14 shows the packet loss in a buffer-less system with 64, 256, and 1024 input/output ports obtained by using both model and OMNeT++ simulations. The packet loss results obtained by applying Equation 2.10 and by using the simulation package are in perfect agreement.

However, it is immediately visible that the packet loss has increased significantly compared to the packet loss of the ideal implementation of the
CRB shown in Figure 2.13. This degradation of the performance is the price paid for the simplicity of the CRB implementation of Figure 2.10. Note also that there is a small increment in the packet loss if the number of ports increases.

The packet loss expressed in Equation 2.10 may be used as a rough measure of the packet loss for the purpose of system and network design. In this respect, the architecture of Figure 2.10 can be considered as “worst case” architecture in terms of packet loss. This is sensible, since it is very thrifty in component usage and control requirements. Adding more components in a smart manner should only improve on Equation 2.10.

Figure 2.13 and 2.14 show the best-case scenario and worst-case scenario for a packet loss in a WDM Spanke-type architecture with a CRB based on wavelength conversion in a buffer-less system.

However, to understand the performance of such architecture in a realistic environment, the system has to be simulated considering buffers with a finite capacity. This is done in the next chapter of this thesis.
2.4 Summary

The challenges that have to be faced in the realization of an OPS architecture that could provide high port-count (>1000), high communication bandwidth and extremely low latency (<1 \(\mu s\)) are presented.

A simple single-chain model that allows computing latency and throughput of a system as function of the particular architecture employed is introduced. The model is employed as a tool to show to the reader that OPS architectures that employ a centralized control are not suited to realize a large port-count OPS, due to the high latency introduced by the switch reconfiguration time.

As a consequence of the results obtained using the model, a novel WDM modular OPS architecture is presented. The proposed OPS is based on a Spanke-like architecture which employs WDM and wavelength conversion for contention resolution purposes. The architecture is based on independent optical modules that allow highly distributed and local control. Ideal and practical implementations of the contention resolution functionality are reported in combination with their performance, in terms of packet loss, in a buffer-less system. On the one hand, the ideal implementation allows for extremely low packet loss, while the practical implementation performances do not look promising. On the other hand, the realization of the ideal implementation is cumbersome due to the high number of components required and to the control complexity of its contention resolution stage, while the practical implementation is easier to realize. In this case the number of components scales linearly with the port-count and the contention resolution functionality is realized in a “decoupled” manner, allowing distributed control also at this stage.

To evaluate the capabilities of the OPS architecture which employs the practical implementation of the contention resolution functionality, input buffers with finite capacity have to be considered.
Chapter 3

Performance investigation

In this chapter\(^1\), the performances of the proposed optical packet switch (OPS) architecture are investigated by means of simulations. Results of latency, throughput and packet loss are presented as function of system load in a simulated intra-DC environment. Synchronous and asynchronous operations are taken into account as well as different traffic patterns. Buffer sizing is studied in systems with fixed and variable packet lengths.

3.1 Simulation environment

The performance of the OPS architecture described in the previous chapter is studied by means of node simulations. OMNeT++ simulation package [74] is the software employed to simulate the studied DC environment.

OMNeT++ is a discrete event simulation environment commonly used

to simulate communication networks. This software tool provides C++ defined components and models. Components and models are then assembled into larger components and models using high-level language.

The schematic of the simulated system is shown in Figure 3.1. It is an inter-cluster communication network in which the total number of input/output ports is $F \cdot M$. $F$ represents the number of input optical fibers of the OPS, and thus a cluster, while $M$ represents the number of wavelength channels carried by each fiber ($\lambda_1$ to $\lambda_M$). Each cluster groups $M$ top-of-the-rack (TOR) switches. Each of the $M$ TOR switches has a dedicated queue ($Queue 1$ to $Queue M$) in the $F$ electronic input buffers as shown in the figure. Each cluster is connected to the OPS architecture by optical fibers. Thus packets present in each queue are electrical to optical converted by using commercial WDM optical transceivers. A host-switch distance of 50 meters is considered in the simulations. Longer or shorter distances would not affect the OPS performance, but the distance between switch and clusters can be regarded as a latency offset introduced in the OPS operation. The system is set to operate at a data rate of 40 $Gb/s$, in accordance with the future DC inter-cluster communication requirements [8].

The OPS processes in parallel the $F \cdot M$ inputs ports by using parallel $1 \times F$ switches, each of them autonomously controlled. The contention resolution block consists of $M \times 1$ wavelength selectors (WSs), also locally controlled, and fixed wavelength converters (FWCs).

It is important to notice that possible contentions between packets coming from different clusters are avoided by the FWCs at the outputs of the WSs as described in Section 2.3. The FWCs convert the packet selected by the WSs to a fixed wavelength and solve these contentions. For this reason, only $M$ inputs/outputs relative to a single input buffer are considered, without compromising the results of the simulations. The subsystem considered is highlighted in Figure 3.1. Contentions occur only between packets at the $M$ inputs of each $M \times 1$ WS.

The optical packets consist of a payload, which carries the real data and an optical label that provides the packet destination port. In the OPS labels are separated from the payload and processed by label processors. The label processors then control the $1 \times F$ optical switches to forward the packets to the appropriate contention resolution block (CRB). At each CRB contentions are avoided by the local control that selects only one packet out of the possible $M$ incoming packets. The selected packet is forwarded to the
Simulation environment

Figure 3.1: Subsystem studied by means of simulations. The system employs the architecture with the practical CRB implementation of Section 2.3.2. Each input buffer consists of \( M \) queues, each of them associated with a different wavelength channel.

destination cluster by means of an optical link and a positive flow control signal acknowledges the reception of the packet. The dropped packets need to be retransmitted.

As demonstrated in the previous chapter, on the contrary of an OPS with centralized control, the proposed architecture control complexity and the reconfiguration time are mainly determined by the label processing time and not by the switch architecture. In fact, the reconfiguration time of \( 1 \times F \) and switches and \( M \times 1 \) WSs is in the order of nanoseconds and on-the-fly operation of the FWC can be considered. Key issue to minimize the overall latency in the system is then the implementation of a labeling technique that allows for extremely low label processing delay. The employed labeling technique is described later in this thesis. For the simulations a round trip time (RTT) of 560 ns is considered, including the delay to process the labels, control the switches and the latency introduced by crossing \( 2 \times 50 = 100 \) m of optical link.

To investigate the performance of the inter-DC network employing the modular WDM OPS with flow control mechanism, all the functionalities of
the optical switch and the cluster buffers queues are implemented one-to-one in the OMNet++ simulation environment.

Figure 3.2 shows the system under investigation implemented in the simulation environment. During the simulations the traffic generators components create packets, according to a certain traffic pattern and load value. Generated packets are stored in the electronic buffer components, and a copy is sent into the $1 \times F$ optical switches via 50 m long optical links. At the OPS node the optical label is extracted from the packet and processed. When a packet reaches the appropriate CRB, if no contention occurs, it is forwarded to the output port and thus to the destination cluster. At each packet delivery, an acknowledgement signal is sent back to the input. Once the acknowledge is received the packet is erased from the input buffer queue. However, it may happen that two or more packets coming from the same cluster (and thus from the same input buffer) have the same destination port. In the case that multiple packets reach the same CRB simultaneously, a contention occurs. In case of contention, only one packet reaches the output cluster, whereas the others are dropped. At the input node only the delivered packet is acknowledged and erased from the buffer queue, while the other ones are kept in the queues and need retransmission. In case of heavy traffic load many contentions may occur. As a consequence, the queues at the input buffers may rapidly grow. The input buffers have finite capacity, if the input buffers are full when new packets are generated by the
traffic generators, they are considered lost and used to calculate the packet loss ratio.

### 3.2 Synchronous operation

At the first stage of the OPS architecture performance evaluation, the OM-NeT++ implementation of the subsystem highlighted in Figure 3.1 is set to operate in discrete time. During the simulation runs the operations are time slotted and the system operates as follows.

At each time slot, each traffic generator creates packets according to a binomial distribution with fixed priority (hereafter called \textit{load}). Thus, there is probability \textit{load} of having a new packet at each queue of the input buffer considered. All traffic generators are programmed to operate independently, but with the same \textit{load}. As a consequence, the \textit{load} set at the beginning of the simulation represents the mean \textit{load} value in the system, while the current \textit{load} may vary from 0 to 1 in each time slot studied during the simulation run. Packets destinations are chosen randomly between all possible outputs according to a uniform distribution. Packet duration is set to be of 40 ns (200 bytes, considering that the system operates at a data rate of 40 Gb/s). A copy of the first packets in each of the buffer queues is sent to the $1 \times F$ switches. At this stage, after RTT/2, the packets labels are read and the packets payloads are forwarded to the appropriate CRB. After RTT, a positive or negative acknowledgement message for each packet is received back at the transmitting inputs. In case of positive acknowledgement, the corresponding packet is erased from the buffer queue. In case of negative acknowledgment, another copy of the corresponding packet is transmitted to the OPS. In order to have uniform queues at the input buffers, in case of output contentions, the CRB are programmed to employ a round robin algorithm to select the contention winner. Thus packet priority is a priori determined, uniformly distributed between the packets of the different queues and no extra latency is added to the system. Input buffer queues have a capacity of 14 packets, according to the rule of thumb that says that the amount of required buffering $B$ is $B = RTT \times \text{PacketRate}$ ($B = 560 \ [ns] \times 1/40 \ [packets/ns] = 14 \ [packets]$). Depending on the traffic \textit{load} considered and the number of contentions, it may happen that a buffer queue is already full when a new packet is generated by the traffic
Synchronous operation

generators. In this scenario, the new packet is immediately dropped and lost, due to buffer overflowing.

The simulations are run under different values of mean load and for different switch sizes. To facilitate the discussion, the number of input fiber \( F \) is set equal to the number of wavelength channels carried by each fiber \( M \). Similar results, not reported, can be obtained for \( M > F \), considering that each WS could select more WDM channels at the same time in this case. With \( M < F \), the architecture implementation requires an extra stage of WSs at each CRB. In this case, the number of \( M \times 1 \) WSs in each CRB is equal to the number of input fibers \( F \). \( F \) is also the number of outputs directed to the FWCs, but only \( M < F \) wavelength channels are available at each output fiber of the OPS. An extra stage of WSs is needed to reduce the number of outputs directed to the FWCs down to \( M \). However, this case is negligible since hundreds of WDM channels can be transported by a single fiber, increasing the system spectral efficiency. The acknowledgement messages are not considered in the system traffic and the lost packets are not considered in the calculation of the average latency that the packets experience in the system. At the output of the system the transmitted packets are collected to calculate the average latency they experience in the system and to compute the system throughput. As already mentioned, lost packets are considered to compute the packet loss ratio.

Figures 3.3 (a), (b) and (c) show the average latency, the packet loss and the throughput of the system, respectively. The simulations are run for \( 10^{10} \) time slots. The performances of switches with 4, 16, 64, 256 and 1024 input/output ports are investigated, assuming \( F = M = 2, 4, 8, 16 \) and 32.

In Figure 3.3 (a), it is visible that the system latency increases with the number of ports. However, the increment becomes smaller and smaller and an asymptotic curve is rapidly reached. In the same figure is also visible that the latency increases linearly with the load until load values around 0.7. After this threshold the latency rapidly increases. This behavior can be explained as follows: on the one hand higher load leads to the presence of more packets in the system and thus higher number of contentions. Having more contentions implies higher probability for a packet to be re-transmitted and thus increased latency. On the other hand over the 0.7 threshold of the load, buffer overflowing is unavoidable. In this scenario, many packets are dropped and others have to be re-transmitted several
times, this process leads to the higher increasing rate of the system latency.

![Figure 3.3: Average latency (a), packet loss and its upper bound (b), and throughput (c) as function of load of the simulated system. Switch matrices with 4, 16, 64, 256 and 1024 input/output ports have been considered ($F = M$).](image)

Figure 3.3 (b) shows the packet loss of the system. In agreement with the results of latency, the packet loss of the system increases for load increments. The figure shows also the upper bound expressed in Equation 2.10 of Section 2.3.3, i.e. the packet loss in a buffer-less system for a switch with 1024 in/out ports. As expected, the packet loss of the simulated system is lower than the packet loss calculated in a system in which the input buffers are not considered. The benefits of using input buffers are higher under low load operation conditions, while the use of buffers becomes less important.
with load values close to 1.

Figure 3.3 (c) shows the throughput of the system. The throughput is computed as the rate between the used bandwidth over the total system bandwidth. In this way it is possible to compare subsystems with different port-count and thus with different bandwidths. The ideal curve represents the maximum throughput achievable under the considered load value. The throughput increases linearly with load until load values around 0.6. After this value the throughput increases much slower and for values over 0.8 it starts saturating. This behavior can be explained by the buffer overflowing that plays a dominant role for load larger than 0.8. Figure 3.3 (c) also shows that in its linear range the throughput is port-count independent, while small decrements in throughput are visible increasing the port-count for load higher than 0.6. Before this threshold, the input buffers allow no or very low losses regardless the port-count and the throughput follows its ideal curve, as expected considering uniform destinations for the packets.

Figure 3.4 shows the mean buffer population as function of load obtained during the simulations. It is visible that with load $\geq 0.8$ the buffer mean population is over 13 packets. As a consequences, the buffer is often full and overflowing, causing high packet loss, high latency and explaining the saturation of the throughput of the system. However, it is important to remark the consideration about the DC link utilization of Section 1.3.1: the typical DC load is below 0.3.

![Figure 3.4: Mean buffer queue as a function of load expressed in number of packets.](image-url)
3.2.1 Simulations vs Model

The results of these simulations differ from the results presented in Section 2.3.2 obtained using the single-chain model. A comparison between the curves in Figure 3.5 (a) shows that the model allows computing the throughput of the system accurately. On the other hand, there are discrepancies in the latency results, as shown in Figure 3.5 (b). The latency results are in agreement for load ≤ 0.4. The differences for load > 0.4 can be explained by the assumption of input buffer with infinite capacity considered in the mathematical model. The buffer capacity plays an important role in the calculation of the system latency. For load values higher than 0.4, the model no longer describes the system accurately. In fact, under these traffic conditions buffer overflowing plays a dominant role, which affects the latency of the system.

![Figure 3.5: Results comparison between OMNeT++ simulations and single-chain model in terms of throughput (a) and latency (b) for a 1024-port system.](image)

The results presented in this section reveal that the proposed architecture allows for small average latency in the system and relatively high throughput regardless of the port-count and the input load. However, under heavy input load (> 0.7), the buffers are continuously overflowing, and the system cannot be sustained any longer, due to the high losses.

3.2.2 Buffer sizing

As already mentioned, the buffer queue capacity has been set according to the rule of thumb that says that the amount of required buffering $B$
Synchronous operation

is \( B \{\text{packets}\} = \text{RTT} \{\text{s}\} \times \text{PacketRate} \{\text{packets/s}\} \). It has been experimentally demonstrated that in core routers this value can be significantly reduced [77] and that it may be useful to employ limited sized buffers at the expense of lower throughput [78]. However, it is interesting to investigate if an higher buffer capacity would decrease the packet losses under heavy traffic load conditions. For this purpose, the simulation set-up considering 1024-port system is run with \( B = 15, 17, 19, 24, 34 \) and 64 packets. The results in terms of packet loss and latency are then compared to the ones obtained considering a buffer queue capacity of 14 packets.

Figure 3.6 (a) and (b) show the packet loss and the latency of the system as function of load and for different buffer queues capacities. Figure 3.6 (a) shows that increasing the buffer size results in decreased packet loss, but only for load lower than 0.7. Over this threshold, an increased buffer capacity does not allow significant improvements on the packet loss performance of the system. Moreover, Figure 3.6 (b) shows that increasing the buffer capacity does not affect the system latency for \( \text{load} \geq 0.5 \). However, for higher load the system latency rapidly increases with the input buffer capacity.

![Figure 3.6: Packet loss (a) and latency (b) as function of load. The results refer to a system with 1024 input/output ports considering input buffer queue capacity (B) of 14, 15 17, 19 24, 34 and 64 packets.](image)

The node simulations reveal that the proposed architecture, considering synchronous operation, allows latencies below 1 \( \mu \text{s} \) regardless of the switch port-count and the load. However, traffic in a DC is not time slotted and packet size may vary from the minimum to the maximum ethernet frame. As a consequence, the next step of the performance investigation of
the OPS architecture studied is to simulate a system considering a traffic pattern that realistically represents DC traffic. This is done in the next section of this chapter.

3.3 Asynchronous operation

To evaluate the performance of the OPS architecture under investigation, a data center-like traffic pattern is considered. The simulation set-up remains the same of Figure 3.1. However, the functionalities of traffic generator and CRB OMNeT++ components are modified.

3.3.1 Self-similar traffic

During these simulations, the traffic generators are programmed to create packets with any arbitrary length, and send them to the $1 \times F$ switches according to specific inter-arrival time parameters. Each of these components emulates the aggregated traffic of a TOR switch. Each of the $M$ inputs of a single cluster receives the input traffic generated by 200 simulated servers. The amount of input load is normalized and is scaled from 0 to 1. The generation of DC traffic is based on referred publications regarding this topic [25, 26]. Moreover, data from more general studies about Internet traffic are also taken into account [79, 80]. According to these publications, packet length in a realistic DC scenario shows a bimodal distribution around 40 and 1500 bytes. These values match the minimum and the maximum Ethernet frame length, respectively. Figure 3.7 (a) and (b) show the cumulative distribution function (CDF) and the histogram of the packet length generated by the traffic generators during the simulations.

Packet inter-arrival time is modelled matching ON/OFF periods (with and without packet retransmission). The length of these ON/OFF periods is characterized by heavy-tailed self-similar distributions. In the simulation set-up studied, a Pareto distribution has been employed to model the length of the ON/OFF periods. Figure 3.8 (a) and (b) show the ON and OFF period length CDFs. ON periods follow the same length distribution regardless of the chosen traffic load value. However, the time between them (i.e. the OFF period) is inversely proportional to load. When a higher load is selected they become shorter and vice versa. As a consequence, the traf-
Asynchronous operation

Figure 3.7: Cumulative distribution function (CDF) (a) and histogram (b) of the generated packets length.

fic complexity remains constant regardless the chosen load, while it is the traffic density that varies in simulations with different load values.

Flattening the DC network architecture at cluster level allows to assume that the communication performances in the system do not depend on the distance or hierarchical position of the destination nodes. In addition, it is assumed that the traffic in the system is equally spread through all the resources. This is a natural assumption considering the aggregated traffic of hundreds of servers at cluster level instead of the traffic at server level. In other words, during the simulation, all packets in every transmission (ON periods) are randomly sent with uniform distribution to one of the possible destinations.
DC traffic characteristics are strongly application-dependent. This means that they may vary significantly from one environment to another. It is then difficult (if even possible) to define a concept such as a typical DC traffic pattern. The traffic model employed here is just a generalized approximation to the very different scenarios that can be found in a real DC.

Table 3.1 summarizes the typical DC traffic conditions, as can be found in [25, 26]. It can be said that, during 95% of the operations, traffic load in DC does not exceed 10% of the maximum network capacity. Moreover, for more than 99% of the time, traffic load is not beyond the 30% of the total capacity. In the upcoming sections, the values reported in Table 3.1 are addressed as typical DC traffic conditions.
Table 3.1: Summary of the typical DC traffic conditions and their probabilities.

<table>
<thead>
<tr>
<th>Probability (% of time)</th>
<th>Normalized load</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td>&lt; 0.1</td>
<td>Normal traffic</td>
</tr>
<tr>
<td>99%</td>
<td>&lt; 0.3</td>
<td>Heavy traffic</td>
</tr>
</tbody>
</table>

3.3.2 Latency, throughput and packet loss

The scalability in terms of port-count of the studied OPS architecture is investigated simulating an asynchronous system with $F = M = 2, 4, 8, 16$ and 32 that correspond to systems which employ an OPS with a total of $F \cdot M = 4, 16, 64, 256$ and 1024 input/output ports.

The system characteristics remain the same of the synchronous case. The host-switch distance is set to 50 m, the data-rate at each input is of 40 Gb/s and the RTT is set to be of 560 ns. The CRBs have been programmed, in this case, according to a first-come/first-choice service policy. The input buffer capacity is set to 20 kB. Average system latency, throughput and packet loss are calculated as function of the traffic load. Figures 3.9 (a), (b) and (c) show the results obtained by simulating the DC environment under asynchronous operation assumption.

Figure 3.9 (a) shows the average latency that the packets experience in the system. As expected, the average latency increases in systems with higher port-count. This behavior is due to the increasing contention probability at the CRBs of the OPS architecture. For example, considering load = 0.4, incrementing the number of ports from 4 to 16 corresponds to 250 ns of added delay. However, beyond 256 input/output ports the latency discrepancies are minimal. Similar behavior is also shown in the packet loss results depicted in Figure 3.9 (b). Higher port-count implies higher contention probability at the CRB stages of the OPS architecture. This translates in more packet retransmissions and higher buffer occupancy, compromising the packet loss performance of the system. For example, a 16-port system starts losing packets with load 12% lower than a 4-port one. The throughput of the simulated system is port-count independent for load ≤ 0.5 as shown in Figure 3.9 (c). However, for load ≥ 0.5, the throughput saturates to higher values in smaller switch matrices. Once again, the differences become minimal and negligible for port-count ≥ 256.
The results reported show that the WDM OPS architecture under investigation allows to obtain a packet loss lower than $10^{-6}$ and an average latency around 900 $\text{ns}$ for load up to 0.3. Considering the DC typical traffic condition reported in 3.1, these results confirm that the investigated architecture could handle DC typical traffic loads, regardless of the system port-count.

### 3.3.3 Buffer sizing

As already mentioned in Section 3.2, the electronic input buffers store the packets while they are travelling through the OPS. In this way, in case of output contentions, the dropped packet at the CRBs can be retransmitted.
Setting input buffers with large capacity should, in principle, help avoiding losses. However, large buffers capable of working at 40 Gb/s are still technologically difficult to obtain and heavy parallelization is required. Electrical buffer should be large enough to hold packets in case of contentions. As already seen from the results of the synchronous operation case, a larger buffer does not improve packet loss performance under heavy traffic load conditions, compromising, in addition, the average latency of the system. However, it is interesting to investigate the effects of different buffer sizes also assuming asynchronous operations. For this purpose, a system with 1024 ports ($F = M = 32$) is simulated considering different buffer sizes. Figures 3.10 (a) and (b) show the results of these simulations in terms of packet loss and average latency, respectively. Buffer capacities of 5, 10, 15, 20, 25 and 50 kB have been considered.

In Figure 3.10 (a) it is visible that, as the buffer size increases, the system can handle heavier traffic load without packet losses. If the buffer capacity is too small, like 5 or 10 kB, typical DC traffic cannot be sustained (see Table 3.1). The minimum studied buffer size that allows good performance is 15 kB. In this case, the system could handle an input load up to 0.3 providing a packet loss lower than $10^{-6}$. Increasing the buffer capacity beyond this value would provide better performance in terms of packet loss.

Since the architecture relies on the packets retransmission mechanism, the real amount of traffic inside the switch will always be higher than the
input traffic. For this reason, even a very large size buffer, would not achieve lossless operation considering $load \geq 0.7$. Beyond this $load$ value packet loss is buffer capacity independent.

The buffer capacity affects also the average system latency: if more packets can be held in the queues, they will end up waiting a longer time before being transmitted. Figure 3.10 (b) shows the effect of having different buffer capacity on the average system latency. According to those results, a buffer of 20 $kB$ should be enough to assure a packet loss below $10^{-6}$ for typical DC traffic and up to $load = 0.35$ providing sub-microsecond latencies ($\sim 900$ ns ). This is the reason why 20 $kB$ has been intentionally used in the previous section to investigate the performance of different sized systems. Higher buffer capacity could be considered to extend lossless operation to higher $load$. For example, with 50 $kB$ buffer per port, the system could handle $load \leq 0.45$. However, the trade-off between packet loss and latency implies that the packets would suffer on average a latency higher than 2 $\mu s$. In a real environment, an optimal buffer capacity should be chosen depending on the application requirements.

### 3.4 Discussion

The performances of the proposed optical packet switch architecture implementation in terms of latency, throughput and packet loss are investigated by means of OMNeT++ simulations. Two different traffic pattern and system operations are considered. In the time slotted case, fixed size packets are generated according to a binomial distribution. In the asynchronous case, packets with a variable length are generated according to a self-similar distribution with inter arrival time based on the Pareto distribution. In both cases, uniform distributed destinations are considered.

**Synchronous vs Asynchronous simulations**

As expected, the performance results in terms of latency and throughput obtained assuming asynchronous operation of the system are worse than the synchronous case. This is due to the fact that the probability of contention is higher in a system in which the packets can be transmitted at any arbitrary time and in which the packet length may considerably vary.
In the latency results, that strongly depends on the number of packet contents, this phenomenon is evident. In the synchronous simulations, the system latency is below 1 $\mu$s regardless the switch port-count and the load value considered. In the asynchronous case, the latency increases rapidly with the input load and is higher than 1 $\mu$s for load around 0.4 for all the switch port-count considered.

The throughput results show a similar behavior in both the operation methods considered. However, the performance slightly worsen in the asynchronous operation, again for the higher contention probability of the asynchronous system.

It is difficult to compare the packet loss results. These results are slightly dependent on the port-count of the considered system, but strongly dependent on the buffer capacity. The two considered cases (synchronous and asynchronous) differ for packet size, fixed and set to 200 bytes in the first case and variable between 40 bytes and 1500 bytes in the second case. As a consequence, different buffer capacities have been considered, compromising a possible comparison in terms of packet loss between the two systems. However, in both cases, no losses are shown for load < 0.3.

**Buffer sizing**

The effects of the buffer capacity on the system latency and packet loss are investigated for both the synchronous and asynchronous case. The results are very interesting. In both cases, increasing the buffer capacity does not correspond to a decrement of the packet loss under heavy traffic conditions (load > 0.7). An higher buffer capacity improves the system performance only for load values lower than this threshold. Moreover, the latency results reveal that the latency of the system is strongly dependent on the buffer capacity. The latency is higher in systems with larger buffers. This is due to the longer queueing times that the packets experience in buffers with a larger capacity.

In general, buffer sizing leads to a trade-off between packet loss and system latency. In this context, the system requirements in a DC environment are strongly application dependent. However, in all simulated operations the investigated OPS architecture allows providing promising results. No packet loss, high throughput and sub-microsecond latency can be achieved under relatively high traffic loads (< 0.4) and employing limited input
buffer capacities.

3.5 Summary

The performance of the modular WDM OPS with highly distributed control are investigated by means of node simulations. Two sets of simulations are considered. First, the DC inter-cluster network is assumed to operate in discrete time. The results in terms of average latency, throughput and packet loss, assuming synchronous operation, reveal that the investigated OPS architecture allows for sub-microsecond latencies, while providing low packet loss and high throughput, regardless the system port-count.

Second, asynchronous operation and variable packets lengths are considered during the simulations. The traffic pattern is modelled to describe the typical DC traffic conditions. Also in this case, results of average latency, throughput and packet loss as function of input load are reported. A degradation of the system performance is visible comparing these results with the ones obtained by considering synchronous operation. However, the OPS would be able to handle DC-like traffic providing an average latency around 900 $\text{ns}$, a packet loss lower than $10^{-6}$ and relatively high throughput.

In both synchronous and asynchronous cases, the effects of the buffer capacity on the system performance are investigated. The results are similar for both of the simulation sets considered. On the one hand, a limited buffer size (14 packets for the synchronous case and 20 kB for the asynchronous case) is able to guarantee lossless operation for typical DC traffic load. On the other hand, increasing buffer capacity does not help improving the packet loss performance of the system under heavy traffic load conditions ($\text{load} \geq 0.7$), compromising also the average packet latency.
Chapter 4

Dynamic switching

In this chapter\(^1\), all functionalities of the studied optical packet switch (OPS) architecture, with the only exception of the flow control mechanism, are implemented and experimentally evaluated. Labeling technique, label processor, \(1 \times F\) optical switch, switch controller and CRB implementations are described in this chapter. The dynamic switching capabilities, considering synchronous operation, are first investigated injecting 8 WDM channels at 40 \(Gb/s\) into a single \(1 \times 8\) optical module. In this case, discrete switching components are used to implement the \(1 \times 8\) optical module. Results of the dynamic operations and bit-error-rate measurements are reported. The experiment reveals that it is feasible to realize a \(64 \times 64\) OPS based on the proposed architecture. The possible integration of a \(16 \times 16\) OPS based on the studied architecture is then investigated injecting 4 WDM optical channels at 40 \(Gb/s\) into a \(1 \times 4\) integrated optical cross-connect.

Dynamic switching operations are shown in combination with bit-error-rate measurements of the 4 WDM channels after switching and after wavelength conversion.

4.1 Labeling technique

The OPS under investigation is based on a strictly non-blocking architecture. This means that there is a guaranteed optical path for every input to every output, thus no connection map has to be established. In addition, the modular structure of the proposed switch allows controlling independently and in a parallel manner all the $1 \times F$ optical switching modules. As a consequence, the switch reconfiguration time is mainly determined by the packet label processing time.

It is of key importance employing a labeling technique that allows parallel label processing and that adds extremely low latency (in the $\text{ns}$ order) to the system. Moreover, the labels should be able to address a large number of destination ports ($\geq 1000$).

![Spectrum of packet labelled with in-band optical labels (\(\lambda_1\) to \(\lambda_L\)) carrying RF tones (\(RF_1\) to \(RF_T\)) (a). Payloads have the same duration of the labels (b).](image)

In [84] a labeling technique that allows for parallel label processing, scaling to over thousand ports and provides $\text{ns}$ order label processing time has been presented. This is the labeling technique employed in this work. As shown in Figure 4.1 (a) the optical label wavelengths are placed within the payload spectrum bandwidth, indicated by $\lambda_1$ to $\lambda_L$. Each label wavelength carries $T$ radio frequency (RF) tones, indicate by $RF_1$ to $RF_T$. The RF tones are binary coded and each of them represents a label bit. Supposing
that $L$ label wavelengths are inserted within the packet payload, this labeling technique allows addressing $2^{L\cdot T}$ possible destinations. Moreover, the labels have the same duration of the packet payload, as shown in Figure 4.1 (b). $f_i^j$ represents the frequency of the RF tone $i \in [1,\ldots,T]$ travelling on the wavelength $j \in [1,\ldots,L]$. This labelling method allows handling packets with variable length in asynchronous operations. Labels and RF tones generation, scalability of the labeling technique and experimental demonstration for different packet formats are reported in the referred article.

At each input of the system under investigation, up to $M$ packets (and attached labels) may be travelling on $M$ different wavelength channels (indicated by $\lambda_1$ to $\lambda_M$). As shown in Figure 4.2 for a single input, to detect the packet destinations, the wavelength channels travelling in each input fiber of the system under investigation have to be demultiplexed. Once the optical channels are separated, the wavelength labels inserted in the packet payload have to be extracted and optical-to-electrical converted, while the packet payloads remain in the optical domain and are fed into the optical switch module. Subsequently, the converted label signals are sent to the switch controller, where they are processed in order to detect the carried RF tones, and thus to recover the label bits. Once the label bits are recovered, the switch controller drives properly the $1 \times F$ optical switch module and the packets are forwarded to the appropriate contention resolution block (CRB).

![Figure 4.2: Schematic of the label extraction and switch control for a single input of the system.](image)

As shown in Figure 4.3 for a single output, at each CRB, the packets are fed into the wavelength selectors (WSs). It is the WS controller that
decides which packet is forwarded to the fixed wavelength converter (FWC) in case that multiple packets are present at its inputs. The selected packet is then converted to the right wavelength and forwarded to its destination port.

![Diagram](image)

Figure 4.3: Schematic of the CRB and CRB controller for a single output of the system.

As already mentioned, it is extremely important to carry out these operations in $ns$ time scales. The modular structure of the studied architecture allows high parallelization of the operations. The label extractors, the switch controllers and the CRB controllers operate in parallel and independently. For this reason, the overall delay added by the OPS is reduced to the time to extract and process a single label plus the time needed to drive a single $1 \times F$ switch and to solve contentions in a single CRB. The implementation of the packet label processor, the CRB controller, the switch controller and the investigation of the system in dynamic switching experiments represent part of the work carried out in this thesis.

### 4.2 8-port optical module subsystem

The dynamic switching capabilities of the OPS architecture are investigated in an experimental set-up. The employed subsystem set-up is depicted in Figure 4.4. The label extractors and processors are implemented for all the WDM channels considered, while only the CRB of output port 1 of the subsystem is considered.
8 transmitters generate $8 \times 40 \text{ Gb/s}$ synchronous WDM packets. A single wavelength label is inserted in the payload spectrum bandwidth. Each wavelength label, one for each of the 8 WDM channels, carries three binary encoded RF tones. The three frequencies of the RF tones correspond to the three destination bits of the packets. The 8 channels are coupled and sent to the OPS module. At the input of the switch the 8 WDM channels are demultiplexed by an array waveguide grating (AWG). The label wavelengths are extracted from the payloads by fiber Bragg gratings (FBGs) in combination with optical circulators. The label extracting process is shown in Figure 4.5. The payload is fed into the $1 \times 8$ optical switch based on broadcast and select architecture and implemented by using discrete LiNbO$_3$ switches. At the same time, the extracted optical labels are optical-to-electrical converted and processed. To recover the three label bits carried by each wavelength label, three parallel band-pass filters (BPFs) centered at the RF tones frequencies and subsequent three parallel envelope detectors (EDs) are used (see Figure 4.5).

Each ED consists of a RF amplifier and a low-pass filter (LPF). The recovered label bits of all the 8 WDM packets are fed into a Virtex 4 FPGA (with 100 $MHz$ clock speed) directly employing the FPGA digital inputs. No power-hungry, high-speed analog-to-digital converters (ADCs) are required at this stage. The FPGA-based switch controller implements the label bits decoder and the contention resolution functionality. Logically, the FPGA first decodes the three label bits of each label wavelength in parallel. The 8 decoder outputs identify the packet destinations. The
outputs of each decoder are connected to the CRB as shown in Figure 4.5. Each CRB has 8 outputs represented by FPGA output digital pins, that are used to control the optical gates of the $1 \times 8$ broadcast and select switches. Each CRB controls the gates directed to the same output port of all the $1 \times 8$ switches. In this case, the wavelength selective functionality of the OPS architecture is directly performed driving the $1 \times 8$ switches according to a priority scheme. The employed priority policy is that $Ch_1 (\lambda_1)$ has the highest priority, then $Ch_2 (\lambda_2)$ and so on up to $Ch_8 (\lambda_8)$, that is the channel with lowest priority. The selected packet payload in then sent to the FWC.

The FWC is based on nonlinear polarization rotation in semiconductor optical amplifier (SOA) [85]. At the FWC output the forwarded packets are detected and evaluated by a bit-error-rate tester (BERT).

4.2.1 Experimental results

The 8 WDM channels injected in the system have central wavelength from $\lambda_1 = 1551.72 \text{ nm} (Ch1)$ to $\lambda_8 = 1557.36 \text{ nm} (Ch8)$ spaced by 100 $\text{GHz}$. The packet payload have a duration of 307 ns with a guard band of 42 ns. A single wavelength label ($L_1$ to $L_8$) is inserted in the payload spectrum bandwidth with an offset of 0.26 nm with respect of the payload central wavelength. Figure 4.6 (a) shows the WDM payloads with the correspond-
ing labels. The frequencies of the three RF tones carried by each label wavelength are \( f_1 = 175 \text{ MHz} \), \( f_2 = 365 \text{ MHz} \) and \( f_3 = 560 \text{ MHz} \). The average optical power of the payload and the label is 3.6 \( \text{dBm} \) and -6.7 \( \text{dBm} \), respectively. The FBGs employed in the experiment to separate the labels from the payload have a -3 dB bandwidth of 6 GHz centered at the label wavelengths. Figure 4.6 (b) shows the optical spectra before and after the label extraction.

![Figure 4.6: Optical spectra of payload and label of the 8 WDM channels (a) and optical spectra before and after the label extraction (b).](image)

Figure 4.7 (a) shows the electrical spectrum of a label wavelength after the extraction and the optical-to-electrical conversion. Figure 4.7 (b) shows the detected RF tones and the corresponding label bits.

Figure 4.8 (a) shows (from top to bottom) the input payloads (line 1), the outputs of the FPGA based decoders (from line 2 to 9) and the switched packets (from line 10 to 17). It is visible that the CRB performs as expected. When two or more packets are destined to the same output port (in this case output 1) the switch controller allows only one of them to be forwarded to the output enabling only one of the gates of the 1 \( \times \) 8 optical switches. The priority is given to \( Ch_1 \), then \( Ch_2 \) and so on.

Each channel is amplified to 3 \( \text{dBm} \) after the 1 \( \times \) 8 switches before being converted by the FWC. In the FWC, a continuous light wave at 1550.9 nm with 5 \( \text{dB} \) optical power is employed as a probe light. The employed SOA (CIP-XN-1550) has a gain recovery time of 10 ps and is biased at 500 m.A. Figure 4.8 (b) shows the measured bit-error-rate (BER) performance for...
the 8 WDM channels in back to back, after switching and after wavelength conversion. Error free operation with 0.6 and 1.6 \(dB\) is achieved after the \(1 \times 8\) switches and after the wavelength conversion, respectively.

The parallel implementation of the label extractors and the label processors allows to have extremely low switching times. The measured delay after the label extraction and the label bits detection is of 15 \(ns\) (measured after the LPFs), while the delay added by the decoding and the contention resolution process performed by the FPGA based switch controller is of 1
FPGA clock-cycle (10 $ns$). Thus, the overall control time of the system employing the investigated OPS architecture results to be of only 25 $ns$ and independent of the port-count.

### 4.3 Integrated 4-port optical module subsystem

To scale the studied OPS architecture to a large port-count (over 1000 ports) photonic integration is a necessity. For example, building a $1024 \times 1024$ input/output ports system (32 input fibers each carrying 32 WDM channels) employing the investigated architecture requires: 32 $1 \times 32$ optical switching modules with 32 label extractors, 32 switch controllers, 32 WS controllers and 32 wavelength converters. The employment of integrated devices, both electronic and optic components, is the only way to reduce the switch footprint, the total power consumption and thus the costs of the system.

Figure 4.9 shows the set-up used to investigate the dynamic switching capabilities of the studied OPS architecture in which an InP integrated $1 \times 4$ optical cross-connect is used as switching module.

![Experimental set-up of the $1 \times 4$ sub-system employing a $1 \times 4$ integrated cross-connect.](image)

4 laser diodes (LDs) at different wavelengths (from $\lambda_1$ to $\lambda_4$) are modulated by a 40 $Gb/s$ pattern generator (PG). The $4 \times 40$ $Gb/s$ WDM channels are first coupled and then sent to the integrated module. Synchronously, the 4 labels associated to each packet are generated by an arbitrary waveform generator and sent in parallel to the label processor and switch controller in
the electrical domain. Optical labeling and optical label extraction are not studied in this experiment. The FPGA based switch controller decodes the recovered label bits and drives the integrated device according to the packets destinations. At the 4 outputs of the optical cross-connect, the transmitted packets (after that the contention resolution has been performed) are wavelength converted by the same FWC employed in the experiment of the previous section. At the output of the FWC, the packets are detected and analyzed by a BERT.

4.3.1 Integrated optical module

The integrated device employed in the set-up as optical switching module is shown in Figure 4.10 (a). It is a InP monolithically integrated 4 × 4 optical WDM cross-connect [86]. However, only one input of the device is used during the experiment, exploiting its 1 × 4 functionalities. Figure 4.10 (b) shows the paths utilized during the experiment.

![Photograph of the 4 × 4 integrated optical cross-connect](image)

The chip has a total footprint of 4.2 mm ×3.6 mm. The cross-connect consists of a broadcast stage (BS) and four parallel wavelength selective stages (WSSs). The BS is realized with two cascaded 1 × 2 multi-mode interference (MMI) splitters. At each of the four outputs of the BS a 750 μm long SOA is placed to compensate the splitting losses. The BS outputs are fed into the inputs of the WSSs. Each WSS consists of a 1 × 4 cyclic AWG with a channel spacing of 3.2 nm (400 GHz) and four 140 μm long
SOAs. The SOAs at the AWG outputs operates as optical gates. The four outputs of a single WSS are then combined by a cascade of MMI combiners.

The SOAs gates at each output of the WSSs are used to perform the contention resolution functionality of the OPS architecture. In case that there are multiple packets present at the input of a single WSS and destined to the same output port, the switch controller drives only one SOA in order to forward just one packet to the FWC. The schematic block diagram of the integrated cross-connect stages with the switch controller is shown in Figure 4.11. The priority policy is again fixed: the packet travelling at $\lambda_1$ have the highest priority, the ones at $\lambda_4$ the lowest.

Figure 4.11: Schematic block diagram of the integrated cross-connect and its controller.

### 4.3.2 Experimental results

The 4 WDM channels are centered in $\lambda_1 = 1548.1\ nm$, $\lambda_2 = 1551.4\ nm$, $\lambda_3 = 1548.1\ nm$ and $\lambda_4 = 1557.7\ nm$. The packets have a payload duration of 290\ ns with 40\ ns guard time. The packet labels have the same payload duration and are generated according to the aforementioned RF tones labeling technique employing an arbitrary wavelength generator. Each label carries two RF tones in order to address the 4 output ports of the integrated cross-connect. All packets are synchronously transmitted with the
corresponding electrical labels. The four WDM packets, each with 6 $dBm$ of input power, are coupled and fed into the $1 \times 4$ optical module by using a commercial available lensed fiber array. The temperature of the integrated cross-connect is kept constant at $18^\circ C$.

At the same time, the label bits are detected by a label parallel processing circuit based on surface-mount device (SMD) electrical components mounted on a PCB board in order to reduce the footprint of the needed circuitry. The detected label bits are then sent to a Virtex 4 FPGA board for decoding and contention resolution purposes. The 750 $\mu m$ long SOAs are DC biased with 80 $mA$, while the WSS SOAs are directly driven by the FPGA output digital pins. The digital pins can provide around 10 $mA$ that are sufficient to properly drive the SOA gates of the WSSs.

Figure 4.12: Recovered label bit patterns, generated control signals and output packet time traces.
Figure 4.12 shows (from top to bottom) the label bit patterns of the four WDM channels (from $ch_1$ to $ch_4$) detected by the label processor circuit, the generated control signals and the switched packets. In the figure the labels with address $[0, 1]$, corresponding to the output port 3 of the integrated cross-connect, are highlighted to understand the contention resolution process.

As already mentioned, $ch_1$ (at $\lambda_1$) has the highest priority, then $ch_2$ (at $\lambda_2$), $ch_3$ (at $\lambda_3$), while $ch_4$ (at $\lambda_4$) has the lowest priority. As programmed in the FPGA based switch control, a control signal is generated only for the packets that win the contentions. These control signals are directly used to drive the SOA gates at the WSSs. As a consequence, only the selected packets are forwarded to the output of the integrated device, as visible from experiment results of Figure 4.12. Figure 4.13 (a), (b), (c) and (d) show the optical spectra after static switching of $ch_1$ (a), $ch_2$ (b), $ch_3$ (c) and $ch_4$ (d).

It is visible that there is a contrast ratio between the channels that goes from $12.4$ dB up to $17.3$ dB. The relatively low contrast ratio is mainly caused by the wavelength misalignments in the AWGs and by the limited length of the WSS SOAs. Longer SOAs at the WSSs would provide larger
absorption and higher gain, improving the extinction ratio. However, they would require also higher driving currents.

In Figure 4.14 the BER measurements in back-to-back, after switching and after wavelength conversion are reported. BER curves are taken switching continuously a single channels out of the 4 at the time. Error free operation with 1 dB and 4 dB penalty is achieved after the integrated cross-connect and after wavelength conversion, respectively.

![Figure 4.14: BER measurements in back-to-back (b-2-b), after switching and after wavelength conversion in the tested set-up employing an integrated cross-connect as optical switching module.](image)

The total chip insertion losses are around 27 dB, including the fiber coupling losses. The power of each WDM channels is 6 and -21 dBm at the input and output of the chip, respectively. The BS SOAs provide a gain around 9 dB/channel, while the WSS SOA gates contribute with a gain around 2 dB. The output signals, before being wavelength converted, are amplified to 3 dBm to exploit the nonlinear polarization rotation effect of the SOA-based wavelength converter. As a consequence, the optical signal-to-noise ratio (OSNR) is degraded, as visible from the eye diagrams of the insets of Figure 4.14. Performance improvements can be anticipated if the excess losses in the integrated optical module are substantially reduced.
4.4 Discussion

For the first time, a subsystem of the studied modular OPS architecture is implemented and tested in two different experimental set-up. The subsystem allows to investigate the dynamic switching performances of the overall architecture since each module (and the corresponding controllers) operates independently.

Labeling technique

It is important to mention the RF tones in-band labeling technique employed in this work. The technique relies on in-band (with the packet payload) label wavelengths each of them carrying binary-coded RF carriers. Using this method it is possible to address a large number of destination ports. Besides the spectral efficiency of the used labeling technique, the RF tones in-band labels allow for fast and parallel label detection and processing, performed by a label bits detection circuitry and by a FPGA based label processor and switch controller in the experiments. The overall time to detect the label bits, decode the packet destinations, solve the contentions and generate the control signals to drive the optical packet switch module is in the nanosecond scale (25 ns) and independent of the module port-count. Since all the modules can operate in parallel and autonomously, this is also the time to configure the overall switching matrix.

Experimental results

In the first experiment, the switching performances of a $1 \times 8$ broadcast and select optical module subsystem build up with discrete optical components are investigated. Label detection, label processing, $1 \times 8$ switch controlling and dynamic switching of 8 WDM channels at 40 $Gb/s$ are demonstrated to operate correctly. Error free operation is achieved with 0.6 $dB$ and 1.6 $dB$ penalty after switching and after wavelength conversion, respectively. This experiment is a proof of concept of the feasibility of the proposed novel OPS architecture and its control system. To scale the architecture to a large port-count, reducing the optical module footprint and energy consumption, and thus to reduce the system costs, photonic integration is a crucial necessity.
For this reason, the implementation of the novel OPS architecture employing an integrated $1 \times 4$ optical cross-connect is studied in a second experiment. In this case, the full functional module subsystem is shown to operate with 4 WDM channels at 40 Gb/s. Error free operation is achieved with $1 \text{dB}$ and $4 \text{dB}$ penalty after switching (at the integrated optical cross-connect outputs) and after wavelength conversion, respectively. Comparing these results with the ones obtained in the previous experiment, it is visible that around $2.5 \text{dB}$ of extra penalty are obtained employing the integrated cross-connect. This is due to the low power at the output of the chip and the subsequential OSNR signal degradation due to high signal amplification. Reducing the input/output coupling losses, will reduce the performance differences between the two studied systems. SOA-based switches are preferable over other technologies (such us Mach-Zehnder switches) due to the high number of switching elements required by the broadcast and select switch architecture considered. Moreover, SOA are easily integrable, can provide gain to compensate the splitting losses and can be driven in nanosecond time scales between the on and off state with limited driving currents. The broadcast and select architecture of the switching module is required to enable multicast and broadcast operations (not investigated in this work although feasible).

**Scalability**

Semiconductor optical amplifiers (SOAs) are commonly employed as basic switching elements for building larger switching matrices. They can provide optical gain in the ON state, large absorption in the OFF state and $ns$ switching times [87]. Moreover, SOA based switches can be integrated. Different switch architectures have been employed to realize SOA based switches, like tree or Clos architectures [88, 89]. However, scaling these switches to large port-count is still an open challenge. Many considerations can be made about the possibility of scaling these architectures to a high number of ports, like split/combine losses at the input and output, large number of SOA gates required or large number of cascaded SOAs in each stage with subsequent signal distortions. The employed switch architecture plays a dominant role in this context.

The proposed OPS architecture relies on the employment of $1 \times F$ WDM optical switching modules. Each optical switching module is based on two
cascaded stages: a broadcast stage (BS) and $F$ parallel wavelength selective stages (WSSs). The scalability of such modules is mainly limited by the splitting losses of the BS. These losses are proportional with the port-count of the system. Consider, for example, the integrated $1 \times 4$ optical cross-connect of Section 4.3.1. The BS contributes with a minimum of $6 \, dB$ on the total on-chip losses due to the two cascaded $1 \times 2$ MMI splitters. Increasing the number of ports would imply adding more cascaded MMI splitters and thus more losses. The losses of the BS scale as $3 \log_2 (F) \, (dB)$. In the employed integrated cross-connect, SOAs are employed to compensate the splitting losses of the BS. Longer SOAs could be employed in the design of a larger port-count switching module.

At the outputs of the employed chip, the WDM channels are combined by means of cascaded $2 \times 1$ MMI couplers. According to the OPS module operations, the MMI couplers of the WSS outputs couple WDM channels travelling on different wavelengths (see Figure 4.5). It is possible then to reduce the losses at this stage by replacing the couplers with an AWG. Using an AWG, the losses at the WSSs can be reduced to $3-4 \, dB$ independently of the number of wavelength channels that have to be multiplexed. However, also in this case, there are technological and fabrication impairments to consider. In fact, AWGs with high extinction ratio and low inter-channel crosstalk are required.

However, employing WDM in the OPS module operation helps reducing the number of ports required by each optical switching module. For example, scaling the proposed OPS architecture to 1024 input/output ports requires the employment of 1024 optical switching modules. Considering, for example, that each input carries 32 distinct wavelength channels the number of output ports of each $1 \times F$ optical switch is reduced to 32. In general, considering the number of input fibers $F$ equal to the number of wavelength channels $M$ carried by each fiber ($F = M = N$), the optical switching module port-count scales as $\sqrt{N}$. A $1 \times 32$ optical switch module translates in $15 \, dB$ losses in the broadcast stage, and in the need of 32-channel $100 \, GHz$ spaced AWGs with flat low loss operation and over 1000 SOAs integrated on the chip. Also, the SOAs employed to compensate the splitting losses should not introduce non-linear effects. In addition, optical signal-to-noise ratio (OSNR) degradation and chip-coupling losses have to be considered.

However, the implementation of the proposed architecture set a promis-
ing breakthrough to the realization of a large port-count integrated OPS that could be employed in a DC environment.

4.5 Summary

The dynamic switching capabilities of the studied OPS architecture are investigated for the first time in two different experimental set-up. All the functionalities of the novel architecture, with the exception of the flow control mechanism, are implemented.

The results of two experiments are presented. In the first experiment, a $1 \times 8$ switching subsystem is investigated employing LiNbO$_3$ switches as switching modules. $8 \times 40$ Gb/s WDM packets with RF tones in-band optical label are injected into the studied system. Labels extraction (based on FBGs and optical circulators), labels processing, switch control and contention resolution are shown to operate in only 25 ns. Error free operation is achieved with 0.6 and 1.6 dB penalty after switching and after wavelength conversion, respectively. The results indicates that it is feasible to realize a $64 \times 64$ (8 input fibers each of them carrying 8 WDM channels) OPS employing the tested $1 \times 8$ switching sub-system.

In the second experiment, the feasibility of a $16 \times 16$ (4 input fibers each of them carrying 4 WDM channels) OPS which employs integrated optical switching modules is demonstrated. $4 \times 40$ Gb/s WDM packets are injected into an $1 \times 4$ InP integrated optical cross-connect. Error free operation is shown with 1 and 4 dB penalty after the switch and after the wavelength conversion, respectively. The extra penalty obtained testing the $1 \times 4$ subsystem which employs the integrated cross-connect is mainly caused by the high on-chip losses.
Chapter 5

Flow control implementation

In this chapter\(^1\), the implementation of the flow control functionality is investigated. Two experiments are presented. In the first set-up, the flow control functionality is implemented in the electronic domain by buffering the packet labels. Real-time operation of a random packet generator with variable load, FIFO queue packet storing, buffer management for packet retransmission, contention resolution and fast switch configuration control are implemented by using an FPGA. The optical packet payloads are implemented with pseudo-random bit sequences from a pulse pattern generator. The flow control mechanism is successfully tested with periodic input packet sequences of 4 WDM channels at 40 Gb/s employing a fixed priority algorithm. Packet loss rate and buffer occupancy in a system with 4 and 8 input channels are then investigated under random traffic conditions and employing a round robin algorithm as priority policy. In the second set-up, an all-optical, low latency and spectral efficient flow control is implemented and tested. 2 × 40 Gb/s WDM packets with RF tones in-band optical labels are sent into a bidirectional system. The optical power of the extracted

wavelength labels is split in two parts. The first part is sent to the label processor, while the second one is re-modulated by an SOA driven by the acknowledgement message generated by the switch controller. The switch controller employs a fixed priority scheme in this case. The robustness of the flow control implementation is investigated and the packet payloads are evaluated by a BERT after the label extraction.

5.1 FPGA-based flow control

Flow controlled optical packet switch operations are investigated in the experimental set-up shown in Figure 5.1.

A time slotted system is considered. The set-up can be split in two layers. The upper layer, where the traffic generation, the packet buffering, the switch control, the contention resolution and the flow control mechanism are implemented in the electronic domain by using an FPGA. In the bottom layer, the packet payload generation and the optical switching module are fully implemented in the optical domain. At each time step, the packet payload generators create optical packets at 4 distinct WDM channels (indicated as $ch_1$, $ch_2$, $ch_3$ and $ch_4$). Synchronously, four independent traffic generators, implemented on the FPGA, generate four labels (one for each WDM channel) with uniform distributed destinations and variable input load. The input load represents the probability of the packet time slot occupancy. For example, an input load of 0.5 indicates that there is a probability of 50% of having a new packet at each time slot. Each label is stored in a FIFO queue. The queues have a limited capacity. As a consequence, if the queues are full and a new label is generated, the label and the associated packet payload are lost. The buffer managers transmit the stored labels and simultaneously provide gate signals to the optical gates of the corresponding channel. At the OPS node, the electrical labels, which determine the packet destination, are processed by the label processors in the FPGA, while the gated packet payloads are forwarded to the destination ports on the set-up optical layer. If multiple packets contend at the same time slot the same destination port, the contention solver selects, using a priority policy, only one packet while the others are dropped. Dropped packets need to be retransmitted. The contention solver generates the control signals that configure the optical module. The mod-
Figure 5.1: fpga based flow control set-up.
ule consists of a broadcast and select $1 \times 4$ optical switch. The contention solver also generates and sends to the appropriate buffer manager a positive acknowledgement message (ACK) for each packet that is successfully delivered. Each buffer manager removes the label from the queue in response to ACK. If no ACK is received the label and the payload are retransmitted after re-signalling the input optical gates, emulating packet retransmission.

5.1.1 Experimental results

The $4 \times 40 \text{Gb/s}$ OOK NRZ WDM optical payloads have central wavelengths at $ch_1 = 1548.1 \text{ nm}$, $ch_2 = 1551.4 \text{ nm}$, $ch_3 = 1554.5 \text{ nm}$ and $ch_4 = 1557.7 \text{ nm}$. Each payload consists of 1500 bytes ($300 \text{ ns}$) with 30 ns of guard band generated by modulating four LDs with a 40 Gb/s pattern generator. The corresponding packet labels are generated in the FPGA according to the RF tones labeling technique described in Section 4.1. Each label consists of two bits encoding the four OPS module outputs. The employed broadcast and select switch is the integrated $1 \times 4$ integrated cross-connect already described in Section 4.3.1 and operating under the same conditions of Section 4.3.2.

To evaluate the performance of the implemented flow control mechanism, the traffic generators are programmed to create a periodic label sequence under specific system conditions. The traffic generator associated with $ch_1$ operates at $load = 1$, thus in every time slot there is a new incoming packet on this channel. The traffic generators associated with $ch_2$, $ch_3$ and $ch_4$ are programmed to operate at $load = 0.3$. All traffic generators are set to independently assign the labels according to a 14 time slots periodic pattern. Thus, every 14 time slots the 4 different label patterns are repeated. This is done to record the packet labels using an oscilloscope. The employed input label pattern associated with $ch_1$ and the input packet patterns of all the channels are shown in Figure 5.2 (from top to bottom). The numbers inserted in the packet time traces indicate the packets destination ports.

The contention resolution algorithm employed in the contention solver is based on a fixed priority policy. Packets travelling on $ch_1$ have the highest priority, afterwards packets on $ch_2$ and so on. The contention solver generates the control and the acknowledgement signals after 3 FPGA clock cycles ($30 \text{ ns}$). Injecting the initial packet sequence in the system a periodic output sequence is achieved and recorded. Figure 5.3 shows, from top to
Figure 5.2: Label bit pattern associated with ch1 and input packet period sequence for the four WDM channels.

Bottom, the control signals for the optical gates of the four WDM channels generated by the buffer managers (including the retransmissions), the packets transmitted to the optical switching module (including the retransmissions), the control signals for the WSS 4 of the integrated cross-connect generated by the contention solver and the packets at the output 4 of the 1 × 4 integrated switch.

The packets transmitted to the OPS module and destined to the output 4 of the system are highlighted in the figure to understand and evaluate the flow control and the retransmission mechanism. The contention solver performs as expected. The packets travelling on ch1, having higher priority are directly forwarded to the output of the systems. Also, the flow control mechanism is shown to work properly, the packets that lose contentions are correctly retransmitted to the optical switching module and, in case no more contentions occur, are forwarded to the output. The load increment caused by the retransmission process is evident. For example, ch3 with a starting load of 0.3 has a total load including the retransmission equal to 1. This behavior has been anticipated in the real traffic definition (load_R) of Section 2.2.

Clearly the system configuration employed does not allow equal utilization of the four FIFO buffer queues. The buffer occupancy and the packet loss rate are investigated employing a different system configuration. The traffic generators are programmed to create random sequences of packet labels, according to 2^{31} – 1 pseudo-random sequences with distinct seed states. Moreover, to obtain queues of equal length at the input buffers,
a round robin algorithm is employed as priority policy in the contention solvers. The buffer occupancy and the packet loss rate are investigated to understand the effects of the port-count on the buffering requirements in systems with 4 and 8 output ports and considering a buffer queue capacity \( B \) of 16 and 64 packets. Figures 5.4 (a) and (b) show the typical buffer occupancy in a system with 4 output ports \( (B=16) \) and the packet loss ratio as function of load for systems with 4 and 8 ports, considering \( B = 16 \) and 64, respectively.

In Figure 5.4 (a) is visible the \( \text{ch}1 \) buffer occupancy time evolution in the first 150 time slots considering the 4 WDM channels operating at input load is equal to 1, 0.7 and 0.5. The maximum number of queued packets is 15,
due to the fact that the served packet is not considered. Equivalent results, not reported, are obtained for the other three WDM channels. Balanced buffer utilization is a consequence of the employment of the round robin algorithm in the contention solver. It is clear that with load = 1 the buffer is rapidly filled up and there are not empty time slots to drain the buffer queue, causing buffer overflowing. With load = 0.7 the buffer queue
increases as function of the contentions. The buffer queue decreases when there are empty time slots that can be used to successfully retransmit the dropped packets. With $load = 0.5$, the buffer is never full during the studied time slots, suggesting that a buffer queue capacity of 16 packets is enough to avoid packet losses in the system under this traffic load.

Figure 5.4 (b) shows the packet loss performance of the system considering $10^{12}$ time slots. Once again, the Figure refers to $ch1$ and the equivalent results for the other channels are omitted. The curves are obtained considering an OPS switching module with 4 and 8 output ports and considering a buffer capacity of 16 and 64 packets. The results reveal that, for $load \geq 0.8$, the packet loss is port-count and buffer capacity independent. Considering lower $load$, the packet loss performance of the system improves increasing the buffer capacity and degrades increasing the number of ports. For example, with $load \approx 0.6$ a buffer capacity of 16 packets allows a packet loss around $10^{-2}$ for both the 4-port and the 8-port cases. At the same $load$ value a buffer capacity of 64 packets ensures a packet loss around $10^{-7}$ for a 4-port system and a packet loss around $10^{-5}$ for a 64-port system. This behavior is in agreement with the results of the simulated systems of Chapter 3.

5.2 All-optical flow control

Although having separated electrical paths that enable the flow control functionality of the studied OPS architecture is a valid solution, an efficient all-optical flow control implementation is preferable to reduce the number of wired interconnections in a large scale system.

The idea is to use the same wavelength utilized to transmit the packet label to send back to the transmitting node the optical acknowledgement message in order to limit the number of required interconnections. This is investigated employing the bidirectional system depicted in Figure 5.5. $2\times40$ Gb/s WDM channels, indicated with $ch1$ (at $\lambda_1$) and $ch2$ (at $\lambda_2$) with RF tones in-band optical labels are sent into the OPS packet switch system under investigation after being multiplexed by an AWG. At the OPS input the WDM channels are demultiplexed. The label of each channel is extracted from the payload by employing an optical circulator and a FBG with central wavelength at the label wavelength. The packet payloads re-
main in the optical domain and are fed into the $1 \times F$ optical switching modules. 80% of the optical power of the extracted optical label is O-E converted and processed by the FPGA switch controller. The other portion of the extracted label power (20%) is fed into a SOA, which is employed as optical modulator. The SOA-based modulator is driven by the electronic acknowledgement signals (ACK) generated by the switch controller. Using this set-up, it is possible to implement an all-optical flow control mechanism. The optical ACKs travel back to the transmitting node on the same wavelength employed to transmit the labels and using the same optical paths. This is possible due to the employment of a second group of optical circulators and FBGs with similar characteristics as the ones utilized for the labels extraction. On the transmitting node side, employing an optical circulator, the optical ACK is sent to the ACK receiver (ACK RX), where is O-E converted and easily retrieved by using a low-pass filter (LPF).

![Figure 5.5: Experimental set-up of the all-optical flow control implementation under investigation.](image)

### 5.2.1 Experimental results

The $2 \times 40$ Gb/s WDM payloads are centered in $\lambda_1 = 1544.9 \text{ nm}$ and $\lambda_2 = 1548.0 \text{ nm}$ and have a duration of 300 ns. The corresponding optical labels have central wavelengths at 1544.1 nm and 1548.2 nm, respectively. Each wavelength label carries two RF tones at the frequencies $f_1 = 280 \text{ MHz}$ and $f_2 = 650 \text{ MHz}$. The average optical input power is 2.5 dBm for the payloads and 0 dBm for the labels.

At the OPS input the labels are extracted employing FBGs centered
on the same label wavelengths and having a 3 dB bandwidth of 6 GHz in order to avoid spectral distortion of the optical payloads. Figures 5.6 (a) and (b) show the optical spectra of the packets payload and label, before and after the label extraction, of the two WDM channels \( ch_1 \) and \( ch_2 \).

![Optical spectra before and after the label extraction of the WDM packets on channel 1 (a) and channel 2 (b).](image)

The power of the extracted optical label is split by a 20:80 coupler. The 80% of the power (-8 dBm) is used to process the label in the switch controller implemented as in Section 4.2. Figures 5.7 shows the electrical spectrum of the detected optical label. The two RF tones at 280 MHz and 650 MHz are clearly visible.

![Electrical spectrum of the recovered optical label.](image)

The contentions are solved in the switch controller employing a fixed priority policy. Packets on \( ch_1 \) have always higher priority than the ones
travelling on \( ch_2 \). The switch controller generates the control signals for driving the \( 1 \times F \) optical switching modules and the acknowledgement messages for the flow control process. These signals are used to drive the SOAs employed to generate the optical ACKs. The typical driving current is 30 \( mA \), which can be provided directly by the FPGA digital pins. Figure 5.8 shows from top to bottom: the label bit patterns of the two channels considered, the electrical ACKs generated by the switch controller and the ACK signals detected at the transmitting node after the LPF. The label bits where contentions occur are labelled in the figure. It is visible that in these time slots the switch control generates a positive acknowledgement message for \( ch_1 \) and a negative ACK for \( ch_2 \). The ACK signals are correctly transmitted back to the input nodes.

![Figure 5.8: From top to bottom: the label bit patterns of the two channels considered, the electrical ACK signals generated by the switch controller and the ACK signals detected after the LPFs.](image)

The electrical spectra of the recovered optical ACK before and after the low pass filter (LPF) are shown in Figures 5.9 (a) and (b), respectively.

The packet payloads are evaluated, after the label extraction, by a BERT. Figure 5.10 (a) shows the bit-error-rate measurements in back-to-back (b2b) and after the label extraction for the two WDM channels considered. It is visible that error free operation is achieved with a power penalty of only 0.5 \( dB \). This suggests that there is not distortion due to
Figure 5.9: Optical spectra of the recovered optical ACKs before (a) and after (b) the low pass filter.

Figure 5.10: Bit-error-rate measurements in back-to-back (b2b) and after the label extraction of the two WDM channels considered (a) and eye-opening factor and amplitude of the detected ACK signal as function of label power fed into the SOA based modulator and the modulator driving current (b).

The eye-opening and the amplitude of the detected ACK as function of the label power fed into the SOA-based modulator and the SOA driving current are shown in Figure 5.10 (b). This is done to understand the effects of power fluctuations (for example caused by different host-switch
distances) on the implemented flow control mechanism. Figure 5.10 (b) shows that driving the SOA-based modulator with a current of 30 \( mA \) ensures an eye-opening factor higher than 0.8 with input power ranging from -30 \( dBm \) to 18 \( dBm \). Even larger dynamic range is achievable with higher driving currents. As a conclusion, the system is robust to power fluctuations and thus to distance variations within a DC network.

5.3 Discussion

Two distinct methods to implement the flow control functionality required to enable packet retransmission are presented in this chapter.

An FPGA-based flow control implementation is tested successfully under dynamic switching operation. The effects of packet retransmission on the system load is clearly visible adopting a fixed priority policy in the contention resolution process. A channel with low priority with a starting load of 0.3 is rapidly full loaded due to the multiple packet contentions. This phenomenon highlights the need of a more complex priority algorithm or the need of a balanced traffic pattern within the investigated system. A round robin algorithm is then used to investigate the performance in terms of packet loss of the implemented system under variable system load. The round robin algorithm allows balanced buffer occupancy and reduces the maximum latency that a packet can experience in the system. In fact, it guarantees that each packet (except the ones lost due to buffer overflowing) is delivered to the appropriate output port after a maximum of three lost contentions, since within the fourth attempt it will be travelling on the channel with highest priority.

The system studied shows no losses for input load up to 0.4, this value is higher than the typical DC traffic load. Increasing the system port-count (from 4 to 8), the packet loss performance degrades. However, this degradation becomes negligible under heavy traffic load (\( \geq 0.6 \)). On the contrary, better packet loss results are achieved considering a larger buffer capacity. Increasing the buffer capacity from 16 to 64 packets lead to lossless systems for load up to 0.5 and 0.6 for the 8-port case and the 4-port case, respectively.

However, an all-optical solution is preferable since it does not require a dedicated interconnections network to be implemented. A large-scale sys-
tem such as a data center requires a large number of wired interconnections. Exploiting added functionalities on the already deployed optical cables is a valid solution.

The proposed all-optical implementation of the flow control functionality is based on the reutilization of part of the optical power of the optical packet label. Few extra components are required for its implementation and the same optical path used for the packet transmission is employed also for the transmission of an optical acknowledgement back to the transmitting node. This solution is resource and spectral efficient.

This implementation is experimentally tested and the results show that it does not degrade the packet payload quality as confirmed by the bit-error-rate curves. Also, the limited portion of optical label power (20%), employed to generate the optical acknowledgement, does not affect the label detection.

A limited number of extra components are required for this implementation: a semiconductor optical amplifier (per channel) is required to modulate the optical label wavelength and a fiber Bragg grating in combination with an optical circulator is needed to couple the generated signals back in the packet transmission path. This implementation requires also an additional receiver (in combination with a low-pass filter) at the transmitting node to detect the optical acknowledge message. The all-optical flow control implementation is also shown to be robust to power fluctuations.

The functionalities implemented by using fiber Bragg gratings and optical circulators (label extraction and optical acknowledgement generation) can be implemented using integrated microring resonators (MMRs). The implementation of an integrated label extractor based on MMRs is discussed in the next chapter of this thesis.

5.4 Summary

Two different implementations of the flow control mechanism that allows packet retransmission in a DC inter-cluster network which employs the OPS architecture under investigation are presented. In the first one, a dedicated electronic network is employed for the acknowledgement signals generation and transmission. Traffic generation, packet label buffering, label processors, switch control, contention resolution functionality and acknowl-
Summary

Edgement signals generation and transmission are implemented by using an FPGA. Optical packet payload generation and switching process are performed in the optical domain. 4 × 40 Gb/s WDM packets are sent into an integrated optical cross-connect. The flow control functionality is successfully implemented and demonstrated. The system is tested injecting periodic packet patterns under constrained system configurations to obtain a periodic response, needed to record the signals. Buffer requirements and packet loss performance are investigated considering a random input traffic and different buffer capacities. The results reveal that the performances of the system are port-count and buffer-size independent for load ≥ 0.8. For lower load the performance degrades increasing the port-count and decreasing the buffer capacity.

Having a dedicated electrical network for the flow control functionality may be cumbersome considering the number of wired interconnections required in a large scale data center. For this reason, the second experimental set-up is used to investigate an all-optical solution. The employed bidirectional system utilizes part of the power of the extracted optical packet label to generate the optical ACK. A SOA-based modulator driven by the FPGA controller and optical circulators in combination with FBGs are used to generate and transmit the optical acknowledge signals employing the same wavelength and the same optical path of the optical packet label. At the receiver side (the packet transmitting node) the optical ACK is O-E converted and retrieved by employing a LPF. The all-optical implementation is shown to operate correctly injecting 2 × 40 Gb/s WDM packets with in-band labels in the system. The packet payloads are evaluated, after the label extraction, by means of BER measurements. Error free operation is achieved with a penalty of only 0.5 dB. The robustness of the optical ACK, in terms of amplitude and eye-opening, as function of the power of the extracted label used for the ACK generation and as function on the SOA-based modulator driving current is investigated. The results reveal that, considering a driving current of 30 mA, the eye-opening factor is larger than 0.8 with an input power range between −30 dBm and −18 dBm. Even larger power ranges can be achieved driving the SOA based modulator with higher currents. Thus, this flow control implementation results to be robust to possible power fluctuations in a DC environment.
Chapter 6

Towards the total OPS integration

In this chapter, the photonic integration of the novel OPS architecture is investigated. Toward this goal, the performances of a SOI integrated label extractor are investigated in an experimental set-up. The studied device consists of four cascaded microring resonators. Three in-band RF tones label wavelengths are coupled with a 160 Gb/s packet payload and sent into the chip for the label extraction process. Extracted labels are processed by a FPGA-based switch controller, while packet payload is evaluated by a BERT after the filtering process. In a second experimental set-up, a SOI integrated device which includes an AWG for WDM channel demultiplexing and microring resonators for label extraction is tested. 3 × 40 Gb/s WDM payloads are coupled with an optical label at 1.3 Gb/s and fed into the chip. The performances of the integrated device in terms of channel demultiplexing and label extraction are investigated and reported.

6.1 Introduction

In this chapter, the focus is on the total integration of the novel OPS architecture. Figure 6.1 shows the functionalities required by each module of the novel architecture, considering the all-optical implementation of the flow control system.

![Figure 6.1: Functionalities required by a single module of the novel architecture.](image)

Each module requires several functional blocks:

- Channels demultiplexer: this functionality can be easily implemented by using an array waveguide grating (AWG). It is employed to demultiplex the $M$ WDM channels carried by each input fiber. It allows processing the packets travelling on different wavelength in parallel.

- Label extractor: it allows separating the optical label from the packet payload. The extracted label is then optical-to-electrical (O-E) converted and sent to the label processor. Also, part of the extracted label power is employed to generate the optical acknowledge (ACK) needed to implement the flow control functionality.

- $1 \times F$ optical switch: it forwards the packet to the appropriate output port.

- Label processor and switch controller: it is used to identify the packet destination, to drive the optical switch according to the packet destination, to solve the contentions, to drive the wavelength selectors (WSs) and to control the flow control ACK generation.

- $M \times 1$ wavelength selector (WS): it selects one out of a maximum of $M$ packets in case that multiple packets have the same output destination.
• Fixed wavelength converter (FWC): it converts the packets to a fixed wavelength. It allows avoiding contentions between different modules of the studied OPS architecture.

6.2 Integrated optical switch

In Section 4.3, the implementation and the switching capabilities of the proposed OPS architecture employing a $1 \times 4$ integrated switching module are shown. Moreover, the employed device allows performing the wavelength selection functionality. In Section 4.4, the scalability of such module is discussed. Recently, a $8 \times 8$ integrated optical interconnect based on the same architecture of the $1 \times 4$ module employed in this work has been designed, fabricated and tested [94]. In this device, shown in Figure 6.2, an extra stage of parallel SOAs is employed at each input to pre-amplify the incoming optical signals.

![Figure 6.2: 8 × 8 integrated optical cross-connect photograph [94].](image)

Similar to its 4-port counter part, the chip features a MMI-based broadcast stage (BS) and 8 (4 in the 4-port chip) parallel wavelength selective stages (WSSs). Compared to the 4-port device, longer SOAs are used to compensate the extra MMI-based shuffle network losses. Each WSS consists of a $8 \times 8$ AWG and the SOAs at each AWG output operate as optical gates to enable the wavelength selective functionality. The total footprint of the $8 \times 8$ cross-connect is $14.6 \text{ mm} \times 6.7 \text{ mm}$. The employment of 8 optical modules like the one described, operating in parallel and autonomously, would allow the implementation of a $64 \times 64$ OPS based on the proposed ar-
chitecture, considering that 8 WDM channels can be simultaneously routed by the switching module.

However, the integration of the $1 \times F$ switching modules is just the first step towards the total integration of the studied OPS architecture. To have a finished product with a compact design, all the functionalities required by the novel architecture need to be integrated in a single device. Towards this goal, further steps are taken in this chapter.

### 6.3 Integrated label extractor

In this section, the performances of an integrated label extractor, which is suitable for the implementation of the studied OPS architecture, are experimentally investigated. The employed integrated device, shown in Figure 6.3, consists of four high-quality factor (Q) add-drop single-ring microring resonators integrated on silicon on insulator (SOI) [95]. Each microring is defined as a rib-type waveguide using a partially etch process of 70 nm, height and width of the rib are 220 nm and 600 nm, respectively.

![Figure 6.3: Photograph of the SOI integrated microring resonators label extractor.](image)

The device features four drop ports (one for each microring resonator) and one pass-through port. Each microring is designed with a radius of 24 $\mu m$ and a straight racetrack length of 4 $\mu m$ to obtain a free spectral range (FSR) of 3.86 $nm$ and low bending losses. Although equally designed, the resonance wavelengths of these adjacent rings may vary up to 1 $nm$ due to local non-uniformities of the etch process. Figure 6.4 shows the wavelength response of the employed microring resonators array.

Two of the four resonance frequencies of the microrings (channel 2 and channel 3) are overlapped for the aforementioned reason. This limits the number of label wavelengths that can be used during the experiment to three. However, integrated heaters such as the ones employed in [96] can
solve this problem and expand the number of useful label wavelengths to four.

### 6.3.1 Dynamic label extraction

The time-slotted set-up employed to investigate the performance of the SOI integrated label extractor is shown in Figure 6.5.

The system operates as follows. The 160 Gb/s packet payload is generated by time multiplexing $4 \times 40$ Gb/s modulated signals. Synchronously, three optical in-band label wavelengths are generated modulating three lasers by an arbitrary waveform generator according to the RF-tone labeling technique described in Section 4.1. After being coupled, payload and label wavelengths are fed into the input port of the microring resonators array. The label wavelengths are extracted by the microring resonators and directed to the drop ports of the label extractor, while the packet payload is forwarded to the $1 \times F$ optical switch through the pass-through port. Once coupled out of the chip, the extracted labels are optical-to-electrical converted and fed into the label processor. Once the label bits are detected and processed, the FPGA-based switch controller generates the control signals
for the broadcast and select $1 \times F$ optical space switch. The label generation, label processor and the FPGA based switch controller are implemented as described in Section 4.2. The dynamic label extractor operations are evaluated by injecting a periodic label sequence with a period of 8 time slots. At the pass-through port the packet payload is detected and evaluated by a BERT.
### 6.3.2 Experimental results

The 160 Gb/s packet payload has central wavelength at 1552.5 \( nm \), while the three label wavelengths have central wavelengths at 1549.8 \( nm \), 1550.45 \( nm \) and 1550.67 \( nm \), respectively. Each of the label wavelengths carries three RF tones with frequencies \( f_1 = 90 \text{ MHz} \), \( f_2 = 280 \text{ MHz} \) and \( f_3 = 650 \text{ MHz} \). As a consequence, up to \( 2^{3 \times 3} = 2^9 = 512 \) output ports of a \( 1 \times F \) optical switching module can be addressed \( (F_{\text{max}} = 512) \).

The packets have a duration of 330 \( ns \) and the labels are generated according to a 8 time slots periodic pattern. Packet payload with 2.5 \( dBm \) of input power and packet labels with -4.7 \( dBm \) of input power (per wavelength) are coupled into the SOI integrated chip by means of grating couplers.

The spectrum of the signal at the input of the SOI integrated label extractor and the spectrum of the signal at the pass-through port are shown in Figures 6.6 (a) and (b), respectively.

![Figure 6.6](image)

Figure 6.6: Spectrum of the signal at the input of the SOI integrated label extractor (a) and spectrum of the signal at the pass-through port (b).

The power of the payload and the labels at the output of the chip are -11 \( dBm \) and -17.5 \( dBm \), respectively. Thus the total on-chip losses are around 13.5 \( dB \) including 12 \( dB \) of the input/output grating couplers and 1.5 \( dB \) due to crossing the four microrings resonators. The coupling losses could be reduced to 3.2 \( dB \) as shown in [97].

Figures 6.7 (a), (b) and (c) show the spectra of the signals at the drop ports 1, 2 and 4, respectively. In the figures are clearly visible the extracted
label wavelengths and the portion of payload extracted by the second FSR of the microring filters, which could be used to insert more label wavelengths and scale further the number of addressable output ports.

Figure 6.7: Spectra of the signals at the drop port 1 (a), drop port 2 (b) and drop port 4 (c).

Figure 6.8 shows, from top to bottom, the recovered bit label pattern and the switched packets at the output of a $1 \times 8$ optical broadcast and select space switch. In the experiment only 3 label bits, corresponding to a single label wavelength, are processed by the FPGA based switch controller. They are sufficient to address $2^3 = 8$ output ports of the employed optical switch.

Figure 6.9 shows the BER measurements of the four demultiplexed 40 Gb/s signals in back-to-back ($b - 2 - b$) and at the pass-through port of the SOI integrated label extractor.

Error free operation is achieved with only 1 $dB$ penalty. This result
Figure 6.8: From top to bottom: recovered label bits of a single label wavelength and packet switched at the 8 output ports considered.

Figure 6.9: BER measurements in back-to-back (b-2-b) and at the chip pass-through port after the labels extraction. In the insets: eye-diagram of the demultiplexed 40 Gb/s in back-to-back and after the label extraction.

is promising compared to the results obtained using fiber Bragg gratings and optical circulators for the label extraction process. In [84] 0.3 dB of penalty are obtained on a 40 Gb/s signal after the carving of one fiber Bragg grating, i.e. extracting only one label wavelength. Clean open eye-diagram is visible after the label extraction. However, a minimal pulse broadening
is visible after the filtering process, as shown in the insets of the figure.

6.4 Integrated demultiplexer and label extractor

According to the OPS architecture operations, at each input of the switch, an optical fiber carries $M$ distinct wavelength channels. Before extracting the optical labels corresponding to the packets travelling on distinct wavelengths, the WDM channels need to be demultiplexed.

In this section, the possible integration of a device that allows channel demultiplexing and label extraction is experimentally investigated. The required device is schematically depicted in Figure 6.10 (a), while the SOI integrated circuit employed to demonstrate the possible integration of the required functionalities is schematically shown in Figure 6.10 (b).

The studied device is fabricated on a SOI platform that features a 220 nm thick silicon layer on a 2 $\mu$m thick buried oxide layer. The chip consists of a $1 \times 32$ AWG with 200 GHz channel spacing and narrow band microring resonator (MRR) for label extraction at the AWG input waveguide. Rib waveguides are used to minimize backscattering which limits the MRR Q-factor. A shallow etch process of 70 nm is used to realize the rib, input/output fiber grating couplers and the AWG star-couplers. Shallowly etched waveguides apertures are used as inputs and outputs of the AWG star-couplers to reduce reflections. An array of 90 waveguides is used to realize the AWG. The MRR has a radius of 35 $\mu$m that corresponds to a FSR of 2.92 nm. Figure 6.11 shows the spectrum of the isolated MRR.
Integrated demultiplexer and label extractor

The extinction ratio of the through-port is of -22.5 dB and the insertion loss of the drop port is less than 0.5 dB. The 3 dB bandwidth of the filter is very narrow (77 pm) to avoid payload distortion.

6.4.1 Dynamic demultiplexing and label extraction

The set-up employed to investigate the performances of the SOI integrated channel demultiplexer and label extractor is shown in Figure 6.12.
The system operates as follows. Three packet payloads on three distinct WDM channels (indicated by $Ch_{-1}$, $Ch_0$ and $Ch_{+1}$) are modulated at 40 Gb/s by a bit pattern generator. At the same time, an in-band optical label is generated in the spectrum of $Ch_0$ by modulating a laser diode (LD) with a 1.3 Gb/s bit pattern generator. Label and payloads are coupled and the resulting signal is fed into one of the 32 AWG outputs (corresponding to the central wavelength of $Ch_0$) of the SOI integrated chip. The goal of the experiment is to successfully demultiplex $Ch_0$ from the other two channels and extract the in-band optical label. The channels are demultiplexed by the integrated AWG. At the AWG input, the demultiplexed $Ch_0$ is fed into the label extraction stage of the integrated device. The label is extracted by the MRR and sent at the output of the chip through the corresponding drop port, while the payload is coupled out by means of the corresponding pass-through port. The signals at the drop port and pass-through port are detected and evaluated by a BERT.

### 6.4.2 Experimental results

The $3 \times 40$ Gb/s packet payloads have central wavelengths at 1548.16 nm ($Ch_{-1}$), 1549.67 nm ($Ch_0$) and 1551.15 nm ($Ch_{+1}$). The 1.3 Gb/s optical label is inserted in the $Ch_0$ optical spectrum and centered in 1549.08 nm. The input power of each optical payload is of 1.8 dBM and the label is transmitted with an input power of -6 dBM. Figures 6.13 (a) and (b) show the optical spectrum of the characterization of the cascaded AWG demultiplexer and label extractor and the spectrum of the signal at the input of the SOI integrated chip, respectively.

In Figure 6.13 (a) the resonance peak of the MRR is clearly visible in $Ch_0$ response of the AWG. The resonance peak of the microring based label extractor has an offset of 0.58 nm with respect of the central wavelength of the packet payload on $Ch_0$.

Figures 6.14 (a) and (b) show the spectra of the signal at the pass-through port and the signal at the drop port, respectively.

In Figure 6.14 (a) is visible that $Ch_0$ is successfully demultiplexed. At the pass-through port, $Ch_0$ shows an extinction ratio relative to the adjacent channels ($Ch_{-1}$ and $Ch_{+1}$) of -20.4 dB and -17.6 dB, respectively. The label has an extra -23 dB drop compared with the starting conditions, for a total of -30.9 dB difference with the $Ch_0$ payload power. The total
Figure 6.13: Optical spectra of the characterization of the cascaded AWG demultiplexer and label extractor (a) and of the signal at the input of the chip (b).

Figure 6.14: Optical spectra of the signal at the pass-through port (a) and at the drop port (b).

insertion loss for the studied WDM channel is of 14.8 dB, which includes the insertion loss of two fiber grating coupler (±6 dB), waveguide losses (1 dB) and AWG losses (1.8 dB). In Figure 6.14 (b) is visible a 12 dB power ratio between the extracted label and the payload. At the drop port, the extinction ratio between the label and the other payloads is more than 30 dB.

Figures 6.15 (a) and (b) show the BER measurements of the packet payload at the pass-through port and the BER curve of the packet label.
at the drop port, respectively. In both cases, error free operation with a penalty around 0.5 $dB$ compared to the back-to-back (b-2-b) measurements is achieved. Clear open eye-diagrams are visible in the insets of the pictures. These results confirm that the channels demultiplexing and the label extraction functionalities can be implemented on photonic integrated circuits with limited penalty on the packet payload and no effects on the label detection.

### 6.5 Discussion

In this thesis the integration of some functional blocks required by each module of the studied OPS architecture has been demonstrated. The integration of the $1 \times F$ optical switch has been shown in Chapter 4 employing a InP integrated $4 \times 4$ cross-connect. Moreover, a $8 \times 8$ InP integrated device based on the same architecture is mentioned in Section 6.2. In this chapter, it is shown that channel demultiplexing and label extraction can be performed with limited penalty employing SOI photonic integrated circuits.

Concrete opportunities for the integration of the other functional blocks required by the studied OPS architecture are now discussed.
6.5.1 Label detection

In this chapter the possible integration of the channels demultiplexing and the label extraction functionalities has been demonstrated by employing two SOI integrated circuits.

The SOI platform allows the integration of high-efficiency photodetectors and trans-impedance amplifiers. In [98] a photonic integrated device with Germanium photo-detectors and trans-impedance amplifiers (TIAs) on SOI is presented. Alternatively, heterogenous integration of high quality III-V materials on top of SOI by means of molecular or adhesive bonding is possible. In [96], a 4-channel WDM demultiplexer with III-V photodiodes integrated on SOI is presented.

A photonic integrated circuit with array waveguide gratings (AWGs), microring resonators (MRRs), photodetectors and CMOS amplifiers will allow the extraction, the detection and the amplification of the in-band optical labels considered in this work. Main requirements in such a device are high extinction ratio in the AWG to avoid channel crosstalk during demultiplexing and narrow band for the MRR-based filters to avoid payload distortions in the label extraction process.

SOI integrated microring resonators can be employed to implement also the bidirectional signals transmission required by the all-optical flow control described in this thesis.

6.5.2 Label processor and switch controller

In the experiment presented in Chapter 4 the label processor and the switch controller have been implemented by using a dedicated label bits detection circuitry and a field-programmable gate array (FPGA) board. The circuitry employed to detect the label bits is based on surface-mount device (SMD) components mounted on a printed circuit board (PCB). The PCB is connected to the digital pins of the FPGA board. The FPGA board processes the detected label bits and, after contention resolution has been performed, drives the optical switch (including the wavelength selectors) according to the decoded label destinations.

The label processor and switch control requires limited computational power. The modular structure of the studied OPS architecture and the employed labelling technique allow the parallel processing of each packet.
label. Moreover, each label bit can be processed in parallel.

CMOS technology, compatible with SOI photonic integrated circuits, can be employed to integrate the label processor and switch controller. The only limiting requirement of this control system is the large number of required inputs/outputs. Many signals need to be processed in parallel and the number of needed in/out pins is dependent on the switch port-count.

6.5.3 Fixed wavelength converter

In the experiment carried out in this thesis, wavelength conversion has been performed exploiting non-linear effects in semiconductor optical amplifiers (SOAs). Non-linear polarization rotation in a SOA has been used to convert the WDM packets. High data-rate wavelength conversion employing this technique has been demonstrated [32].

Other non-linear effects in SOA, such as cross gain modulation (XGM) or cross phase modulation (XPM) for example, can be used to perform SOA-based wavelength conversion. The integration of the wavelength conversion functionality is possible. In [99], a fully InP monolithically integrated wavelength converter has been presented. This device combines a SOA and a fast tunable multi-frequency laser, and error free operation at 40 Gb/s for different wavelengths is demonstrated.

6.5.4 Conclusions

Photonic integration of a large port-count switch including all the required devices and functionalities on a single chip remains technologically demanding either for the fabrication and the electrical wiring.

In this thesis, two material platforms, namely SOI and InP, are considered. Hybrid integration will allow to take advantage of the benefits provided by the different material platforms. CMOS compatibility and limited footprint size of devices fabricated on SOI platform can be combined with the opportunity of integrating active functionalities on III-V materials.

It is important to highlight that the photonic integrated circuits employed in the experiments reported in this work have not been designed and fabricated specifically for the proposed optical packet switch architecture. Better results, in terms of total system losses could be achieved by
the design and the fabrication of dedicated photonic integrated circuits.

6.6 Summary

Further steps towards the total integration of the OPS architecture under investigation are taken in this chapter. The possible integration of the label extractor is demonstrated. The performances of a SOI integrated label extractor are investigated in a experimental set-up. The SOI chip consists of an input port, four cascaded microring resonators, a pass-through port and four drop ports. The chip is tested injecting at its input a 160 Gb/s WDM payload coupled with three label wavelengths each of them carrying three RF tone bits. The label bits can address up to 512 output ports. The label wavelengths are successfully extracted by the microring resonators and coupled out through the drop ports of the chip. Three label bits, associated with one of the label wavelengths, are then processed by a FPGA based label processor and switch controller. The packet payload, after the label filtering process, is coupled out of the chip through the pass-through port and evaluated by a BERT. Dynamic label detection and processing is shown to operate correctly and error free operation of the system is shown with 1 dB penalty on the packet payload.

In a second experiment presented in this chapter, the channel demultiplexing functionality needed at each input of the studied OPS architecture is investigated in combination with the label extraction employing a SOI integrated demultiplexer and label extractor. The employed SOI chip features a 1 × 32 AWG with 200 GHz channel spacing that is used to demultiplex 3 × 40 Gb/s WDM packet payloads. At the AWG input, a microring resonator allows to implement the label extraction functionality on chip. A single label wavelength, inserted in the spectrum of one of the WDM payloads, is extracted by the microring based label extractor at the input of the AWG. BER measurements on the packet payload and the packet label are used to investigate the performance of the integrated packet demultiplexer and label extractor at the drop port and at the pass-through port of the chip, respectively. In both cases, error free operation is achieved with 0.5 dB penalty compared to the back-to-back measurements, confirming the possible integration of the demultiplexing and label extraction functionalities on a single device.
The integration of the other functionalities required by the proposed optical packet switch architecture is discussed in this chapter. The total OPS integration is still an open challenge. Scaling the OPS switch port-count, multiple devices and functionalities need to be integrated on the same photonic device. Improvements in photonic integrated technologies may allow the fabrication of such large-scale photonic chip.
Chapter 7

Conclusions

The goal of this work was to identify, study and implement an optical packet switch architecture for high performance data center networks. Such a switch is required to handle a large number of high data-rate ports to enhance the resources sharing capabilities in these systems. Also, it should add negligible latency to the system to improve the system performance.

A novel single-chain model that allows computing a multi-port system throughput and latency as function of the employed switch architecture and as function of the switch port-count is used as a tool to identify a suitable architecture. This study highlights the need of a distributed control to obtain low switch reconfiguration time, while scaling to a large port-count. As a result, a novel architecture is presented. It is a modular strictly non-blocking architecture. Wavelength division multiplexing is used to scale the switch port-count. The total number of inputs/outputs is given by the number of input fibers $F$ multiplied by the number of wavelength channels carried by each fiber $M$. Wavelength conversion is used to partially solve packet contention “on the fly” in the optical domain. Also, it allows parallel and autonomous operations of the architecture modules. Flow control functionality is provided in case packet contentions cannot be solved in the optical domain. The modular structure of the novel architecture and the parallel modules operations enable the employment of highly distributed control, which in turn allows for port-count independent nanosecond-scale switch reconfiguration time.

The performances of the novel architecture in a data center scenario are
investigated by means of simulations. The architecture is shown to handle synchronous and asynchronous traffic providing sub-microsecond average end-to-end latency, low packet loss and high throughput under relatively high traffic load and using a limited buffer capacity. The simulations results in terms of throughput are in agreement with the ones obtained by applying the single-chain model, while discrepancies are evident for the latency results. This is due to the fact that the model does not consider the buffer capacity that strongly affects the system latency. The effects of the buffer capacity on the system performances are investigated with an interesting result. Under heavy traffic conditions, increasing the buffer capacity does not improve the performance of the system in terms of packet loss, resulting also in increased system latency due to longer queueing times.

The architecture module subsystem is implemented for the first time in experimental set-up, not considering the flow control functionality. The feasibility of a $64 \times 64$ optical switch based on the novel architecture is demonstrated by testing the dynamic switching capabilities of a $1 \times 8$ subsystem based on discrete optical components. The integration of the $1 \times F$ optical switch required by each architecture module is exploited employing a $1 \times 4$ InP integrated optical cross-connect. $4$ parallel chips like the one employed would allow the implementation of a $16 \times 16$ optical packet switch. The two subsystems are studied injecting $8$ ($4$) WDM channels at $40 \, \text{Gb/s}$. Each $1 \times 8$ ($1 \times 4$) module can provide $320 \, \text{Gb/s}$ ($160 \, \text{Gb/s}$) of maximum throughput. This translates in a maximum throughput of over $2.5 \, \text{Tb/s}$ ($640 \, \text{Gb/s}$) considering a $64 \times 64$ ($16 \times 16$) switch based on the novel architecture. The radio frequency tone in-band label detection, after label extraction, is performed by a dedicated printed circuit board based on pass-band filters and envelope detectors. The label processor, the switch controller, the contention resolution functionality and the wavelength selector controller are implemented by programming an FPGA. In both cases studied, the switch subsystem configuration takes place in only $25 \, \text{ns}$ due to the employed parallel labelling and label processing techniques, regardless the number of label wavelengths and tones employed to encode the packet address. This result confirms that the switch configuration time is port-count independent. This is a remarkable result, since this time delay includes the time it takes to extract and process the labels, to solve the contentions and to drive the $1 \times 8$ (or $1 \times 4$) optical switch. Moreover, due to the parallel and autonomous operations of the switch modules, this is
the time to configure the whole switch architecture. The configuration time can be further reduced simply employing a faster FPGA, since 10 ns out of the total 25 ns are due to the single clock-cycle (100 MHz) required by the FPGA to perform its functionalities.

The flow control functionality is investigated by employing two distinct implementations. A FPGA-based implementation is tested successfully in an experimental set-up. Contended packets are correctly retransmitted and the round robin algorithm employed as priority policy ensures balanced buffer utilization. This implementation would require a dedicated electrical network, solution that is not practical in a large-scale system as a DC network. As a consequence, an all-optical implementation is considered. To reduce the number of required optical transmitters, part of the optical power of the extracted in-band label is employed to generate an optical acknowledge message. A semiconductor optical amplifier driven by the switch controller is used as modulator to generate the optical acknowledge. This signal is sent back employing the same optical path used for the packet transmission. An extra receiver is needed at the transmitting node to detect the optical acknowledgement. However, no extra optical links are required.

The integration of some of the functional blocks required by the proposed optical switch architecture is demonstrated in this work. Integration of the optical switch, label extractor and channels demultiplexer is experimentally investigated by using photonic integrated circuits. Channels demultiplexing and labels extraction functionalities are demonstrated by employing SOI integrated chips. Error free operations on the packet payload and the packet labels with 0.5 dB penalty is demonstrated. Also, the integration of the other functionalities required by a single architecture module is discussed.

**Port-count scalability**

The port-count scalability of the proposed architecture needs be to discussed considering two main aspects: the control and the practical implementation.

From a control point of view the port-count does not represent a problem. The employed radio frequency in-band labelling technique allows addressing a large number of ports. In this work, it has been shown that
512 distinct destinations can be addressed employing 3 label wavelengths each of them carrying 3 tones. If a larger number of ports needs to be addressed more tones can be encoded in each label wavelength or more label wavelengths can be inserted in the signal spectrum, without affecting the packet payload. The limits of this technique have been presented in [84]. Moreover, the modular structure of the architecture in combination with the parallel label processing employed allows the switch configuration time to be port-count independent.

Different considerations can be given about the optical switch implementation. Basically, the port-count scalability of the architecture is limited by the splitting losses of the broadcast stage of the optical switch required in each module. As already discussed in Section 4.4, wavelength division multiplexing is the key to scale the proposed architecture to a large port-count. However, as the number of wavelength channels transmitted in a single fiber increases, high extinction ratio and low cross-talk AWGs are required on chip. Nonlinear effects in semiconductor optical amplifiers and optical signal-to-noise degradation have also to be considered.

Latency

Main advantage of the novel architecture is that the overall switch configuration time is reduced to the time needed to extract and process a packet label (all labels can be extracted and processed in parallel and autonomously), to solve the contentions in a single $M \times 1$ wavelength selector and to drive a single $1 \times F$ optical switch. This leads to a promising result: 25 $\text{ns}$ are required to carry out all the needed operations, regardless the number of employed label wavelengths and radio frequency tones. Parallel label bits detection performed by the dedicated electronic circuitry takes 15 $\text{ns}$, while label decoding, contention resolution and switch control are performed by the employed FPGA in only 1 clock-cycle (10 $\text{ns}$). Alternative implementations of the label bit detection circuitry may lower the required processing delay. In addition, the employment of an FPGA with a faster processor will further reduce the control time. However, the packet latency may result in considerably larger time due to possible contentions and subsequent packet retransmission. Since the buffering is performed at the edges of the switching matrix the total system latency is strongly
dependent on the system round trip time and thus on the host-switch distance. Positioning the electrical buffers and the optical transceivers as close as possible to the optical switch would mitigate this limitation.

**Data-rate**

The maximum data-rate that can be employed is limited by two functional blocks of the proposed architecture. The optical switch maximum data-rate depends on the wavelength selective stage pass-band. For example, the employed $1 \times 4$ has a 3-dB bandwidth greater than 100 $GHz$ [100]. This switching module is data-format agnostic. However, in this context it is the SOA-based wavelength conversion that limits the operation to intensity-modulated signals. The employment of optoelectronic wavelength converters will solve this issue.

**Throughput**

Employing a centralized large port-count optical packet switch at the cluster layer of the data center architecture would enhance the resource sharing between the servers of a datacenter. In this scenario, all the cluster of the data center have equal available bandwidth and equal logical distance from each other. This will allow to spread the workload between the massive hardware equipment available in a DC network in a balanced manner. In theory, a perfect balanced traffic in a system employing the proposed architecture will allow to reach 100% of bandwidth utilization, i.e. maximum throughput.

**Integration**

In this thesis the possible integration of the required $1 \times F$ optical switch, the channel demultiplexer and the label extractor has been demonstrated. It is important to remark that the employed photonic integrated circuits have not been designed and fabricated specifically for the proposed architecture. Available chips have been used to demonstrate that it is possible to integrate the functional blocks required by the novel architecture. Chips
with a dedicated design would allow better results in terms of total system losses and chip footprint size, reducing both fabrication and operation costs.

The SOI platform also allows the integration on the same chip employed as label extractor of high-efficiency photo-detectors and trans-impedance amplifiers that can be used to detect and amplify the packet labels before the required processing. The integration of SOA-based wavelength converters is also feasible.

To take advantage of both material platforms investigated, namely SOI and InP, an hybrid integration of the required switch functional blocks seems to be the optimal solution.

The simplicity and the limited computational requirements of the employed control system (label processor and switch controller) allow easily integration of the architecture control. The basic limitation is the number of required input/output pins to process all signals in parallel, number that depends on the switch port-count.

**Power consumption**

Due to the early stage of the switch architecture implementation, few considerations can be made about the power consumption. For example, the $1 \times 4$ integrated optical cross-connect employed in this work has been shown to consume 3 $\text{pJ/bit}$ while routing 40 $\text{Gb/s}$ signals (not considering the power supply and the temperature controller) [100]. In the system level experiments carried out in this thesis, this value is significantly increased by the switch controller. Consider that only the employed FPGA development board consumes 25 $\text{W}$. A fair power consumption analysis would require an experimental set-up employing integrated electronic and optical components specifically designed for the proposed architecture. It can be said that scaling the switch port-count will correspond to increased power consumptions due to the higher driven current required to bias longer semiconductor optical amplifiers. However, the power consumption is not data-rate dependent, thus higher power efficiency can be achieved by using higher data-rate.
Alternative approaches

The approach employed in this thesis to design a high performance data center network is to employ a centralized, large port-count optical packet switch at the cluster level of the network. Such a switch, employing WDM, could guarantee the large port-count and high communication bandwidth required in the cluster interaction process, while providing low transmission latency due to its modular structure and distributed control system. Moreover, the proposed data center network could benefit of an enhanced resource sharing and would allow the opportunity of spreading the servers workload in a balanced manner.

Due to the bimodal distribution of the data center traffic, hybrid electrical-optical solutions have been considered [101–103]. Basically, the data center traffic consists of short lived flows (small control packets) and long lived flows (fragmented portions of large files). The hybrid approach relies on the employment of standard electronic packet switches to handle the short lived flows and optical circuit switches to forward the long lived flows to their destination. A parallel optical network is then available in case large bandwidth is required in the inter-rack or inter-cluster communication process. However, this approach needs traffic control and provisioning which is rather difficult in a data center environment and may lead to undesired system latency and reduced throughput.

An all-optical solution in which both packet and circuit switching are performed by two parallel optical networks represents the best solution to face the increasing bandwidth requirements of the data center networks, as confirmed by recent projects such as [104]. The proposed optical packet switch is suited for this hybrid implementation.

The study of optical packet switching reported in this thesis is focused on the physical implementation of an optical packet switch and its controller. The data center network management has not been considered. In this contest, the employment of software-defined networking (SDN) is of continuously increasing interest [105–107]. It consists in decoupling the control plane, where the routing decisions are made, and the data plane of the network, where the packets are effectively forwarded. The control plane is directly programmable and a more sophisticated traffic management is enabled by the physical separation from the forwarding plane. The control is centralized in software-based controllers that can control or modify the
flow table of network devices (physical or virtual) through a standard interface, OpenFlow is the dominant protocol [108]. In the contest of this thesis, SDN could be used to control, redirect and balance the traffic flows of the ToR and cluster switches maximizing the throughput of the centralized optical packet switch.

**Recommendations**

The implementation of the novel optical packet switch architecture described and studied in this thesis is still in its infancy. This work shows promising results towards the fabrication of a fast large port-count integrated optical packet switch for high performance data center networks. However, the realization of a prototype that can be tested in a real data center environment requires further research. The design and fabrication of dedicated photonic integrated circuits will allow to exploit the real potentialities of the novel architecture.

The experiments results presented in this work are obtained considering synchronous traffic, considering packets with fixed size. Testing the switch architecture under asynchronous operations, considering variable packet size would require enhanced functionalities in the switch controller and in the contentions solving process. This may affect the control complexity and, as a consequence, the system latency.
Bibliography


munications. Optical Society of America, 2011.


no. 4, pp. 6–20, 2007.


A. Rohit, A. Albores-Mejia, J. Bolk, X. Leijtens, and K. Williams, “Multi-path routing in an monolithically integrated 4x4 broadcast and select wdm cross-connect,” in *European Conference and Exposition on Optical Communications*. Optical Society of America, 2011.


E. Aw, A. Wonfor, M. Glick, R. Penty, and I. White, “An optimized


## Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACK</td>
<td>acknowledge message</td>
</tr>
<tr>
<td>ADC</td>
<td>analog-to-digital converter</td>
</tr>
<tr>
<td>AWG</td>
<td>array waveguide grating</td>
</tr>
<tr>
<td>ASIC</td>
<td>application-specific integrated circuit</td>
</tr>
<tr>
<td>B2B</td>
<td>back to back</td>
</tr>
<tr>
<td>BER</td>
<td>bit-error-rate</td>
</tr>
<tr>
<td>BERT</td>
<td>bit-error-rate tester</td>
</tr>
<tr>
<td>BPF</td>
<td>band-pass filter</td>
</tr>
<tr>
<td>BS</td>
<td>broadcast stage</td>
</tr>
<tr>
<td>CDF</td>
<td>cumulative distribution function</td>
</tr>
<tr>
<td>CFP</td>
<td>C form-factor pluggable</td>
</tr>
<tr>
<td>CMOS</td>
<td>complementary metal-oxide-semiconductor</td>
</tr>
<tr>
<td>CRB</td>
<td>contention resolution block</td>
</tr>
<tr>
<td>DC</td>
<td>data center</td>
</tr>
<tr>
<td>DFB</td>
<td>distributed feedback</td>
</tr>
<tr>
<td>E-O</td>
<td>electrical to optical</td>
</tr>
<tr>
<td>ED</td>
<td>envelope detector</td>
</tr>
<tr>
<td>EDFA</td>
<td>erbium-doped fiber amplifier</td>
</tr>
<tr>
<td>FBG</td>
<td>fiber Bragg grating</td>
</tr>
<tr>
<td>FDL</td>
<td>fiber delay line</td>
</tr>
</tbody>
</table>

*continued on next page*
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPGA</td>
<td>field-programmable gate array</td>
</tr>
<tr>
<td>FSR</td>
<td>free spectral range</td>
</tr>
<tr>
<td>FTTH</td>
<td>fibre to the home</td>
</tr>
<tr>
<td>FWC</td>
<td>fixed wavelength converter</td>
</tr>
<tr>
<td>HPC</td>
<td>high-performance computing</td>
</tr>
<tr>
<td>LD</td>
<td>laser diode</td>
</tr>
<tr>
<td>LED</td>
<td>light emitting diode</td>
</tr>
<tr>
<td>LPF</td>
<td>low-pass filter</td>
</tr>
<tr>
<td>MAN</td>
<td>metro area network</td>
</tr>
<tr>
<td>MEMS</td>
<td>micro electro-mechanical system</td>
</tr>
<tr>
<td>MMF</td>
<td>multi-mode fibre</td>
</tr>
<tr>
<td>MMI</td>
<td>multi-mode interference</td>
</tr>
<tr>
<td>MMR</td>
<td>micro ring resonator</td>
</tr>
<tr>
<td>NRZ</td>
<td>non-return to zero</td>
</tr>
<tr>
<td>O-E-O</td>
<td>optical-to-electrical-to-optical</td>
</tr>
<tr>
<td>O-E</td>
<td>optical to electrical</td>
</tr>
<tr>
<td>OOK</td>
<td>on-off keying</td>
</tr>
<tr>
<td>OPS</td>
<td>optical packet switch</td>
</tr>
<tr>
<td>OSNR</td>
<td>optical signal-to-noise ratio</td>
</tr>
<tr>
<td>PBF</td>
<td>pass-band filter</td>
</tr>
<tr>
<td>PCB</td>
<td>printed circuit board</td>
</tr>
<tr>
<td>PIC</td>
<td>photonic integrated circuit</td>
</tr>
<tr>
<td>PLC</td>
<td>photonic lightwave circuit</td>
</tr>
<tr>
<td>QSFP</td>
<td>quad SFP</td>
</tr>
<tr>
<td>RAM</td>
<td>random access memory</td>
</tr>
<tr>
<td>RF</td>
<td>radio frequency</td>
</tr>
<tr>
<td>ROADM</td>
<td>rearrangeable add/drop multiplexer</td>
</tr>
<tr>
<td>RTT</td>
<td>round trip time</td>
</tr>
</tbody>
</table>

*continued on next page –*
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RX</td>
<td>receiver</td>
</tr>
<tr>
<td>SDN</td>
<td>software-defined networking</td>
</tr>
<tr>
<td>SerDes</td>
<td>serializer/deserializer</td>
</tr>
<tr>
<td>SMD</td>
<td>surface-mount device</td>
</tr>
<tr>
<td>SMF</td>
<td>single-mode fibre</td>
</tr>
<tr>
<td>SFP</td>
<td>small-factor pluggable</td>
</tr>
<tr>
<td>SNR</td>
<td>signal-to-noise ratio</td>
</tr>
<tr>
<td>SOA</td>
<td>semiconductor optical amplifier</td>
</tr>
<tr>
<td>SOI</td>
<td>silicon on insulator</td>
</tr>
<tr>
<td>TIA</td>
<td>trans-impedance amplifier</td>
</tr>
<tr>
<td>TOR</td>
<td>top-of-the-rack</td>
</tr>
<tr>
<td>TPM</td>
<td>trial partition machine</td>
</tr>
<tr>
<td>TX</td>
<td>transmitter</td>
</tr>
<tr>
<td>VCSEL</td>
<td>vertical cavity surface emitting laser</td>
</tr>
<tr>
<td>WC</td>
<td>wavelength converter</td>
</tr>
<tr>
<td>WS</td>
<td>wavelength selector</td>
</tr>
<tr>
<td>WSC</td>
<td>warehouse-scale computer</td>
</tr>
<tr>
<td>WSS</td>
<td>wavelength selective stage</td>
</tr>
<tr>
<td>WDM</td>
<td>wavelength division multiplexing</td>
</tr>
<tr>
<td>XGM</td>
<td>cross gain modulation</td>
</tr>
<tr>
<td>XPM</td>
<td>cross phase modulation</td>
</tr>
</tbody>
</table>
List of publications

Journals


International conference


• N. Calabretta, S. Di Lucente, Y. Nazarathy, O. Raz, and H.J.S. Dorren, “Scalable optical packet switch architecture for low latency and high load computer communication networks”, in Proc. of the 13th


Symposia


I can finally write the last section of my thesis, late, as usual. It is time to thank all those who have made this thesis possible, directly or indirectly. I have met people from all over the world in the last four years and I have learned from all of them, it has been a great experience.

First of all, I would like to thank my first promotor prof. Harm Doren for giving me the opportunity to work in the ECO group. I want to thank him for his professional supervision, for the patience shown during hundreds of meetings, for his advices and for guiding me to this important achievement. I hope I have fulfilled his expectations.

I would like to thank my copromotor dr. Nicola Calabretta. I have been extremely lucky to work under his guidance. His experience and his energy have motivated me everyday. I thank him for his support for work-related and private matters, I have learned a lot from him.

I would like to thank the committee members for reviewing my thesis and for accepting this role despite their tight schedule. I am particularly grateful to prof. Gabriella Cincotti. This experience is also due to her suggestions.

I would like to thank Chigo, Fausto, Patty and Julio for the useful discussions and, more important, for their friendship. I am thankful to Aaron, Karen, Prometheus, Nikos, Vlado, Jim, Haoshuo, Pang, Pinxiang, Roy, Cac and Yan. I will miss our international dinners, birthday parties and conference trips. I am grateful to Abhinav, Jun and Peter for the great time spent in the lab. Their help during the measurements has been essential for the results presented in this thesis. Many thanks to Jose, Brigitta and Jolanda for their assistance in the administrative matters and for making my work easier. I would like to thank all the other ECO members, it is just
impossible to mention one by one by name.

I am thankful to all my basketball teammates and coaches of Tantalus and Almonte for the tough practices, games and parties. I thank my “brother from Lebanon” Elie for being always present in the happy and the sad moments, I will always treasure his friendship.

I would like to thank my parents, my brother and my grandmother for their constant support and encouragement. I have dedicated this thesis to my niece Sara and my nephew Marco for the time I could not spend with them.

It has been a difficult year. It has been even more difficult being close to me. I am particularly indebted with my girlfriend Gabriella for her infinite patience. Her unconditional love has supported and inspired me continuously.

Stefano Di Lucente
Eindhoven, December 2013
The Netherlands
Stefano Di Lucente was born in Rome, Italy, in July 23rd, 1982. He received his BSc degree in electronic engineering from Roma Tre University in 2006. His BSc thesis was entitled "Development of a wireless device for biomedical data and signals acquisition: data transmission module". He received his MSc degree cum laude in electronic engineering from Roma Tre University in 2009. His MSc thesis was entitled "Scrambling techniques for OCDMA networks".

Since January, 2010, Stefano is a PhD researcher in the Electro-Optical Communications group (ECO), department of Electrical Engineering, Eindhoven University of Technology, under the supervision of Professor H. J. S. Dorren and Dr. N. Calabretta. Stefanos research was conducted under the Communication Technologies Basic Research and Application (COBRA) research program, funded by the Dutch government.

His research activities are focused on optical packet switching, optical labeling techniques and optical packet switch control systems. He has worked with integrated photonic devices, specifically InP optical integrated cross-connect and SOI integrated label extractor based on microring resonators. He has worked with FPGA boards to develop and implement switch control systems.

Stefano Di Lucente has published in international conferences and peer-reviewed journals and has acted as supervisor for MSc students. He has received the best student award OECC/PS for the paper entitled: “160 Gb/s Optical Packet Switch Module Employing SOI Integrated Label Extractor” during the 18th OptoElectronics and Communication Conference/Photonic in Switching Conference in Kyoto, 3rd July 2013.